MICRODEVICES
Physics and Fabrication Technologies

Semiconductor Materials

An Introduction to Basic Principles

B. G. Yacobi

Semiconductor Materials

MICRODEVICES
Physics and Fabrication Technologies

Series Editors: Ivor Brodie and Arden Sher
SRI International
Menlo Park, California

Recent volumes in this series:

COMPOUND AND JOSEPHSON HIGH-SPEED DEVICES
Edited by Takahiko Misugi and Akihiro Shibatomi

ELECTRON AND ION OPTICS
Miklos Szilagyi

ELECTRON BEAM TESTING TECHNOLOGY
Edited by John T. L. Thong

ORIENTED CRYSTALLIZATION ON AMORPOUS SUBSTRATES
E. I. Givargizov

PHYSICS OF HIGH-SPEED TRANSISTORS
Juras Polzela

THE PHYSICS OF MICRO/NANO-FABRICATION
Ivor Brodie and Julius J. Muray

PHYSICS OF SUBMICRON DEVICES
David K. Ferry and Robert O. Grondin

THE PHYSICS OF SUBMICRON LITHOGRAPHY
Kamil A. Valiev

RAPID THERMAL PROCESSING OF SEMICONDUCTORS
Victor E. Borisenko and Peter J. Hesketh

SEMICONDUCTOR ALLOYS
Physics and Materials Engineering
An-Ban Chen and Arden Sher

SEMICONDUCTOR DEVICE PHYSICS AND SIMULATION
J. S. Yuan and J. J. Liou

SEMICONDUCTOR LITHOGRAPHY
Principles, Practices, and Materials
Wayne M. Moreau

SEMICONDUCTOR MATERIALS
An Introduction to Basic Principles
B. G. Yacobi

SEMICONDCUTOR PHYSICAL ELECTRONICS
Sheng S. Li

Semiconductor Materials

An Introduction to Basic Principles

B. G. Yacobi
University of Toronto
Toronto, Ontario, Canada

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47942-7
Print ISBN: 0-306-47361-5

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

Preface

The technological progress is closely related to the developments of various materials and tools made of those materials. Even the different ages have been defined in relation to the materials used. Some of the major attributes of the present-day age (i.e., the electronic materials' age) are such common tools as computers and fiber-optic telecommunication systems, in which semiconductor materials provide vital components for various microelectronic and optoelectronic devices in applications such as computing, memory storage, and communication.

The field of semiconductors encompasses a variety of disciplines. This book is not intended to provide a comprehensive description of a wide range of semiconductor properties or of a continually increasing number of the semiconductor device applications. Rather, the main purpose of this book is to provide an introductory perspective on the basic principles of semiconductor materials and their applications that are described in a relatively concise format in a single volume. Thus, this book should especially be suitable as an introductory text for a single course on semiconductor materials that may be taken by both undergraduate and graduate engineering students. This book should also be useful, as a concise reference on semiconductor materials, for researchers working in a wide variety of fields in physical and engineering sciences.

B. G. Yacobi
Toronto

This page intentionally left blank

Contents

This page intentionally left blank

1

Introduction

During the recent decades, advances in semiconductor materials resulted in the development of a wide range of electronic and optoelectronic devices that affected many aspects of the technological society. From semiconductors to microelectronic and optoelectronic devices (i.e., integrated circuits and devices for the generation and detection of light) for information applications (i.e., computing, memory storage, and communication), these advances and applications were catalyzed by an improved understanding of the interrelationship between different aspects (i.e., structure, properties, synthesis and processing, performance, and characterization of materials) of this multidisciplinary field.

The main objective of this book is to provide an overview of the basic properties, applications, and major types of semiconductors, as well as characterization methods that are routinely employed in the analysis of semiconductor materials. For details on a wide variety of specific topics, a reader is encouraged to refer to the provided Bibliography section.

At this juncture, it would be useful to define a semiconductor. However, without the basic details related to the electronic energy band structure of solids, any definition would be incomplete. Thus, although a more adequate definition and description of semiconductors are given in the following chapters, we can now only say that (i) semiconductors have electrical resistivity in the range between those of typical metals and typical insulators (i.e., between about 10^{-3} and 10^{9} Ω-cm), (ii) they usually have negative temperature coefficient of resistance, and (iii) the electrical conductivity of semiconductors can be varied (in both sign and magnitude) widely as a function of, e.g., impurity content (e.g., doping), temperature, excess charge carrier injection, and optical excitation. (Note that in all these cases it is implied that charge is carried by electronic particles; this is to differentiate from materials that have high ionic conductivity.) These factors may affect the electrical conductivity of a given semiconductor to vary by several orders of magnitude. Such a capability of varying (or controlling) the electrical conductivity over orders of magnitude in semiconductors offers unique applications of these materials in various electronic devices, such as transistors, and various optoelectronic devices for generation and detection of electromagnetic radiation, including data transmission through fiber-optic networks (i.e., photonics). The present computer

technology is essentially based on the ability of transistors to act as fast "on" or "off" switches, whereas lightwave communication systems rely on semiconductor-material-based (i) lasers as the source for photons on the one end and (ii) photodetectors on the other end.

An important (and distinctive) property of a semiconductor is its temperature dependence of conductivity, i.e., the fact that the conductivity in semiconductors increases with increasing temperature, whereas the conductivity in metals decreases with increasing temperature. One of the important parameters that often determine the range of applications of a given semiconductor is the *fundamental energy band gap,* or as it is referred to in the subsequent description, the *energy gap,* E_g (i.e., the energy separation between the valence and conduction bands), which is typically in the range between 0 and about 4 eV for semiconductors. It should be noted, however, that for semiconductors the boundaries for both the resistivity (between about 10^{-3} and 10^{9} Ω-cm) and the upper limit of the energy gap (of about 4 eV) are only approximate. For example, some materials (e.g., diamond, having the energy gap of about 5.5 eV) also exhibit semiconducting properties if properly processed (e.g., doped) for applications in semiconductor devices.

Some examples of common semiconductors that are widely used in electronic and optoelectronic applications are group IV elemental semiconductors (e.g., Si and Ge), group III–V semiconductor compounds (e.g., AlAs, GaAs, GaP, GaN, InP, InAs, and InSb), group II–VI compounds (e.g., ZnS, ZnSe, ZnTe, CdS, CdSe, and CdTe), and group IV–VI compounds (e.g., PbS, PbSe, and PbTe). In addition to these elemental and binary semiconductors, materials such as ternary (e.g., $Al_xGa_{1-x}As$, $GaAs_{1-x}P_x$, and $Hg_{1-x}Cd_xTe$) and quaternary (e.g., $Ga_xIn_{1-x}As_yP_{1-y}$) alloys with "tunable" (adjustable) properties are also used in specific device applications. Among these semiconductors, Si is one of the most important materials for electronic devices (e.g., integrated circuits), since Si-related fabrication technology is most advanced and a nearly defect-free material is readily obtainable. Another important semiconductor for (high-speed) electronic and optoelectronic device applications is GaAs, which has superior electron mobility. A wide variety of semiconductor compounds, mentioned earlier, are commonly used in optoelectronic applications, such as light-emitting devices and radiation detectors. It should be noted that it is the energy gap of a semiconductor that determines the energy (or wavelength) of the emitted or absorbed electromagnetic radiation (in the ultraviolet, visible, or infrared ranges); the availability of a wide variety of semiconductors with appropriate energy gaps makes various semiconductor devices suitable for the detection and emission of electromagnetic radiation in these ranges. In addition, in the ternary and quaternary alloys, the energy gap is tunable by alloying various semiconductors, and that allows the flexibility of producing materials with desired properties by varying the composition (i.e., x and y in the chemical formulas).

The topics covered in this book are organized in chapters as follows. Chapter 2 presents a brief description of interatomic bonding, crystal structure, and defects. These concepts are important in understanding semiconductors as three-dimensional structures that are composed of assemblies of atoms, which are held together by the interatomic forces and which inevitably contain one or more types

of irregularities. This is followed, in Chapter 3, by an introduction to the basic principles of quantum mechanics, which is essential for the description of the electronic properties of semiconductors by using the energy band theory of solids that is also outlined in Chapter 3. These two chapters (i.e., 2 and 3) provide the necessary basis for further discussions in the following chapters dedicated to the basic properties, applications, types, and characterization of semiconductors.

Chapter 4 describes the basic principles and properties of semiconductors. These include concepts of electrons and holes in semiconductors, intrinsic and extrinsic semiconductors, donors and acceptors, majority and minority carriers, and equilibrium and nonequilibrium properties of carriers.

Chapter 5 provides an overview on semiconductor junctions and devices, which include a p–n junction, Schottky barrier, heterojunctions, diodes, transistors and optoelectronic devices.

Chapter 6 provides a classification scheme of semiconductors. In this scheme, some overlap between the sections is inevitable. The properties, specific applications and limitations of various types of semiconductors are discussed. These include some recent developments in various areas, such as wide energy-gap semiconductors and low-dimensional semiconductors. An introductory section of this chapter also provides an overview on semiconductor growth and processing methods.

In the final chapter (Chapter 7), basic characterization techniques are described. The characterization of various properties of semiconductors is essential, before their usefulness in various applications can be fully ascertained. Substantial effort is devoted to the development of submicron structures and devices for continuing miniaturization of the electronic technology. Recent advances in materials processing also led to a new field, nanotechnology or nanoengineering, i.e., manufacturing of various structures and devices on a nanometer scale. This development requires a better understanding of materials and device properties with corresponding high spatial resolution. The development of scanning probe techniques (such as scanning tunneling microscopy, which is capable of imaging surfaces on the atomic scale), offers a variety of methods for examining materials properties with the nanometer-scale resolution. Such novel characterization techniques, as well as various electrical and optical characterization methods, and structural and surface analysis methods are outlined in Chapter 7.

In most cases, equations and parameters are presented in SI units. However, for convenience and typical use in various applications (and in the literature), units such as electron-volt (eV), centimeter (cm), or Angstrom (Å) are also employed in the present text. Also note that some symbols throughout the book have various meanings depending on the context.

This page intentionally left blank

2

Interatomic Bonding, Crystal Structure, and Defects in Solids

2.1. INTRODUCTION

Semiconductors can be distinguished by several different ways depending on their properties and applications. For example, we can classify them on the basis of their electronic band structure, or the periodic table (e.g., groups IV, III–V, and II–VI compounds), or crystalline structure, or electrical properties. None of these classification schemes would be completely acceptable in all cases; therefore, we do not adhere to any specific and rigid scheme, but we use these various schemes as they become useful for the description of different materials and their applications.

In order to understand and to describe semiconductors, it is important to consider their interatomic bonding configurations, structural properties, and various imperfections present in the material.

In this chapter, we first discuss types of interatomic bonding present in different solids, and then introduce some definitions of the crystalline structure. Thus, although we can, in principle, use either the interatomic bonding or structural symmetry concepts for the classification purposes, such methods are not sufficient for the description of the physical properties and behavior of semiconductors. For the classification and detailed description of the physical properties of semiconductors, it is essential to consider the electronic band structure that is described in the following chapters.

The fundamental categories of solids, based on their structural order, are *crystalline*, *polycrystalline*, and *amorphous.* It should be noted that the majority of semiconductors used in electronic applications are crystalline materials, although some polycrystalline and amorphous semiconductors have found a wide range of applications in various electronic devices. In crystalline materials, the atoms are arranged in a periodic, regularly repeated three-dimensional pattern. For the description of crystals, one can define a *lattice* as a periodic array of lattice points

in three dimensions; each point in such an array has identical surroundings. The entire solid can be recreated by repetitive translation of the so-called *primitive cell* in three dimensions. For geometrical convenience, it is also possible to choose a larger atomic pattern, a *unit cell,* as a building block of the crystalline material. As described in the subsequent chapters, the description of semiconductors as solids based on the concept of periodicity is of great importance, since such a description becomes manageable by considering a single unit cell only, instead of an enormous collection of about 5×10^{22} atoms cm^{-3}.

In polycrystalline materials, numerous crystalline regions, called *grains*, have different orientation and are separated by a *grain boundary.* In contrast to crystalline materials, amorphous semiconductors have only short-range order with no periodic structure.

In this chapter, we also discuss various types of defects, i.e., all types of deviation from ideal crystal structures, present in a semiconductor. One of the major efforts of semiconductor technology is to reduce the undesirable defect densities to such low levels that do not influence semiconductor properties or device performance. Although the presence of defects in semiconductors should always be anticipated, and their possible effects on various semiconductor properties should always be considered, we first have to develop an ideal picture of the solid state as a system containing atoms in an infinite regular array. (However, note that even a free surface of a finite ideal solid is a defect, since it is a discontinuity containing defects such as dangling bonds.) The real solids can be understood on the basis of the effects that various irregularities have on both the nearly perfectly-ordered crystal structure and physical properties of the material.

It should be emphasized that not all the deviations from the ideal crystal structure are detrimental. On the contrary, some are deliberately introduced to produce materials and devices with desired properties. These are, e.g., *n*- and *p*-type doped regions and epitaxial multilayers in various devices which are discussed in the following chapters. We, thus, regard impurity atoms in crystals as point defects if they are detrimental in the utilization of the material; on the other hand, we refer to them as donors (or acceptors) and recombination centers if they are deliberately incorporated in the material in order to control electrical conductivity or optical properties.

For the description that follows in this chapter, some definitions and equations should be recalled: momentum, $p = m\mathrm{v}$; velocity of a wave $\mathrm{v} = \nu\lambda$, where ν is the frequency and λ is the wavelength of the wave; angular frequency, $\omega = 2\pi\nu$. Also, recalling the application of a wave approach to free particles (see Chapter 3), one can relate the *momentum p* to the wave vector k ($k = 2\pi/\lambda$) using the de Broglie relationship $p = h/\lambda$, where h is Planck's constant; thus, $p = hk/2\pi$ (note that in many cases, the reduced Planck's constant $\hbar = h/2\pi$ is used), and thus $p = \hbar k$.

2.2. INTERATOMIC BONDING

As mentioned earlier, solid-state materials can be distinguished by several different ways. An important classification scheme is based on the bonding

configurations of atoms. In a solid, atoms are held together by a balance between the attractive and the repulsive forces. The attractive forces are dominant at larger interatomic distances, whereas the repulsive forces are dominant at smaller distances. The energy $E(r)$ of a crystal can be described, as a first approximation, in terms of separate mechanisms of attraction and repulsion

$$E(r) = \frac{A}{r^n} - \frac{B}{r^m} \tag{2.2.1}$$

where r is the separation between nearest neighbor atoms; A, B, n, and m are constants specific to a given crystal; and the term (A/r^n) represents repulsion and $(-B/r^m)$ represents attraction. The equilibrium between the attractive and repulsive forces occurs at the interatomic spacing which is determined by the condition of minimization of the total energy (see Fig. 2.1). This condition is satisfied if each atom in a solid is in an identical situation. Thus, in crystalline materials the atoms are arranged in a periodic, regularly repeated three-dimensional pattern. As mentioned earlier, such a periodic array of atoms with long-range order is called a lattice, and the entire solid can be recreated by repetitive translation of the primitive cell (or, alternatively, of a unit cell) in three dimensions.

The interatomic forces depend on the distribution of the outer electrons around the nucleus, and they also depend on whether the solid is composed of one type of atom or more than one type of atom. In solids, there are five basic types of bonding. These are (i) *ionic*, (ii) *covalent*, (iii) *metallic*, (iv) *molecular* (*van der Waals*), and (v) *hydrogen types.*

In general, the electrons outside the closed shells, i.e., valence electrons (and their number) determine the type of bonding that a particular atom forms with other atoms. When the number of electrons outside a closed shell is small, these

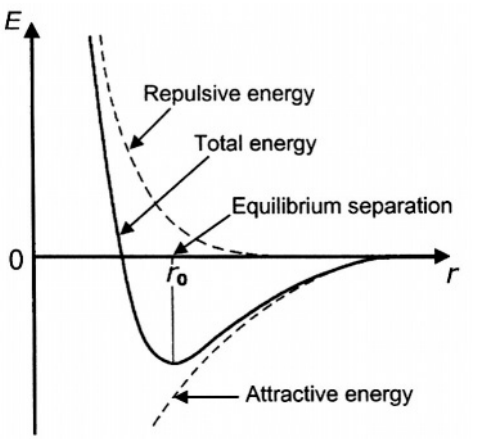

FIGURE 2.1. Schematic diagram of the dependence of the energy of the solid on separation between the atoms. The minimum of the total-energy curve corresponds to the equilibrium separation r_0.

outer electrons in the atom can be relinquished relatively easy, leaving a positive ion. On the other hand, when the atoms need only a few electrons to complete their outer shells, it is favorable to gain extra electrons to form a closed-shell configuration and, thus, to form a negative ion. In solids, the *ionic bonding* is formed when electrons are transferred from one atom to another. Thus, the ionic bonding is due to the Coulombic attraction between the oppositely charged ions. The ionic bond is nondirectional, i.e., a positive ion attracts an adjacent negative ion in all directions equally; this fact has an effect on the structure of the material.

In the cases of the atoms with outer shells about half filled, it is more favorable to share those electrons with similar atoms, leading to the *covalent bonding*, in which the electrons are shared between neighboring atoms. (Note that the charge density in the covalent bond is prevalent in the region between the neighboring atoms.) The covalent bond is directional; such a directional nature of sharing of valence electrons also determines the so-called *bond angle* (i.e., the angle between the covalent bonds). For example, each carbon atom, having four outer electrons in a shell that can accommodate eight electrons, can share its electrons with four equally spaced nearest neighbors in a tetrahedral configuration with a bond angle of 109.5° (this is discussed in the following section), and thus complete a closed-shell of eight electrons. The bonding in group IV semiconductors (e.g., Ge, Si, and diamond) is covalent with equal sharing of outer electrons. In many compounds, the electrons are not shared equally between the atoms, leading to partial ionization and mixed ionic and covalent bonding. Such a bonding is present in group III–V (e.g., GaAs) and group II–VI compounds (e.g., CdTe), where the sharing of outer electrons is unequal.

An important concept for the description of the interatomic bonding is the cohesive energy, or the binding energy, which is defined as the energy required to disassociate a solid into separate atoms, or separate ions in the case of ionic crystals. The covalent and ionic bond strengths are comparable, i.e., they are in the range between about 3 and 8 eV per atom for different semiconductors; e.g., the cohesive energy for silicon is about 4.7 eV per atom, whereas the cohesive energy for diamond is about 7.4 eV per atom.

In the *metallic bond*, since metal atoms easily give up their valence electrons to acquire stable closed-shell electron configurations and in the absence of atoms that would attract and confine them, the valence electrons, being free to migrate, are shared by all the ions in the solid. However, unlike the covalent bonding that is based on electron sharing and is directional, the metallic bond is nondirectional. The bonding in this case can be considered as an electrostatic interaction between the positive array of ions and the negative electron gas. The presence of free electrons (i.e., conduction electrons) elucidates the high electrical conductivity of metals. Although each single metallic bond is relatively weak, the cohesive energy is comparatively large, e.g., between about 1 and 4 eV per atom, since each metal atom interacts with many other atoms in the lattice.

The *molecular* or *van der Waals bonding* is due to the van der Waals attractive forces that are related to dipole–dipole interactions. These may include permanent dipole–permanent dipole force, permanent dipole-induced dipole force, and induced dipole-induced dipole force. This bonding is relatively weak, and in most cases it is

TABLE 2.1. The fractional bond ionicity, f_i, for selected semiconductors

Semiconductor	f_i (%)
Ge	0
Si	0
C	0
SiC	17.7
GaSb	26.0
GaAs	31.0
GaP	32.8
InAs	35.9
InP	42.1
ZnS	62.3
CdTe	67.5
CdS	67.9

overshadowed by one of the stronger bindings present. The van der Waals bond is usually of the order of 0.1 eV per atom or lower.

The *hydrogen bond* is formed in a compound of hydrogen and strongly electronegative atoms such as oxygen; in this case the hydrogen is positively charged and can be shared between the neighboring atoms or molecules. The hydrogen bond is of the order of 0.5 eV per atom.

As mentioned earlier, in many materials, the electrons are not shared equally between the atoms which results in partial ionization and mixed ionic and covalent bonding. It is important in this context to estimate the "degree of ionicity". A theory of the fractional bond ionicity, developed by Phillips, addresses the issue of the fractional ionic or covalent nature of a bond in dielectric solids; it considers a free electron gas with one average energy gap, referred to as the *Penn gap.* The *fractional bond ionicity* (f_i) is listed for selected semiconductors in Table 2.1 (see, Bibliography Section B2, Phillips, 1973; Phillips, Volume 1 of *Handbook on Semiconductors;* Moss, 1992). Phillips has demonstrated that properties such as, the cohesive energy and heats of formation depend linearly on f_i.

As seen in Table 2.1, in many semiconductors the binding is described as a mixture of ionic and covalent interactions.

2.3. CRYSTAL STRUCTURE

As mentioned in Section 2.1, for the description of the crystalline structure, one can define a lattice as a periodic three-dimensional array of points, i.e., lattice points, with identical surroundings. The actual crystal structures are formed by arranging single atoms (or identical group of atoms) on (or near) these lattice points. There are only 14 different ways to arrange the lattice points in three dimensions with each lattice point having identical surroundings. Thus, in principle, all crystal structures can be reproduced by a repetitive movement of an atom or a group of atoms at each point of one of the 14 point lattices, called

Bravais space lattices, which belong to one of the seven systems of axes (seven crystal systems). In a Bravais lattice, the points closest to a specific point are referred to as its *nearest neighbors*, and the number of nearest neighbors is called the *coordination number*.

The entire solid can be recreated by repetitive translation of the so-called primitive cell in three dimensions. For geometrical convenience, a larger pattern (a unit cell) can be chosen as a building block of the crystalline material. The unit cell is defined by three lattice translation vectors (i.e., vectors connecting lattice points). The unit vectors along the axes, **a**, **b**, and **c**, are called the *primitive vectors* and may be taken as the edges of a unit cell. The unit cell is translated by integral multiples of these vectors to form a lattice, so that the lattice sites can be defined in terms of a translation vector **R** as

$$\mathbf{R} = m_1\mathbf{a} + m_2\mathbf{b} + m_3\mathbf{c} \tag{2.3.1}$$

where m_1, m_2, and m_3 are integers.

The unit cell that describes seven crystal systems and 14 Bravais lattices is shown in Fig. 2.2, whereas Table 2.2 lists features of the crystal systems and Bravais lattices. The size and shape of the unit cell are defined by *lattice parameters* (or *lattice constants*) that include the length of unit cell edges and the angles between crystallographic axes.

An important treatment of crystal structures also involves the concepts related to symmetry properties (operations) applied to fixed lattice points. Concepts, such as *point symmetry*, e.g., reflection in a plane (i.e., *mirror symmetry*), inversion symmetry, rotation axes, and rotation–inversion axes are useful in both the understanding and description of crystalline materials, since an atom or a group of atoms in the lattice has specific symmetry properties. These symmetry operations involve a coordinate transformation. Thus, the mirror symmetry about the *yz*-plane is described by the transformation $y' = y$, $z' = z$, $x' = -x$, and the presence

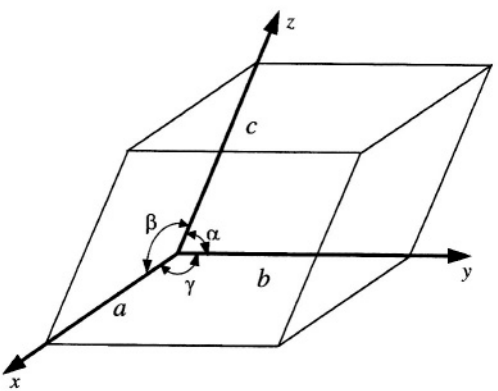

FIGURE 2.2. Crystallographic axes and lattice parameters.

TABLE 2.2. Description of the seven crystal systems and 14 Bravais lattices

Crystal system	The unit cell parameters	Bravais lattices
Cubic	$a=b=c$, $\alpha=\beta=\gamma=90°$	Simple, body-centered, face-centered
Tetragonal	$a=b\neq c$, $\alpha=\beta=\gamma=90°$	Simple, body-centered
Orthorombic	$a\neq b\neq c$, $\alpha=\beta=\gamma=90°$	Simple, body-centered, face-centered, base-centered
Trigonal (Rhombohedral)	$a=b=c$, $\alpha=\beta=\gamma\neq 90°$	Simple
Hexagonal	$a=b\neq c$, $\alpha=\beta=90°$, $\gamma=120°$	Simple
Monoclinic	$a\neq b\neq c$, $\alpha=\beta=90°\neq\gamma$ or $\alpha=\gamma=90°\neq\beta$	Simple, base-centered
Triclinic	$a\neq b\neq c$, $\alpha\neq\beta\neq\gamma\neq 90°$	Simple

of a mirror plane in a crystal structure is denoted by the symbol *m*. *Inversion symmetry* is represented by the coordinate transformation $x'=-x$, $y'=-y$, $z'=-z$, and is denoted by the symbol $\bar{\mathbf{1}}$. *Rotational symmetry* is related to such a rotation through a specific angle about a particular axis that results in an identical structure, whereas *rotation–inversion symmetry* axis represents rotation with simultaneous inversion. The collection of symmetry operations is called a *point group*. The point group operations together with a translation symmetry (in terms of a translation vector **R**), define the *space group* of a crystal.

The simplest three-dimensional unit cell is the *simple cubic* unit cell with an atom positioned at each corner of the cube. Two unit cells that are closely related to the simple cubic cell are the *body centered cubic* (bcc) unit cell and the *face centered cubic* (fcc) unit cell. In the bcc, a cell has an atom added at the center of the cube, whereas in the fcc unit cell each face of the cube contains an atom (in addition to the atoms at each corner).

Typical semiconductor lattices (i.e., diamond and zincblende) can be depicted by an arrangement shown in Fig. 2.3. These illustrations of such lattices depict the arrangement of the atoms in the unit cell; in real solids, however, the atoms, viewed as hard spheres, are considered to touch. Also note that in the simple cubic unit cell, only 1/8 of each of the eight corner atoms is in fact inside the cell, i.e., it contains total of one atom; the bcc unit cell, having an atom added at the center of the cube, in addition to the atoms at each corner, contains a total of two atoms; whereas the fcc unit cell containing an atom at each face of the cube, in addition to the atoms at each corner, but with only 1/2 of each of the six face atoms inside the cell, contains a total of four atoms.

As mentioned earlier, the number of nearest neighbors is called the coordination number, which essentially indicates the efficiency of packing of atoms together. For example, the simple cubic lattice has a coordination number 6, whereas the bcc lattice has a coordination number 8, and the fcc lattice has a coordination number 12 (i.e., the theoretical maximum).

Group IV elemental semiconductors, such as Si, Ge, and C (diamond) have the *diamond structure* that can be described as the fcc structure. Inside the cell,

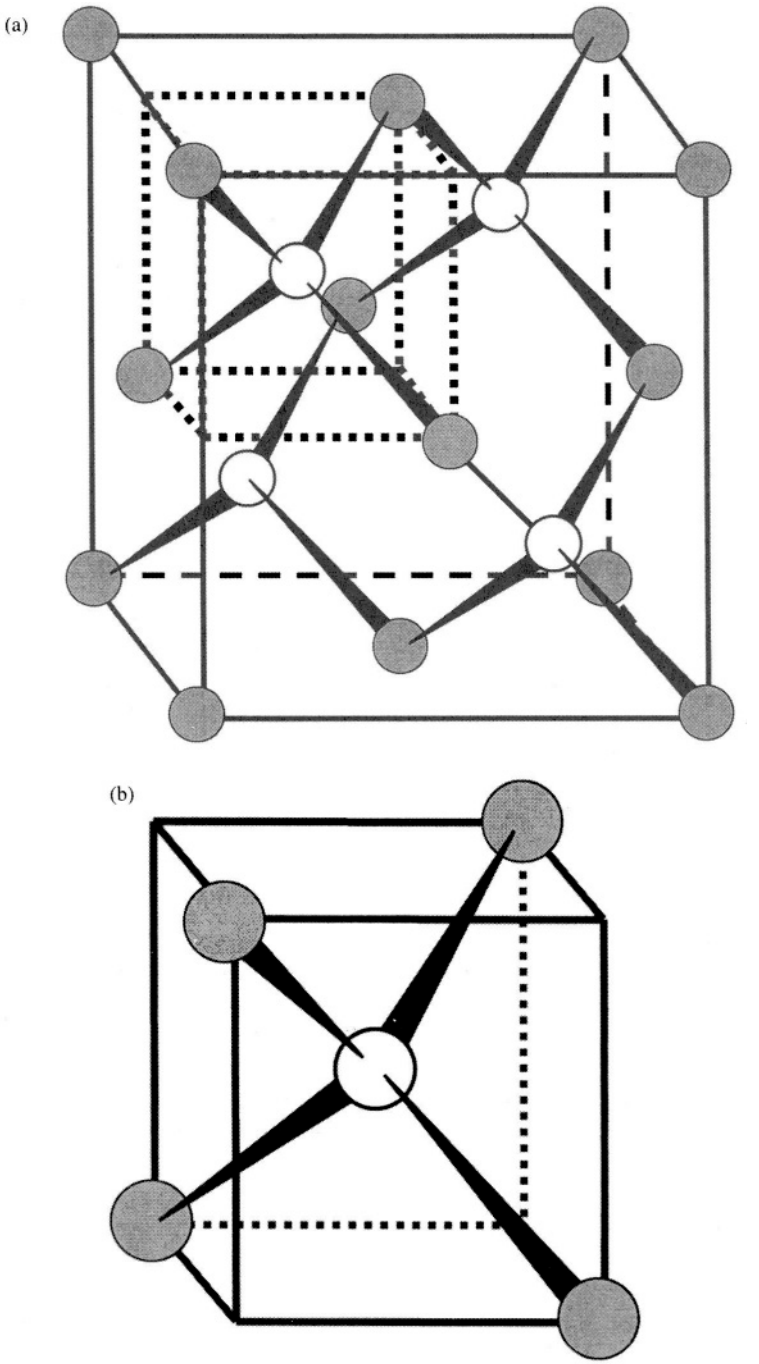

FIGURE 2.3. (a) Ball and stick model of the zincblende structure, (b) Ball and stick model of the tetrahedral unit of the zincblende structure.

however, there are four additional atoms, thus, there are eight atoms in the cubic unit cell; every atom in this arrangement has four equally spaced nearest neighbors in a tetrahedral coordination. The diamond structure can also be viewed as two interpenetrating fcc lattices displaced from each other along the cube body diagonal by one-fourth of its length.

Some group III–V semiconductor compounds (e.g., GaAs, InP, and InAs) and group II–VI compounds (e.g., ZnS, ZnSe, and CdTe) crystallize in the *zincblende structure* which differs from the diamond structure in that the four neighbors are all of the opposite atomic species. In other words, zincblende structure (see Fig. 2.3) can be obtained from the diamond structure by replacing carbon atoms alternately by Zn and S atoms in the case of ZnS (or Ga and As atoms in the case of GaAs). As mentioned earlier, the diamond structure can be described as two interpenetrating fcc lattices displaced from each other by one-fourth of the cube body diagonal distance. Thus, the zincblende structure differs from the diamond, in that the two interpenetrating fcc sublattices in the zincblende structure are of different atomic species, i.e., in the case of ZnS each Zn atom has four S nearest neighbors, and vice versa. One of the important features of the compound semiconductor, which offers great flexibility in "tuning" various materials properties, is the capability of forming the ternary (e.g., $Al_xGa_{1-x}As$) and quaternary (e.g., $Ga_xIn_{1-x}As_yP_{1-y}$) alloys by modifying the composition of each of the two interpenetrating fcc sublattices in the zincblende structure. (In the case of $Al_xGa_{1-x}As$, by varying the composition of group III sublattice, and in the case of $Ga_xIn_{1-x}As_yP_{1-y}$, by varying the compositions of both group III and group V sublattices.)

Some semiconductors can crystallize in several different structures depending on temperature or pressure, leading to polymorphism and existence of different polytypes. For example, ZnS can exist in several different structural forms, such as a cubic (zincblende) structure, or as a hexagonal (wurtzite) structure, or as various polytypes. The basic difference between these two structures (i.e., zincblende and wurtzite) is in the fact that, while the tetrahedral coordination of first-nearest neighbors is the same for both structures, the arrangement of second-nearest neighbors vary. This is clarified further in Fig. 2.4. The configurations of atoms in many crystalline solids can be represented by a close-packed arrangement of identical spheres. Consequently, as shown in Fig. 2.4, in three dimensions, there are three possible positions for close-packed planes on top of a first layer (e.g., A, B, or C). Such possible arrangements of planes leads to various structures with different stacking sequences. For example, ABCABC... stacking produces the fcc structure of one atom at each point of an fcc Bravais lattice, whereas ABAB... stacking produces the hexagonal close-packed structure based on the simple hexagonal lattice. More complex stacking sequences (i.e., polytypes) are also possible. In some compounds (e.g., ZnS and SiC), the zincblende (sphalerite) structure corresponds to the cubic packing sequence ABCABC... (denoted as 3C), the wurtzite structure corresponds to the hexagonal packing ABABAB... (denoted as 2H), whereas the polytypes 4H and 6H can be represented by ABACABAC... and ABCACBABCACB..., respectively.

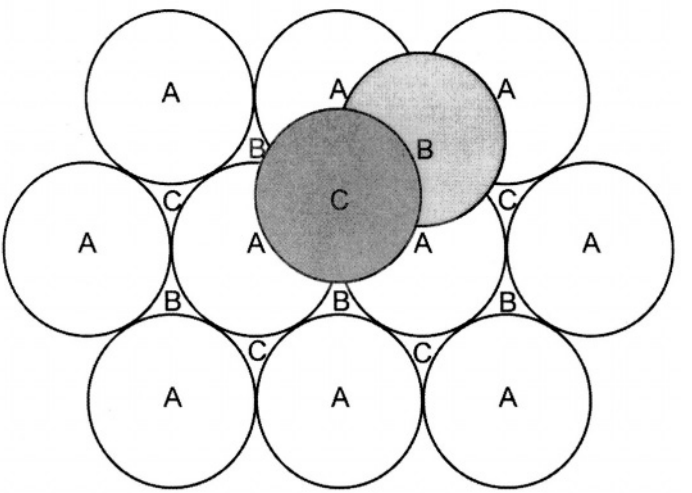

FIGURE 2.4. A close-packed layer of equal spheres; A, B, and C are three possible stacking positions for successive planes of atoms.

Ionic compounds, such as group IV–VI compounds (e.g., PbS, PbSe, and PbTe) crystallize in the *rocksalt structure* (NaCl structure), which can be viewed as a cubic structure consisting of two interpenetrating fcc lattices displaced from each other by one-half of a cube diagonal.

The three lattices, i.e., diamond, zincblende, and rocksalt, correspond to cubic system, and thus a single lattice constant represents the size of the unit cells. In the case of the wurtzite lattice corresponding to the hexagonal system, however, for complete description of the unit cell, a lattice constant c is also required. Table 2.3 presents structural properties of common semiconductors at room temperature (300 K); these include lattice structure, lattice constants (a and c, where applicable), and density.

We basically orient ourselves in a three-dimensional space. For convenience, we use various maps to find our locations. (We may use simple city maps showing alphabetical and numerical grids for x and y coordinates; for navigation, longitudes and latitudes are used.) Analogously, we have to use some means to orient ourselves in relation to the crystal lattice, for which we use notations for describing crystallographic directions and planes in crystals. Such notations are useful considering the fact that many crystals exhibit anisotropy of physical properties; in other words, there are differences in various properties of a material as a function of crystallographic directions.

In principle, crystallographic planes in a three-dimensional lattice can be identified by their intercepts along the crystal axes expressed as integral multiples of the primitive vectors, i.e., $m_1\mathbf{a}$, $m_2\mathbf{b}$, and $m_3\mathbf{c}$. The difficulties arise when a plane is parallel to one (or more) of the crystal axes. To avoid designations involving infinity, the reciprocals of these numbers converted to the smallest set of integers

TABLE 2.3. Structural properties of common semiconductors at room temperature (300 K).

Semiconductor	Lattice structure	Lattice constant a (Å)	Density (g cm^{-3})
Ge	D	5.646	5.327
Si	D	5.431	2.329
C (Diamond)	D	3.567	3.515
SiC (3C)	Z	4.360	3.166
SiC (6H)	W	$a = 3.081$, $c = 15.117$	3.211
InSb	Z	6.479	5.775
InAs	Z	6.058	5.66
GaSb	Z	6.096	5.619
InP	Z	5.869	4.787
GaAs	Z	5.653	5.318
AlSb	Z	6.136	4.218
AlAs	Z	5.662	3.760
GaP	Z	5.451	4.138
AlP	Z	5.464	2.40
InN	W	$a = 3.545$, $c = 5.703$	6.81
GaN	W	$a = 3.189$, c = 5.185	6.10
AlN	W	$a = 3.11$, $c = 4.98$	3.255
CdTe	Z	6.482	6.20
CdSe	W	$a = 4.300$, $c = 7.011$	5.81
CdS	W	$a = 4.136$, $c = 6.714$	4.82
ZnO	W	$a = 3.253$, $c = 5.213$	5.675
ZnTe	Z	6.101	5.64
ZnSe	Z	5.668	5.27
ZnS	Z	5.410	4.075
ZnS	W	$a = 3.822$, $c = 6.260$	4.087
PbSe	R	6.117	8.26
PbTe	R	6.462	8.22
PbS	R	5.936	7.61

Lattice structures: D, Diamond; W, Wurtzite (hexagonal); Z, Zincblende (cubic); R, Rocksalt. Note that some compounds (e.g., SiC and ZnS) can be grown in either W (hexagonal, 2H) or Z (cubic, 3C) structures, and they can also exhibit various polytypic structures (e.g., 6H); in the case of the wurtzite structure, for complete description, a lattice constant c is also required. For more details on a wide range of properties of various semiconductors, see Madelung, 1996.

of the same ratio are used, and these are called *Miller indices* (*hkl*) for the set of parallel planes. Thus, this method of reciprocals relates the intercepts to the unit cell parameters. For example, a plane with intercepts at $m_1 = 2$, $m_2 = 2$, $m_3 = \infty$, has reciprocals 1/2, 1/2, $1/\infty$, which can be designated as (0.5, 0.5, 0), or if converted to the smallest set of integers of the same ratio, as (1,1,0) and thus the Miller indices are denoted as (110), which represents a set of parallel planes. When an intercept along any axis is negative, a bar is placed over that index. Some planes of the cubic system are shown in Fig. 2.5. The set of symmetrically (structurally) equivalent planes is denoted by {*hkl*}; e.g., the set of all six cube faces is designated as {100}, i.e., these are (100), (010), (001), $(\bar{1}00)$, $(0\bar{1}0)$, $(00\bar{1})$. In the hexagonal structure, the planes are referred to three coplanar a-axes, at 120° to each other, and a fourth c-axis normal to them; the indices in this case are (*hkil*), where $i = -(h + k)$ and the index l is related to the intercept on the c-axis.

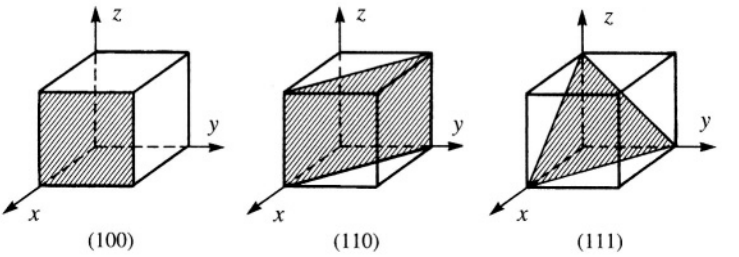

FIGURE 2.5. Miller indices of some planes in cubic lattices.

Directions in the crystal lattice are expressed as sets of three integers having the same relationship as the components of a vector in those directions. The crystallographic directions are specified with the notation [*hkl*], which are obtained by (i) identifying the three vector components expressed in multiples of the basis vectors defining the direction relative to the origin and (ii) reducing the three integers to their smallest values with the same ratio. The structurally equivalent directions are designated as ⟨*hkl*⟩; e.g., the body diagonals in the cubic lattice are designated as ⟨111⟩, and the equivalent directions denoted by ⟨100⟩ correspond to the crystal axes [100], [010], and [001]. In cubic systems, a direction [*hkl*] is normal to the plane (*hkl*); this is, however, not generally correct for other crystal systems.

The identification of the crystallographic planes and directions in the crystalline semiconductors is of great importance in the analysis and processing of semiconductors materials, and especially of the wafers, in which the surface is often chosen with specific orientation relative to a specific crystallographic plane. The choice of a specific surface orientation, e.g., is crucial in the device processing and the epitaxial growth, which are discussed in the following chapters. It is also important to know the crystal cleavage planes in different materials. For example, Si and Ge cleave along the {111} planes, whereas the III–V compound semiconductors with the zincblende structure preferentially cleave along {110} planes.

The determination of spacing between the lattice planes (i.e., *d*-spacing) is important for the crystallographic analysis of solids. This spacing (denoted d_{hkl}) indicates that d_{100} is the spacing between (100) planes. For the cubic lattice, the spacing $d_{100} = d_{010} = d_{001} = a$, where a is the lattice parameter. For each of the crystal systems, there is also a general expression for the *d*-spacing in terms of the parameters a, b, c, α, β, γ, and h, k, l. For example, for the cubic system ($a = b = c$, and $\alpha = \beta = \gamma = 90°$), the expression is

$$d_{hkl} = \frac{a}{(h^2 + k^2 + l^2)^{1/2}} \tag{2.3.2}$$

Thus, in this case, the spacing between the (111) planes is $d_{111} = a/3^{1/2}$. For more complex crystal systems, such a relationship is also more complex.

The analysis of crystal structures is of great importance in the description of materials; such an analysis is typically performed by employing X-ray diffraction

techniques. (Other techniques include electron diffraction method and neutron diffraction method.) The basic information that can be obtained from such diffraction patterns is the d_{hkl}-spacing, the crystal lattice type, and lattice parameters.

The well-known *Bragg diffraction law*, which is instrumental in the analysis of diffraction patterns, is now discussed briefly. In general, two interacting waves interfere constructively if they are in phase, i.e., in step, whereas the waves interfere destructively, i.e., cancel each other out, if they are out of phase. This is called diffraction and, in principle, it is valid for various waves, e.g., X-rays, electrons, neutrons, and visible light. The periodic array of atoms in a three-dimensional lattice diffracts waves, such as X-rays, which are commensurate with the interatomic spacing that is typically about a few Angstroms (see Table 2.3). For example, for X-rays (which have appropriate wavelengths commensurate with the interatomic spacing), incident on a crystalline specimen, strong diffracted beams are observed in the directions depending on the crystallographic structure and the wavelength of the incident radiation. The diffraction condition is illustrated in Fig. 2.6, which demonstrates that the path difference between X-ray waves, which are specularly reflected from two adjoining (parallel) planes, is $2d \sin\theta$. (Note that in specular reflection the angle of incidence equals the angle of reflection.) The constructive interference occurs (and diffracted spots or lines are observed and recorded), when this path difference is an integral number of wavelengths, i.e., diffraction occurs for the following condition

$$2d \sin\theta = n\lambda \tag{2.3.3}$$

where n (called the order of the corresponding reflection) is an integer, λ is the wavelength of the radiation, d is the periodic spacing between the planes, and θ is the Bragg angle. This is Bragg's diffraction law.

Another approach is that of von Laue. In this case, the crystal is considered to be composed of identical sets of atoms in a Bravais lattice, each site of which can reradiate the incident radiation in all directions. In such a case, in directions and at wavelengths, for which the waves scattered from all lattice sites interfere

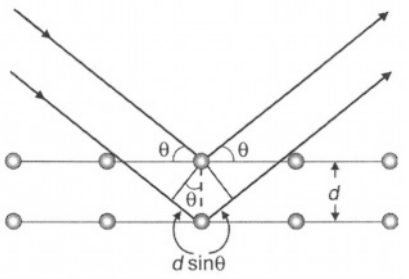

FIGURE 2.6. The Bragg reflection from lattice planes, separated by a spacing d, illustrating the diffraction condition, i.e., $2d \sin\theta = n\lambda$.

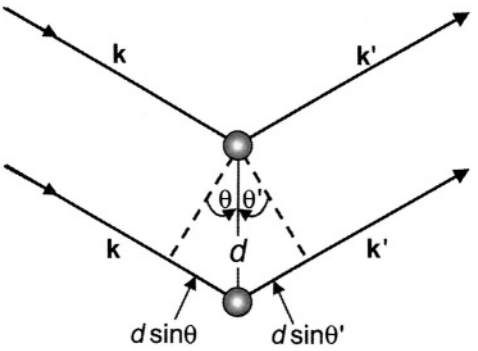

FIGURE 2.7. Schematic diagram of the path difference for two scattering centers separated by d.

constructively, sharp peaks would be observed. The condition for constructive interference in this case is shown in Fig. 2.7. In this case, the wave vectors (and directions) of the incident and scattered radiation are $\mathbf{k}$ (and $\mathbf{r}$) and $\mathbf{k}'$ (and $\mathbf{r}'$), respectively ($\mathbf{k} = 2\pi\mathbf{r}/\lambda$ and $\mathbf{k}' = 2\pi\mathbf{r}'/\lambda$). Thus, the path difference is

$$d \sin\theta + d \sin\theta' = \mathbf{d} \cdot (\mathbf{r} - \mathbf{r}') \tag{2.3.4}$$

The constructive interference occurs for

$$\mathbf{d} \cdot (\mathbf{r} - \mathbf{r}') = n\lambda \tag{2.3.5}$$

where n is an integer and λ is the wavelength of the radiation. Alternatively, one can write (recalling that $\mathbf{k} = 2\pi\mathbf{r}/\lambda$ and $\mathbf{k}' = 2\pi\mathbf{r}'/\lambda$)

$$\mathbf{d} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi n \tag{2.3.6}$$

In order to extend this condition for two scatterers to an array of scatterers (at each site of a Bravais lattice), it should be equally and simultaneously valid for all values of $\mathbf{d}$ that are represented by Bravais lattice vectors (i.e., $\mathbf{R} = m_1\mathbf{a} + m_2\mathbf{b} + m_3\mathbf{c}$), and thus one can write

$$\mathbf{R} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi n \tag{2.3.7}$$

which can also be expressed as

$$\exp[i(\mathbf{k} - \mathbf{k}') \cdot \mathbf{R}] = 1 \tag{2.3.8}$$

For the crystallographic analysis of the orientation of various planes and of the interplanar spacing, and of the diffraction spots in various types of diffraction patterns in crystals, it is convenient to use the *reciprocal space.* As discussed earlier, for the definition of the planes and directions in crystals, the Miller indices were

introduced that also have in their construction the reciprocals of lattice parameters. Thus, the concept of reciprocal lattice is consistent with the representation of Miller indices. In principle, there are two definitions of the reciprocal lattice; one is related to the crystallographic analysis, and the other to the description of the electronic band structure in the solid-state theory. The difference is in the definition of the reciprocal lattice vectors, i.e., whether they are defined as $1/d$ or $2\pi/d$. A factor 2π is introduced to make the reciprocal space the same as **k**-space (or, as sometimes referred to as *wave vector space,* or *momentum space* representation), which is more relevant to the description of the electronic band structure in solids. (This will become apparent in subsequent discussions; recall that the wave vector k is expressed as $k = 2\pi/\lambda$, where λ is the wavelength having the dimensions of distance.)

In analogy with a set of primitive vectors (**a**, **b**, and **c**) and a translation vector (**R**), defined earlier for the direct (real-space) lattice, it is also possible to define the points of the reciprocal lattice in wave vector space as a set of vectors **G** satisfying $\exp[i(\mathbf{G}\cdot\mathbf{R})] = 1$ for all **R**. Such a reciprocal lattice can be constructed by the primitive reciprocal lattice vectors $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$, which are defined by the following relationships

$$\mathbf{a}^* = \frac{2\pi(\mathbf{b}\times\mathbf{c})}{(\mathbf{a}\cdot\mathbf{b}\times\mathbf{c})} \tag{2.3.9}$$

$$\mathbf{b}^* = \frac{2\pi(\mathbf{c}\times\mathbf{a})}{(\mathbf{a}\cdot\mathbf{b}\times\mathbf{c})} \tag{2.3.10}$$

$$\mathbf{c}^* = \frac{2\pi(\mathbf{a}\times\mathbf{b})}{(\mathbf{a}\cdot\mathbf{b}\times\mathbf{c})} \tag{2.3.11}$$

where $V_{uc} = (\mathbf{a}\cdot\mathbf{b}\times\mathbf{c})$ is the volume of the unit cell of the direct lattice, and

$$\mathbf{a}^*\cdot\mathbf{a} = \mathbf{b}^*\cdot\mathbf{b} = \mathbf{c}^*\cdot\mathbf{c} = 2\pi \tag{2.3.12}$$

$$\mathbf{a}^*\cdot\mathbf{b} = \mathbf{a}^*\cdot\mathbf{c} = \mathbf{b}^*\cdot\mathbf{a} = \mathbf{b}^*\cdot\mathbf{c} = \mathbf{c}^*\cdot\mathbf{a} = \mathbf{c}^*\cdot\mathbf{b} = 0 \tag{2.3.13}$$

The reciprocal lattice can be expressed as

$$\mathbf{G} = n_1\mathbf{a}^* + n_2\mathbf{b}^* + n_3\mathbf{c}^* \tag{2.3.14}$$

where n_1, n_2, and n_3 are integers.

Using the above conditions ($\mathbf{a}^*\cdot\mathbf{a} = \mathbf{b}^*\cdot\mathbf{b} = \mathbf{c}^*\cdot\mathbf{c} = 2\pi$) one can write

$$\mathbf{G}\cdot\mathbf{R} = 2\pi(n_1m_1 + n_2m_2 + n_3m_3) \tag{2.3.15}$$

which indicates that $\mathbf{G} \cdot \mathbf{R} = 2\pi \times N$, where N is an integer, and thus, the reciprocal lattice vector satisfies the condition $\exp[i(\mathbf{G} \cdot \mathbf{R})] = 1$, and each vector of the reciprocal lattice is normal to a set of planes in the direct lattice. (Note that this then can be associated with the Miller indices.) Also note that from the earlier discussion, the wave vector $\mathbf{k}$ can be considered as a point in the reciprocal space. (This is useful in the definition of the Brillouin zones in Chapter 3.)

Following the earlier discussion, the von Laue condition for the occurrence of constructive interference $\{\exp[i(\mathbf{k}-\mathbf{k}') \cdot \mathbf{R}] = 1\}$ can now be expressed as $\mathbf{k}-\mathbf{k}' = \mathbf{G}$, i.e., the constructive interference occurs if the change in wave vector is a reciprocal lattice vector.

The concept of a reciprocal lattice facilitates the interpretation of diffraction patterns, since it allows identifying the sets of reflecting planes that cause a diffraction spot in a diffraction pattern. The reciprocal lattice can be constructed by plotting a normal to each set of planes in the direct lattice and specifying points along these normals at distances $1/d$ from the origin. The combined set of all these points produces the basic array of the reciprocal lattice. An interesting feature of the reciprocal lattice is that the fcc lattice in direct space transforms to a bcc lattice in reciprocal space, whereas the bcc lattice in direct space forms a fcc lattice in reciprocal space. This is important to remember, since many common semiconductors have the fcc structure. For the analysis of the diffraction patterns, it is possible to select a two-dimensional section out of the three-dimensional reciprocal lattice. It can be shown that the resulting specific array of reciprocal lattice points corresponds to the array of diffraction spots in the diffraction pattern. This construction also allows labeling individual spots with appropriate Miller indices. The correspondence between a specific section of the reciprocal lattice and the diffraction pattern can be understood by employing the *Ewald construction*, or the *Ewald sphere construction.* In this case, a reciprocal lattice representation of the crystal diffraction of radiation (with wavelength λ) is considered. Through the origin of the reciprocal lattice, a sphere (called the *Ewald sphere*) of radius $|\mathbf{k}| = 2\pi/\lambda$ is drawn. The wave vectors of the incident and scattered radiation, i.e., $\mathbf{k}$ and $\mathbf{k}'$ have equal magnitudes (lengths), since the diffraction is an elastic process, and thus they can be radii of the sphere. Some points of the reciprocal lattice touch the surface of the sphere, and the diffraction occurs at the angle 2θ between $\mathbf{k}$ and $\mathbf{k}'$ (see Fig. 2.8).

It is important to note that the direct image (i.e., microscopic image) of a crystal describes a real crystal structure, whereas the diffraction pattern of a crystal represents its reciprocal lattice.

It is also useful to note that any periodic function in direct (real) space that has a periodicity described by $f(\mathbf{r} + \mathbf{R}) = f(\mathbf{r})$, can be described by the three-dimensional Fourier series, i.e.,

$$f(\mathbf{r}) = \sum_{\mathbf{G}} f_{\mathbf{G}} \exp(i\mathbf{G} \cdot \mathbf{r}) \tag{2.3.16}$$

Since a property of such Fourier series is that they can be inverted, this can also be used for the conversion from one space into the other (i.e., direct and reciprocal).

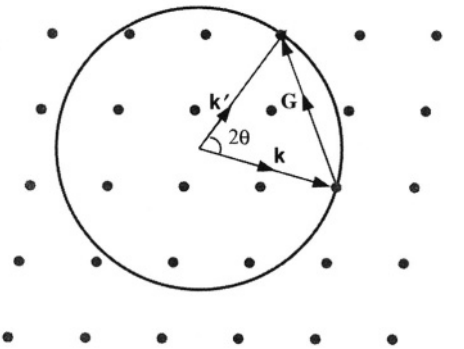

FIGURE 2.8. The Ewald sphere construction; the points correspond to reciprocal lattice points of the crystal; **k** is the wave vector of the incident wave and **k′** is the wave vector of the scattered wave. A sphere (called Ewald sphere) of radius $|\mathbf{k}| = 2\pi/\lambda$ is drawn about the origin of **k**. The diffracted beam, formed if the Ewald sphere intersects any other reciprocal lattice point, is in the direction $\mathbf{k}' = \mathbf{k} + \mathbf{G}$, where **G** is the reciprocal lattice vector.

2.4. DEFECTS IN SOLIDS

In general, defects can be defined as all types of deviation from ideal crystal structures. Semiconductors, as well as other solid-state materials, may contain a variety of defects, which are introduced in the material during, e.g., growth and processing. It is useful at this juncture to introduce general categorization schemes for defects.

We can distinguish between (i) structural defects and (ii) transient defects. In the former case, the correct arrangement of atoms in real crystals is permanently altered, whereas transient defects are elementary excitations such as phonons (i.e., quanta of lattice vibrational energy).

Structural defects can be classified as (i) point defects, such as substitutional and interstitial impurity atoms and vacancies, (ii) one-dimensional or line defects, such as dislocations, (iii) two-dimensional or planar defects, such as surfaces, grain boundaries and stacking faults, and (iv) three-dimensional or volume defects, such as voids and inclusions. It should be emphasized that defects in real solids may interact and form a variety of possible combinations. It is also important to note that defects may also act as attractive centers for free electrons or holes. Among these types of defects, only point defects, having sufficiently low formation energies, may be formed in thermal equilibrium (and their number increases with thermal activation). Thus, at a given temperature, point defects such as vacancies and interstitials are present in crystal structures. Other defects with higher dimensionality (e.g., dislocations) occur during growth and/or processing of semiconductors. Such defects may cause device failure, and they are also of great concern from the reliability point of view. In other words, in some cases, although a defect may not prevent initial device operation, it may do so during a prolonged device operation, and thus cause its failure. It should

be emphasized again that it is most essential to reduce the undesirable defect densities to sufficiently low levels that do not influence semiconductor properties or device performance.

Simple *native* (or *intrinsic*) *point defects* (shown in Fig. 2.9) are a *vacancy* (i.e., unoccupied site in the lattice) and an *interstitial* (i.e., an atom inserted into a space between the crystal structure sites). It should be noted that, although it is not shown in Fig. 2.9, localized regions of distortion are formed in the crystal due to these defects since the surrounding atoms in the lattice have to accommodate them. An isolated vacancy in the lattice is termed a *Schottky defect*, in which case the missing atom has migrated from the interior to the surface of the crystal or is trapped at an extended defect (e.g., dislocation). A vacancy associated with an interstitial atom (i.e., associated vacancy–interstitial pair) is referred to as a *Frenkel defect.* It should be noted that the number of these defects, besides being formed in thermal equilibrium in larger numbers with increasing temperature, may also rise due to such nonequilibrium means as high-energy electron or nuclear particle bombardment-induced displacement of atoms. *Foreign* (or *extrinsic*) *point defects*, shown in Fig. 2.9, are formed by impurities. In some cases, point defects may also form clusters (e.g., a vacancy pair) or precipitates of impurity atoms. The presence of vacancies in solids can elucidate the diffusion of an atom through the material by its movement to a vacancy, resulting in a migration of a vacancy in the opposite direction. The atomic diffusion may also occur when atoms (or ions) change their interstitial positions. The equilibrium density of vacancies (n_V) in a crystal, containing N atoms per unit volume at a temperature T, is given by

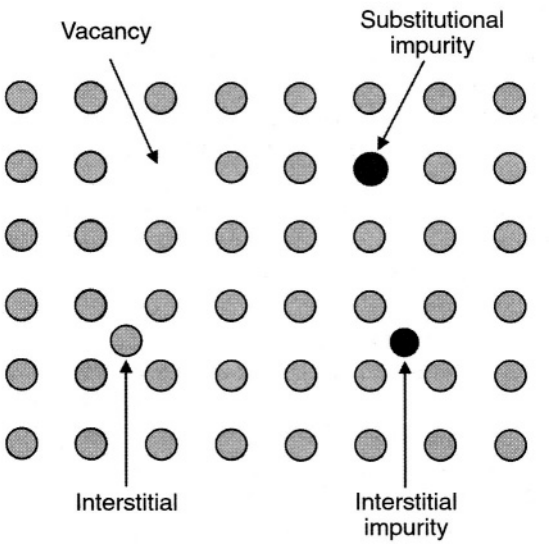

FIGURE 2.9. Simple native and foreign point defects. For simplicity, the distortion of the lattice around these defects is not shown.

$n_V \cong N \exp(-E_V/k_BT)$, where E_V is the energy of formation of a vacancy which is typically of the order of 1 eV. It should be noted that in compound crystals (AB), an antisite defect might form if an atom A occupies an atom B site, or vice versa. In GaAs, a dominant intrinsic defect, As_{Ga} antisite defect (i.e., an As atom located on a Ga site), referred to as EL2, is thought to be accountable for the semi-insulating property of undoped material.

Dislocations are important defects, since they play a very significant (and often harmful) role in the electrical and optical behavior of semiconductors. In addition, dislocations interact strongly with other defects and also lead to their generation. Two elementary types of dislocation are an *edge dislocation* and a *screw dislocation* (see Fig. 2.10). An edge dislocation can be described as the edge of an extra plane inserted into the crystal, whereas a screw dislocation introduces a helical distortion into the crystal. The presence of dislocations (the type and distribution) in crystalline semiconductors influences various properties, e.g., crystal growth, mechanical strength, and electronic properties. Dislocations may be formed, when point defects aggregate at an atomic plane, or they can be introduced when the stress causes atomic planes to slip past each other at high temperatures during growth or processing. Due to their high energy, dislocations do not occur in thermodynamic equilibrium, and large dislocation-free crystals (e.g., Si) have been produced. The nature of a dislocation is specified by the *Burgers vector* **b**. For the determination of **b**, one can use the *Burgers circuit,* which is essentially a geometrical construction (i.e., circuit) around the dislocation line; each step of the Burgers circuit corresponds to a lattice translation vector. In the case of a perfect crystal, the circuit will close at the starting point, whereas if taken around a dislocation the same circuit will not close. The closure vector in this case represents the magnitude of the dislocation. In order to construct the Burgers circuit, one can follow the right-handed (or clockwise) atom-to-atom steps (counting the same number of lattice translation vectors in all directions) around the dislocation line. The closure failure considered from "finish atom" to "start atom"

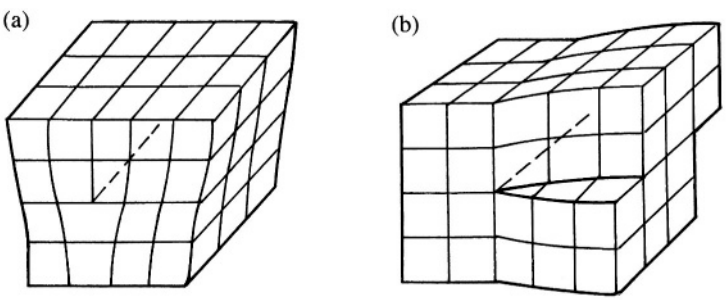

FIGURE 2.10. An edge dislocation (a) and a screw dislocation (b).

is represented by the Burgers vector. Thus, the Burgers vector corresponds to the displacement vector required for closing the Burgers circuit around the dislocation. (In this case, the Burgers vector is referred to as finish-to-start, right-handed Burgers vector.) For an edge dislocation, the Burgers vector is perpendicular to the dislocation line, whereas for a screw dislocation, the Burgers vector is parallel to the dislocation line. In real materials, dislocations are frequently intermediate between these two types of dislocation and they are curved. In order to minimize the total strain energy of the lattice, dislocations interact with one another that may lead to their annihilation. An interaction between point defects and dislocations also occurs resulting in impurity segregation or precipitation at dislocations and having a significant effect on the electronic properties of semiconductors. It should be noted that, in semiconductors, the dangling bonds at the dislocation core could capture electrons from the conduction band of an *n*-type semiconductor and lead to the formation of a cylindrical space-charge region around the dislocation. In commercial Si and GaAs wafers, dislocation densities are lower than 10^4 cm^{-2}. However, in thin epitaxial films, much higher dislocation densities may be present due to the interfacial misfit strain when the epitaxial layers are deposited on substrates with differing lattice constants; the lattice constant mismatch generates dislocations that may propagate into the epitaxial layer and degrade its electronic properties.

Surface defects are (i) free surfaces, which are present in all samples and (ii) interfaces that are formed unintentionally (unlike deliberately formed interfaces such as *p–n* junctions and contacts). In polycrystalline materials, lattice misorientations between the adjoining, randomly oriented crystallites result in *grain boundaries.* It should be noted that such grain boundaries could block the movement of dislocations. When the misorientation between adjacent grains is small (up to about 15°), the boundary is called *low-angle boundary*, which consists of an array of well-separated dislocations. In such a case, low-angle boundaries are called (i) *tilt boundaries* if they consist of edge dislocations and (ii) *twist boundaries* if they consist of screw dislocations. For larger angles, i.e., for *high-angle grain boundaries,* the boundary structure cannot be resolved into dislocations and it must be analyzed as a defect in its own right. In the case of a *twin boundary*, the atomic arrangements on each side of it are mirror images of each other. The grain boundary is considered as a two-dimensional defect, but in reality there is a specific thickness associated with this defect. In general, the grain boundaries contain high density of interface states that may trap free carriers, cause carrier scattering, and act as sinks for the impurity segregation.

Other surface defects are stacking faults. As mentioned in Section 2.3, crystalline structures can be described by their stacking sequence of layers in a close-packed arrangement. For example, the sequence ABCABC. . . corresponds to the fcc packing, and the ABABAB. . . represents the hexagonal close-packed structure. In these sequences, if a layer is missing (ABCACABC. . .) or an extra layer is inserted (ABCACBCABC. . .) in the cubic structure, then a stacking fault is formed. Analogously, a stacking fault is formed in the hexagonal packing when a layer in the C position is introduced or when a C layer replaces an A or B layer.

A *volume defect* is any volume in a semiconductor that differs from the rest of the crystal in composition, structure, and/or orientation. Volume defects are formed from vacancy clusters that may grow and eventually collapse to form dislocation loops. Some impurities may precipitate into a separate phase, and impurity atoms and vacancies may also form large three-dimensional aggregates. Other volume defects are voids, cracks, and inclusions.

It should be emphasized that one of the major objectives in the applications of semiconductors in various electronic devices is to control the influence of surfaces, interfaces, and grain boundaries on the properties of semiconductors and electronic devices. For example, great effort is directed towards processing steps that lead to the passivation of these defects. As discussed later, one of the main advantages of Si over other semiconductors is a relative ease of passivating the surface by oxidizing it in a controlled manner, i.e., forming a layer of stable native oxide, which substantially reduces the surface recombination velocity.

It should also be emphasized that, in general, defects have a crucial effect on semiconductor properties and the operation of semiconductor devices. The presence of defects in semiconductors may lead to (i) the introduction of energy levels in the energy gap that may affect the electronic properties of the material and (ii) a reduced carrier mobility due to increased scattering by defects. In semiconductors, transport properties depend on the presence of various defects which act as scatterers of carriers. These effects are often different as a function of temperature, so that electrical conductivity has a nonlinear dependence as a function of temperature, since the concentration (and electrical activity) of these defects have different temperature dependencies.

As mentioned earlier, not all the deviations from ideal crystal structure are detrimental. Some defects are deliberately introduced to produce materials and devices with desired properties. These are *n*- and *p*-type doped regions and epitaxial multilayers in various devices, which are discussed in the following chapters. To reiterate again, impurity atoms in crystals are considered as point defects if they are detrimental in the utilization of the material or device; but if they are deliberately incorporated in the material in order to control conductivity or optical properties, we refer to them as donors, acceptors, and recombination centers. It should also be emphasized that even trace amounts of impurity atoms can drastically affect the electrical properties of a semiconductor. Therefore, in semiconductors, it is crucial to control the amount of unintentional dopant atoms and to keep it below levels above which they influence device performance. Presently, the compositional purity of Si is such that the content of undesired impurities is less than one atom per billion (10^9) Si atoms. It should be noted, however, that in order to control the electrical properties of a semiconductor, dopant atoms in the range between one dopant atom per 100 million (10^8) semiconductor atoms to one dopant atom per thousand (10^3) semiconductor atoms are intentionally added to the material.

The typical techniques that are commonly employed in the analysis of structural defects include transmission electron microscopy (TEM), X-ray topography, scanning electron microscopy (SEM), scanning probe microscopy

(SPM) including scanning tunneling microscopy (STM), and Rutherford backscattering spectrometry (RBS). (For more details and other techniques, see Chapter 7.)

2.5. LATTICE VIBRATIONS

As discussed in Section 2.2, in solids there is an attractive force between atoms at large separations, whereas at smaller distances repulsion force dominates. The balance between these opposing forces determines the equilibrium separation. In the description of the crystallographic structure, a fixed lattice was considered. In the realistic case, however, even at low temperatures, atoms vibrate around the fixed lattice sites. As a result of such vibrations, electrons and (vibrating) ions in solids collide randomly with very high frequency (of about 10^{13} collisions s^{-1}) under the influence of the Coulomb force between them.

For sufficiently small vibrational amplitudes, the atoms in a solid can be considered as three-dimensional simple harmonic oscillators that can be characterized by a set of normal modes (waves) with specific frequencies. The energy of such modes is quantized. (Note that, as discussed in Chapter 3 in relation to quantum-mechanical description, on atomic-scale, confinement results in quantization, i.e., only certain normal modes are allowed.) The allowed energies are

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega \tag{2.5.1}$$

where $n = 0, 1, 2, 3 \ldots$; ω is the frequency of vibration; and $\hbar$ is the reduced Planck's constant. Note that the ground state (i.e., $n = 0$) of such a system is $E_0 = \frac{1}{2}(\hbar\omega)$ and is referred to as the zero-point energy. In such a system, the change in energy associated with the transition from one energy state to the next highest state is given by $\hbar\omega$. In analogy with the transitions involving a photon (i.e., a quantum of electromagnetic waves), the quantum of energy $\hbar\omega$ is called a *phonon*, i.e., a quantum of lattice vibrational energy. It is also important to emphasize at this juncture that from the definition of momentum (defined as h/λ, i.e., hv/c for a photon, and $\hbar\omega/u$ for a phonon, where c and u are the speed of light and speed of sound, respectively), it follows that a momentum associated with a phonon is relatively large as compared to a momentum of a photon which is very small. (Note that the speed of sound in a solid is much smaller than the speed of light.) Since each atom is coupled to several others, the vibrations of the atoms are not independent of each other, but can be considered as contributing to the vibrations of the whole lattice. The energy of such lattice vibrations is still quantized, and can only be changed by $\hbar\omega$. In thermal equilibrium (at temperature T), the average number of phonons with a specific frequency ω can be described by the *Bose-Einstein distribution function* as

$$n(\omega) = [\exp(\hbar\omega/k_B T) - 1]^{-1} \tag{2.5.2}$$

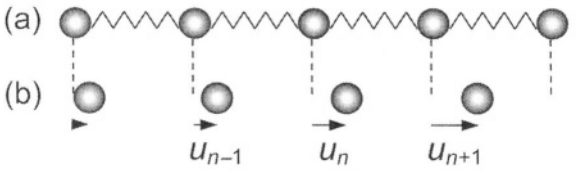

FIGURE 2.11. Schematic diagram of the geometry of a one-dimensional chain of identical atoms of mass m; (a) in equilibrium, (b) displaced from equilibrium.

Since atoms are held in their sites by strong elastic forces, the amplitudes of the vibrational displacements are much smaller than the interatomic spacing. Near the equilibrium, the net force between a pair of atoms is proportional to the departure of their separation from the equilibrium value. As mentioned earlier, since each atom is coupled to several others, the vibrations of the atoms are not independent of each other. For such a coupled system, it is of interest to determine the modes of vibration. For the description of such lattice vibrations, we can employ a model in which atoms are interconnected by springs (see Fig. 2.11); thus we can use Hook's law for the derivation of a relationship between the energy and frequency of the vibrational motion. The longitudinal vibrational motion of a one-dimensional chain of identical atoms of mass m, which are bound to one another by linear forces results in a periodic motion about the equilibrium positions. We assume (i) the displacement of the nth atom from the equilibrium position is u_n, (ii) Hook's law forces between neighboring atoms (i.e., analogous to atoms being bound together by ideal springs), and (iii) only significant force interactions are direct nearest neighbor interactions. Thus, the net force on the nth atom can be expressed as

$$F_n = \beta(u_{n+1} - u_n) - \beta(u_n - u_{n-1}) = \beta(u_{n+1} + u_{n-1} - 2u_n) \tag{2.5.3}$$

where β is Hook's law constant. (Note that in some sources in the literature, a symbol k is used for β.) Thus, we can write

$$m\frac{d^2u_n}{dt^2} = \beta(u_{n+1} + u_{n-1} - 2u_n) \tag{2.5.4}$$

The solution of the equation can be expressed as

$$u_n = A\exp[\mathrm{i}(\omega t - kna)] \tag{2.5.5}$$

and we can also write

$$u_{n+1} = A\exp\{\mathrm{i}[\omega t - k(n+1)a]\} \tag{2.5.6}$$

$$u_{n-1} = A\exp\{\mathrm{i}[\omega t - k(n-1)a]\} \tag{2.5.7}$$

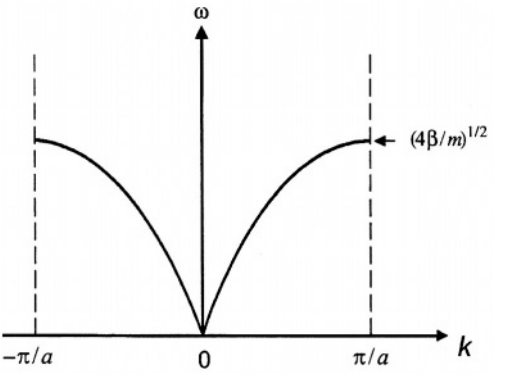

FIGURE 2.12. Dispersion relationship for the monatomic linear lattice.

Differentiating Eqs. (2.5.5)–(2.5.7) twice with respect to time and substituting into Eq. (2.5.4) gives a relationship between ω and k, i.e., the *dispersion equation*

$$\omega(k) = (4\beta/m)^{1/2}|\sin(ka/2)| \tag{2.5.8}$$

The absolute value sign is used, since ω is regarded a positive quantity in any case of the wave propagation to the right or to the left along the chain. The dispersion relation for the monatomic linear chain is presented in Fig. 2.12.

For the one-dimensional diatomic chain with alternating masses m and M (see Fig. 2.13) and with the similar assumptions as earlier, separate equations for light and heavy atoms can be expressed as

$$F_{2n} = m\frac{d^2u_{2n}}{dt^2} = \beta(u_{2n+1} + u_{2n-1} - 2u_{2n}) \tag{2.5.9}$$

$$F_{2n+1} = M\frac{d^2u_{2n+1}}{dt^2} = \beta(u_{2n+2} + u_{2n} - 2u_{2n+1}) \tag{2.5.10}$$

Solutions of the equations can be expressed as (note that since two types of atom with different masses are present, vibration amplitudes differ)

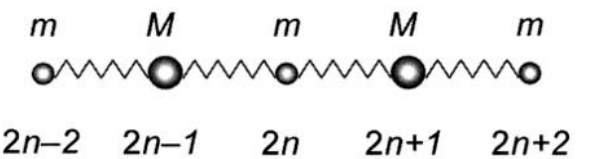

FIGURE 2.13. Schematic diagram of the geometry of the one-dimensional diatomic chain with alternating masses m and M.

$$u_{2n} = A \exp[\mathrm{i}(\omega t - 2kna)] \tag{2.5.11}$$

$$u_{2n+1} = B \exp\{\mathrm{i}[\omega t - k(2n+1)a]\} \tag{2.5.12}$$

We can also write

$$u_{2n+2} = A \exp\{\mathrm{i}[\omega t - k(2n+2)a]\} \tag{2.5.13}$$

$$u_{2n-1} = B \exp\{\mathrm{i}[\omega t - k(2n-1)a]\} \tag{2.5.14}$$

Differentiating Eqs. (2.5.11) and (2.5.12) twice with respect to time and substituting into Eqs. (2.5.9) and (2.5.10), and using Eqs. (2.5.13) and (2.5.14), gives a relationship between ω and k, i.e., the dispersion equation for a one-dimensional diatomic lattice

$$\omega_{\pm}^2 = \frac{\beta(m+M)}{mM}\left\{1 \pm \left[1 - \frac{4mM\sin^2(ka)}{(m+M)^2}\right]^{1/2}\right\} \tag{2.5.15}$$

The dispersion relation for the diatomic linear chain is presented in Fig. 2.14. The two branches of $\omega(k)$ are referred to as an *optical branch* $\omega_+(k)$, and an *acoustic branch* $\omega_-(k)$.

For small k, $\sin(ka) \cong ka$, and we can write

$$\omega_+(0) = \left[\frac{2\beta(m+M)}{mM}\right]^{1/2} \tag{2.5.16}$$

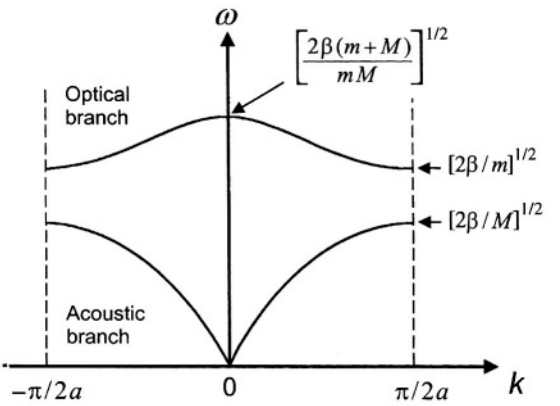

FIGURE 2.14. Dispersion relationship for the diatomic linear lattice; both optical and acoustic branches of the vibrational spectra are shown.

$$\omega_{-}(k) = ka\left[\frac{2\beta}{m+M}\right]^{1/2} (ka << \pi/2) \tag{2.5.17}$$

It should be noted that, although the frequency of vibration does not change much with the temperature, the amplitude of vibration increases with increasing temperature. Since the atoms in the solid are essentially coupled, i.e., they interact with each other, they also tend to vibrate synchronously, i.e., groups of atoms move in the same direction. This is an important fact which helps to understand various vibrational modes. To summarize briefly, two types of phonons are *acoustic phonons* (all atoms in the unit cell vibrate in the same direction) and *optical phonons* (atoms in the unit cell vibrate in opposite directions). The terms acoustic and optical refer to the facts that in the former case, for small values of k, the acoustic branch represents the propagation of sound waves; whereas the term optical implies that in the case of oppositely charged atoms in ionic crystals, this mode can be excited by the electric field related to the electromagnetic radiation of an appropriate frequency. There are two phonon modes, i.e., *longitudinal* (longitudinal acoustic, LA, and longitudinal optical, LO) when atoms move along the axis, and *transverse* (transverse acoustic, TA, and transverse optical, TO) when the displacements are perpendicular to the chain axis.

Phonons play an important role in transport properties of semiconductors. One of the important scattering mechanisms of carriers is due to phonons (see Chapter 4).

2.6. SUMMARY

The basic categories of solids (based on their structural order) are *crystalline, polycrystalline,* and *amorphous.* In crystalline materials the atoms are arranged in a periodic, regularly repeated three-dimensional patterns. In polycrystalline materials, numerous crystalline regions, called *grains,* have different orientation and are separated by *grain boundaries.* Amorphous semiconductors have only short-range order with no periodic structure. For the description of crystals, one can define a *lattice* as a periodic array of *lattice points* in three dimensions. The solid can be recreated by repetitive translation of the *primitive cell* in three dimensions. (For geometrical convenience, it is also possible to select a larger atomic pattern, a *unit cell.*)

In general, the valence electrons (and their number) and the composition (i.e., whether one type of atom or more are involved) determine the particular type of interatomic bonding. The *ionic bonding*, formed when electrons are transferred from one atom to another, is due to the Coulombic attraction between the oppositely charged ions. In the case of the *covalent bonding*, the electrons are shared between neighboring atoms. In the *metallic bond*, the valence electrons are shared by all the ions in the solid; the metallic bond can be considered as an electrostatic interaction between the positive array of ions and the negative electron gas. The relatively weak *molecular* or *van der Waals*

bonding is due to the electric dipole-dipole interactions. The *hydrogen bond* is formed in a compound of hydrogen and strongly electronegative atoms, such as oxygen. The interatomic bonding in group IV semiconductors (e.g., Si) is covalent with equal sharing of outer electrons. In compound semiconductors, such as, group III–V (e.g., GaAs) and group II–VI (e.g., CdTe) compounds, the electrons are not shared equally between the atoms, leading to mixed ionic and covalent bonding. A theory of the fractional bond ionicity, developed by Phillips, addresses the issue of the fractional ionic or covalent nature of a bond (the fractional bond ionicity, f_i, is listed for selected semiconductors in Table 2.1 (see Bibliography Section B2, Phillips, 1973; Phillips, in Volume 1 of *Handbook on Semiconductors,* Moss, 1992). For more details on interatomic bonding, see also (Bibliography Section B1) Ashcroft and Mermin (1976), Blakemore (1985), Burns (1985), and Kittel (1986).

One can define a *lattice* as a periodic three-dimensional array of *lattice points* with identical surroundings; the actual crystal structures are formed by arranging single atoms (or identical group of atoms) on (or near) these lattice points. The crystal structures can be reproduced by a repetitive movement of an atom or a group of atoms at each point of one of the 14 *Bravais space lattices,* which belong to one of the seven *crystal systems.* The presence of periodicity in crystalline solids is an important property that simplifies greatly the theoretical treatment of the solid-state. Many semiconductors crystallize in diamond and zincblende structures, in which each atom has four nearest neighbors in a tetrahedral configuration. The crystallographic planes in a three-dimensional lattice can be identified by *Miller indices* (*hkl*) for the set of parallel planes; the set of symmetrically equivalent planes is denoted {*hkl*}. The crystallographic directions in the crystal lattice are expressed as sets of three integers specified with the notation [*hkl*]; the structurally equivalent directions are designated as $\langle hkl \rangle$. The analysis of crystal structures is typically performed by using X-ray diffraction techniques (and the *Bragg diffraction law*). The basic information that can be obtained from such diffraction patterns is the crystal lattice type and lattice parameters. The analysis of the orientation of various planes and of the diffraction patterns in crystals is greatly simplified by invoking the concept of the *reciprocal space*, which is also important in the description of the electronic band structure in solid-state theory; the reciprocal space is also referred to as the **k**-space (or, as *wave vector space*, or *momentum space*).

An important treatment of crystal structures involves the concepts related to *symmetry properties* (operations) applied to fixed lattice points. The collection of symmetry operations is called a *point group.* The point group operations, together with a translation symmetry (in terms of a translation vector **R**), define the *space group* of a crystal. For more details on crystal structures and symmetry operations, see (Bibliography Section B1) Ashcroft and Mermin (1976), Blakemore (1985), Burns (1985), and Kittel (1986).

Semiconductors, in general, may contain a variety of *structural defects* and *transient defects.* Structural defects include (i) *point defects* (substitutional and interstitial impurity atoms and vacancies), (ii) *one-dimensional* or *line defects* (dislocations), (iii) *two-dimensional* or *planar defects* (surfaces, grain boundaries

and stacking faults), and (iv) *three-dimensional* or *volume defects* (voids and inclusions). Transient defects are elementary excitations such as *phonons* (i.e., quanta of lattice vibrational energy). In real solids, defects may interact and form a variety of possible combinations and they may also act as attractive centers for free electrons or holes. Defects play a crucial role in both (i) understanding the behavior and properties of semiconductors, as well as in (ii) the operation and reliability of semiconductor devices. As discussed in the following chapters, the presence of defects in semiconductors may lead to (i) the introduction of energy states in the energy gap, (ii) a reduction in carrier mobility due to increased scattering by defects, (iii) the changes in the recombination processes of excess carriers in optical phenomena, and (iv) defects may cause a device degradation and failure during a prolonged operation. One of the main objectives in semiconductor synthesis and device fabrication is to reduce the undesirable defect densities to sufficiently low levels that do not influence semiconductor properties or device performance.

PROBLEMS

2.1. Explain the need of taking reciprocals for Miller indices.

2.2. Calculate the distances between atoms in the three cubic structures, i.e., simple, fcc and bcc.

2.3. Find the set of all six cube faces designated as {100}.

2.4. Discuss qualitatively the meaning of the reciprocal lattice and its usefulness.

3

Band Theory of Solids

3.1. INTRODUCTION

In the analysis of semiconductors as solids, we have to consider a semiconductor as a collection of about 5×10^{22} ions cm^{-3} (depending somewhat on specific material) and the equal number of valence electrons which are, at temperatures above absolute zero, in a continual and random interaction with each other. The mass of the ions is much larger than that of the electrons; thus, the crystal structure is determined mainly by the positions of the ions. The behavior of valence electrons (i.e., negatively charged carriers), on the other hand, determines, e.g., the electrical properties of materials. In other words, it is the ability (or inability) of valence electrons to become "free" charge carriers to conduct electricity that distinguishes metals, semiconductors and insulators.

In the analysis of the electronic properties of semiconductors it is important to consider the electron transport processes, i.e., the motion of electrons through the material. For a detailed description of such a process, it would be necessary to consider the interactions between all the electrons and ions in the solid. As mentioned above, since the mass of ions is relatively much larger than that of the electrons, as a first approximation they can be considered in fixed positions. It is impractical to consider the interaction between all the electrons, since it would involve many particles. In order to avoid the so-called "many-body" problem, it is necessary to employ the single (one or independent) electron approximation, and to consider the electronic properties of all the electrons as a sum total of the actions of individual electrons. As mentioned in Chapter 2, the concept of periodicity is of great importance in the description of the crystalline solids, since such a description becomes manageable by considering a single unit cell only, instead of a collection of about 5×10^{22} atoms cm^{-3}.

In a solid crystal, electrons can be considered as moving in a periodic crystal potential $V(\mathbf{r})$, which is periodic with the periodicity of the lattice.

In the Section 3.2, we will first outline the basic concepts of quantum mechanics that are needed for the description of the electronic band structure and the various processes in semiconductors. In general terms, the electronic band theory of solids is concerned with the analysis of grouping of the electronic energy

levels into energy bands and with the description of the various properties and processes based on that analysis.

3.2. PRINCIPLES OF QUANTUM MECHANICS

3.2.1. The Wave – Particle Duality

The developments in instrumentation and experimental methods at the turn of twentieth century led to the discoveries that have established the foundations of the quantum mechanics. Investigations related to (i) the *blackbody radiation* (i.e., the characteristic radiation that a body emits when heated), (ii) the *photoelectric effect* (i.e., electron emission from matter by electromagnetic radiation of certain energy), and (iii) the *Compton scattering* (i.e., increase in the wavelength of X-ray or gamma rays scattered by free electrons) revealed the quantum nature of light, whereas the electron diffraction in crystals demonstrated the wave nature of particles. This *wave-particle duality* can be expressed in terms of the equation connecting the energy E of a photon to its frequency v,

$$E = h\nu \tag{3.2.1}$$

and in terms of the de Broglie expression for matter waves relating the momentum p of a particle to a wavelength λ associated with it

$$p = h/\lambda \tag{3.2.2}$$

where h is Planck's constant.

According to de Broglie hypothesis, various particles such as electrons and protons, or macroscopic (e.g., tennis) balls could also exhibit wave characteristics in certain circumstances. In this context, whenever the value of the de Broglie wavelength of a given particle (or a ball) is much smaller than the dimensions of the apparatus (i.e., its components), we can apply classical mechanics. This is, e.g., a case of a ball having a mass of 0.1 kg moving with the speed of 66 m s^{-1}. The de Broglie wavelength (i.e., $h/m\nu$) of such a ball is 6.63×10^{-34} J s/(0.1 kg $\times$ 66 m s^{-1}) $\cong 10^{-34}$ m, a very small number indeed, as compared to the objects that ball interacts with; thus, classical mechanics is applied to the description of the motion of the ball. On the other hand, if a particle, such as electron, has a speed of about 8×10^3 m s^{-1}, its de Broglie wavelength is about 6.63×10^{-34} J s/(9.11×10^{-31} kg $\times$ 8×10^3 m s^{-1}) $\cong 9 \times 10^{-8}$ m, which is larger than the spacing between atoms; thus, quantum mechanical description applies.

The wave–particle duality is resolved in terms of a wave packet that leads to the uncertainty principle. These considerations have established a new way of looking at atomic-sized objects, as being neither a particle, nor a wave, but entities that may exhibit either wave-like or particle-like properties depending on the experiment observed.

3.2.2. The Heisenberg Uncertainty Principle

According to quantum-mechanical principles there is inherent limitation to the accuracy of a measurement of a quantum-mechanical entity. This limitation is the result of (i) the particle–wave duality and (ii) inevitable interaction between the observer (i.e., his instrument) and the entity observed.

In quantum mechanics, a particle may be described in terms of a wave packet of length Δx, which is the resultant of individual waves with different amplitudes and nearly equal wavelengths. The position of the particle is defined so that the probability of finding the particle at the center of the packet is highest. For smaller Δx, the position of the particle is more accurately defined. However, since the range of wavelengths $\Delta\lambda$ in the packet now must be wider, there is greater uncertainty in the momentum Δp. The particle's uncertainties are related by Heisenberg's uncertainty principle

$$\Delta x\, \Delta p \geqslant \hbar/2 \tag{3.2.3}$$

which implies that on the atomic scale it is impossible to measure simultaneously the precise position and momentum of a particle. (In this expression, $\hbar = h/2\pi$.)

3.2.3. The Schrödinger Wave Equation

As mentioned above, electron diffraction in crystals demonstrates the wave nature of particles, which therefore can be described using the wave equation. The differential equation that describes the spatial dependence of the wave amplitude ψ of a vibrating system can be expressed as

$$\nabla^2\psi + \left(\frac{2\pi}{\lambda}\right)^2 \psi = 0 \tag{3.2.4}$$

where

$$\nabla^2\psi = \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} \tag{3.2.5}$$

Using the de Broglie expression for matter waves (i.e., $p = h/\lambda$ or $\lambda = h/m\mathrm{v}$), the wave equation can be written as

$$\nabla^2\psi + \left(\frac{2\pi m\mathrm{v}}{h}\right)^2 \psi = 0 \tag{3.2.6}$$

This equation can be further modified by replacing the kinetic energy (i.e., $m\mathrm{v}^2/2$) with the total energy (E) and potential energy (V); i.e., $m\mathrm{v}^2/2 = E - V$ (m is the mass of the electron). Thus, the wave equation can be written as (note that $\hbar = h/2\pi$)

$$\nabla^2\psi + \frac{2m}{\hbar^2}(E - V)\psi = 0 \tag{3.2.7}$$

Equation (3.2.7) is known as the (*time-independent*) *Schrödinger wave equation,* or simply the *Schrödinger equation,* which is an equation that describes the wave properties of electrons. Thus, an electron can be described by a *wave function* ψ (for this case, ψ is referred to as time-independent or stationary state wave function) that satisfies the Schrödinger equation, where E is the total energy of the electron, and V is the potential energy. V is zero for a free electron or for an electron inside a potential well, whereas it is a periodic function in the crystalline solid. The physical meaning of ψ is that $|\psi|^2 dxdydz$ represents the probability of finding an electron in the volume element $dxdydz$ in the vicinity of a position (x, y, z). For a one-dimensional case, $dxdydz$ can be replaced by dx, and the probability of finding the electron along the x-axis must be unity, i.e.,

$$\int_{-\infty}^{\infty} |\psi|^2 dx = 1 \tag{3.2.8}$$

which is called normalization condition on ψ, and the quantity $|\psi|^2$ is referred to as the *probability density.* The probability of finding the electron in a given interval $x_1 \leqslant x \leqslant x_2$ can be expressed as

$$P = \int_{x_1}^{x_2} |\psi|^2 \, dx \tag{3.2.9}$$

The Schrödinger wave equation is a second-order differential equation, the solution of which can provide E or ψ.

Being a time-independent equation implies that the properties of the atomic system surrounding the electron do not vary with time. For a case of a time-varying periodic potential, a *time-dependent Schrödinger equation* has to be employed. In three dimensions, a time-dependent Schrödinger equation can be written as

$$-\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi = i\hbar\frac{\partial\Psi}{\partial t} \tag{3.2.10}$$

where $\Psi(x, y, z, t) = \psi(x, y, z)\exp(-iwt)$ depends on both space and time.

Next, some specific applications of the Schrödinger equation are discussed. As shown later, the energy of free electrons is $E = \hbar^2k^2/2m$; whereas the solution of the Schrödinger equation (with appropriate boundary conditions) leads to quantized energy levels for electrons inside a potential well. In crystalline solids, the periodic potential $V(\mathbf{r})$ applies, and the solution of the Schrödinger equation in this case is in terms of the periodic Bloch functions

$$\psi(\mathbf{r}) = u(\mathbf{r})\exp(i\mathbf{k} \cdot \mathbf{r}) \tag{3.2.11}$$

where $u(\mathbf{r})$ is a periodic function (with the period of the crystal structure) which depends on the value of $\mathbf{k}$. As shown later, since interatomic distances and the potential energy distribution in real crystals depend on the direction, the $E(\mathbf{k})$ relationship is also a function of the crystallographic orientation.

At this juncture, it is important to emphasize that electron energies (e.g., in an atom) become quantized as a result of confinement. In other words, unlike free electrons that are allowed to have continuous range of energies, confined electrons are allowed only discrete values of energy. This is best understood by considering an analogy with waves in a string, i.e., two cases of the ends of the string being either fixed or free. In the case of fixed ends, the frequency spectrum of waves of a string is quantized (i.e., only certain normal modes are allowed). Such a restriction is removed if the ends of the string are not fixed. Thus, the energy quantization of electrons in an atom is a result of the fact that they have wavelike properties and behave in a manner analogous to a wave in a string with fixed ends. To summarize briefly, *quantization is a result ofparticle confinement.*

3.3. SOME APPLICATIONS OF THE SCHRÖDINGER EQUATION

3.3.1. Free Electrons

For free electrons, propagating in the x-direction, with no potential barrier restricting the propagation of the electron wave (i.e., $V = 0$), the Schrödinger equation can be written as

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}E\psi = 0 \tag{3.3.1}$$

In general, the solution of this differential equation can be expressed in terms of $\sin kx$, or $\cos kx$, or exponential functions $\exp(ikx)$ and $\exp(-ikx)$, where

$$k = (2m_e E/\hbar^2)^{1/2} \tag{3.3.2}$$

For this specific case, the solution in terms of exponential functions is most appropriate, and it can be written as

$$\psi(x) = A\exp(ikx) + B\exp(-ikx) \tag{3.3.3}$$

where A and B are constants. The first term in this equation for ψ corresponds to a wave traveling in the positive x-direction, whereas the second term corresponds to a wave traveling in the negative x-direction. From the above equation for k it follows that

$$E = \frac{\hbar^2 k^2}{2m_e} \tag{3.3.4}$$

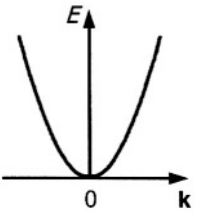

FIGURE 3.1. Plot of electron energy, E, as a function of wave vector $\mathbf{k}$ for free electrons.

which indicates that in the absence of any boundary conditions, all values of energy are allowed for free electrons. Since $E = p^2/2m_e$, the momentum $p = \hbar k$. Also, recalling that $p = h/\lambda = 2\pi\hbar/\lambda$, and thus, $k = 2\pi/\lambda$, where k is a wave vector of the electron. For free electrons, the relationship between energy and momentum, i.e., $E(\mathbf{k})$ relationship, is shown in Fig. 3.1.

3.3.2. Bound Electron in an Infinitely Deep Potential Well

For electrons bound between two infinitely high potential barriers (but free to move inside the well; see Fig. 3.2), the potential energy inside the well $V = 0$, and the Schrödinger equation for this one-dimensional box can be written as

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}E\psi = 0 \tag{3.3.5}$$

This is a differential equation having the general solution that can be expressed as

$$\psi(x) = A\sin kx + B\cos kx \tag{3.3.6}$$

where

$$k = (2m_eE/\hbar^2)^{1/2} \tag{3.3.7}$$

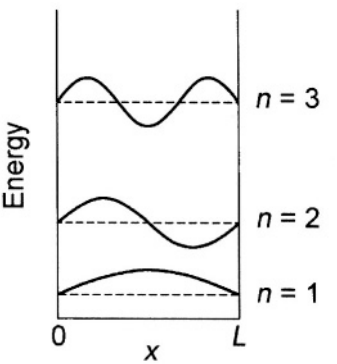

FIGURE 3.2. Schematic diagram of a ground state and two excited state energy levels and associated wave functions for an infinitely deep square potential well.

and A and B are constants, which in this specific case can be determined by considering the boundary conditions: $\psi(0) = 0$ and $\psi(L) = 0$. Thus, for $x = 0$, $\psi(0) = B = 0$, and $\psi(x = L)$ can be written as

$$\psi(L) = A \sin kL = 0 \tag{3.3.8}$$

which is satisfied only if kL is an integral multiple of π, i.e., if $kL = n\pi$, where $n = 0$, 1, 2, 3, ... Thus, since $E = \hbar^2 k^2 / 2m_e$,

$$E_n = \frac{\hbar^2}{2m_e} k^2 = \frac{\hbar^2 \pi^2}{2m_e L^2} n^2 \tag{3.3.9}$$

Therefore, because of the boundary conditions, only certain energy levels are allowed (see Fig. 3.2). The wave functions, correspondingly, can be expressed as

$$\psi_n(x) = A \sin \frac{n\pi x}{L} \tag{3.3.10}$$

where A is the normalization constant, which can be expressed as $A = (2/L)^{1/2}$ (see Problem 3.2).

For the three-dimensional case, i.e., in the case of an electron in a *three-dimensional box,* the expression for its energy is

$$E_n = \frac{\hbar^2 \pi^2}{2m_e} [(n_x/L_x)^2 + (n_y/L_y)^2 + (n_z/L_z)^2] \tag{3.3.11}$$

If the three-dimensional box is a cube of side L, the expression for the energy is

$$E_n = \frac{\hbar^2 \pi^2}{2m_e L^2} (n_x^2 + n_y^2 + n_z^2) \tag{3.3.12}$$

3.3.3. Bound Electron in a Finite Potential Well

For electrons bound in a potential well of finite depth V and width L (see Fig. 3.3), the potential energy inside the well (i.e., region II) $V = 0$, and

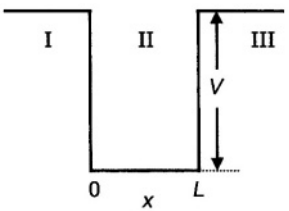

FIGURE 3.3. Schematic diagram for a square potential well of finite depth V.

the Schrödinger equation for this region can be written as (similar to a previous case of a bound electron in an infinitely deep potential well)

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}E\psi = 0 \tag{3.3.13}$$

with the allowed wave functions corresponding to Eq. (3.3.6), i.e., $\psi(x) = A\sin kx + B\cos kx$. In this case, however, the boundary conditions do not require for $\psi = 0$ at the walls, since according to quantum mechanics, there is a finite probability that the electron may be found outside the well (i.e., outside the region II), and thus the wave function is generally nonzero in regions I and III (see Fig. 3.3). The Schrödinger equation for these regions I and III can be written as

$$\frac{d^2\psi}{dx^2} - \frac{2m(V-E)}{\hbar^2}\psi = 0 \tag{3.3.14}$$

where $V > E$. The general solution to this equation is

$$\psi(x) = C\exp\left\{[2m(V-E)/\hbar^2]^{1/2}x\right\} + D\exp\left\{-[2m(V-E)/\hbar^2]^{1/2}x\right\} \tag{3.3.15}$$

where C and D are constants that can be determined by applying boundary conditions. In region I ($x < 0$), the second term must be ruled out (and hence, $D = 0$) in order to avoid an infinite value for ψ for large negative values of x. In region III ($x > L$), on the other hand, the first term must be ruled out (and hence, $C = 0$) in order to avoid an infinite value for ψ for large positive values of x. Thus, for region I ($x < 0$), the solution is

$$\psi_1(x) = C\exp\left\{[2m(V-E)/\hbar^2]^{1/2}x\right\} \tag{3.3.16}$$

and for region III ($x > L$),

$$\psi_{III}(x) = D\exp\left\{-[2m(V-E)/\hbar^2]^{1/2}x\right\} \tag{3.3.17}$$

Thus, these results indicate that for a finite potential well, the wave functions inside the well are sinusoidal, whereas the wave functions for these states decay exponentially with distance into the barrier regions (see Fig. 3.4). It is important to note that the wave functions are no longer equal to zero at the walls of the potential well, resulting in an increase of the de Broglie wavelength inside the well (i.e., in region II) and, thus, in lowering of the energy and momentum of the electron.

3.3.4. Electron Tunneling through a Finite Potential Barrier

In this case, an electron traveling in the positive x-direction in region I (with $V = 0$) encounters a potential barrier of height V (see Fig. 3.5). In region I,

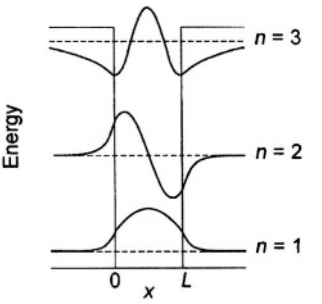

FIGURE 3.4. Schematic diagram of a ground state and two excited state energy levels and associated wave functions for a square potential well of finite depth.

the potential energy $V = 0$, and the Schrödinger equation for this region can be written as

$$\frac{d^2\psi_1}{dx^2} + \frac{2m}{\hbar^2} E\psi_1 = 0 \tag{3.3.18}$$

with the solution corresponding to

$$\psi_1(x) = A \sin kx + B \cos kx \tag{3.3.19}$$

In region II, the Schrödinger equation can be written as

$$\frac{d^2\psi_2}{dx^2} - \frac{2m(V-E)}{\hbar^2}\psi_2 = 0 \tag{3.3.20}$$

where $V > E$. The general solution to this equation is

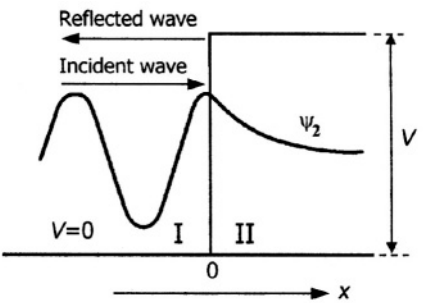

FIGURE 3.5. Schematic diagram of an incident, reflected, and penetrating waves of a particle encountering a potential energy barrier of height V. $\psi_2(x) = D\exp\{-[2m(V-E)/\hbar^2]^{1/2}x\}$.

$$\psi_2(x) = C\exp\left\{[2m(V-E)/\hbar^2]^{1/2}x\right\} + D\exp\left\{-[2m(V-E)/\hbar^2]^{1/2}x\right\} \quad (3.3.21)$$

At $x = 0$, $\psi_1 = \psi_2$; and in order for the wave function to remain finite at $x \to \infty$, C must be zero. Thus, at $x = 0$, we can write

$$A\sin kx + B\cos kx = D\exp\left\{-[2m(V-E)/\hbar^2]^{1/2}x\right\} \quad (3.3.22)$$

and, hence, $B = D$. In addition, at $x = 0$, $d\psi_1/dx = d\psi_2/dx$, and thus,

$$kA = -D\left[2m(V-E)/\hbar^2\right]^{1/2} \quad (3.3.23)$$

Taking all these into consideration, we can write

$$\psi_1(x) = \left\{-\frac{D}{k}\left[2m(V-E)/\hbar^2\right]^{1/2}\right\}\sin kx + D\cos kx \quad (3.3.24)$$

and

$$\psi_2(x) = D\exp\left\{-\left[2m(V-E)/\hbar^2\right]^{1/2}x\right\} \quad (3.3.25)$$

This implies that the electron, having lower energy E than the energy barrier height V, can penetrate the barrier by tunneling.

3.3.5. The Kronig–Penney Model (Electron in a Periodic Crystal Potential)

For an electron moving through a crystal lattice, the effect of the potential well of each ion in its path has to be considered. As mentioned earlier, in crystalline solids, the periodic potential $V(x) = V(x + a)$ applies (e.g., for a one-dimensional case), and the solution of the Schrödinger equation in this case is in terms of the periodic *Bloch functions*

$$\psi(x) = u(x)\exp(ikx) \quad (3.3.26)$$

where $u(x) = u(x + a)$, called a Bloch function, is a periodic function (with the period of the crystal structure) which depends on the value of k. The analysis for the realistic periodic potentials would be very complex, thus some simplifications are applied (see Fig. 3.6). According to the *Kronig–Penney model,* the electron motion is considered for a one-dimensional array of square potential barriers of width a, which are separated by potential barriers V_0 having width b (see Fig. 3.7). As shown later, the nature of the energy variation as a function of wave vector depends on the width, b, and height, V_0, of the barriers.

The solution of the Schrödinger equation has to be considered for two distinct regions (see Fig. 3.7). In region I, for which $0<x<a$, i.e., in the potential well, $V = 0$, and the Schrödinger equation can be written as

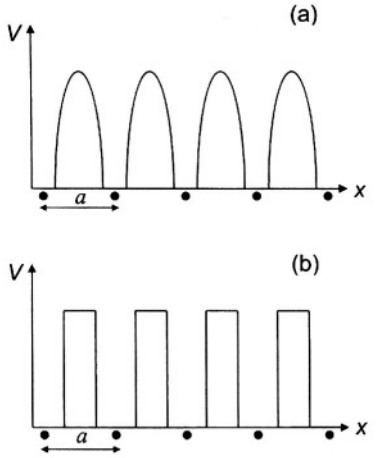

FIGURE 3.6. (a) Schematic illustration of the variation of the potential energy of electrons in a one-dimensional crystalline lattice. (b) An approximation to the realistic potential energy, as depicted in (a), according to the Kronig–Penney model.

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}E\psi = 0 \tag{3.3.27}$$

In region II, for which $-b < x < 0$, i.e., in the potential barrier, $V = V_0$, and the Schrödinger equation can be written as

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}(E - V_0)\psi = 0 \tag{3.3.28}$$

A simultaneous solution of the above equations is in terms of the periodic Bloch functions $\psi(x) = u(x)\exp(ikx)$, where $u(x)$ is no longer a constant amplitude, but it changes periodically with the periodicity of the lattice. (The mathematical solution for this one-dimensional case is based largely on differentiating the Bloch function twice with respect to x, inserting the result into the above Schrödinger equations, and determining the constants of the solution

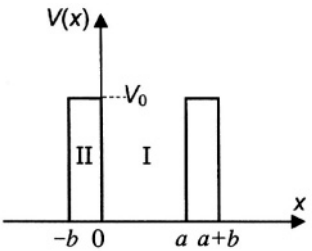

FIGURE 3.7. One-dimensional periodic potential according to Kronig–Penney model.

to these equations by employing the appropriate boundary conditions.) In this case, the condition for solution (with matching boundary conditions) is in terms of the following equation

$$\frac{P \sin \alpha a}{\alpha a} + \cos \alpha a = \cos ka \tag{3.3.29}$$

which relates the energy E (through α) to k. In this equation,

$$\alpha = \frac{(2m_e E)^{1/2}}{\hbar} \tag{3.3.30}$$

and P, which is a measure of the potential barrier strength, is

$$P = \frac{m_e V_0 b a}{\hbar^2} \tag{3.3.31}$$

Thus, the application of the boundary conditions in this case results in an equation (relating E with k) with trigonometric functions, implying that only certain values of α, and hence, E are allowed. This is best visualized by plotting $(P \sin \alpha a)/\alpha a + \cos \alpha a$ as a function of αa, as shown in Fig. 3.8.

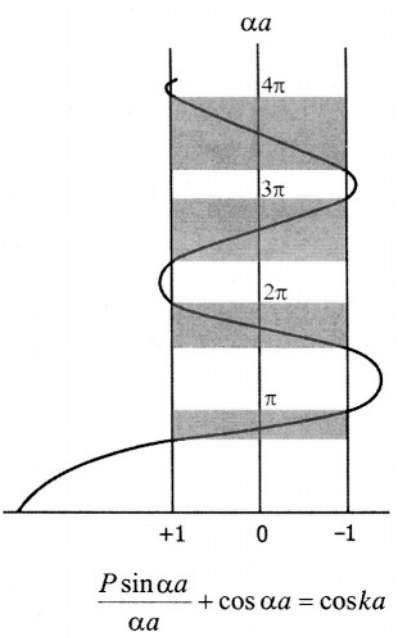

FIGURE 3.8. Plot of $\frac{P \sin \alpha a}{\alpha a} + \cos \alpha a = \cos ka$ as a function of αa.

The cos ka is only defined within the limits between +1 and –1, indicating that, since α is related to E, the solution for energies should also be between those limits. This essentially implies that the electrons, moving in a periodically varying potential field, may possess energies within certain energy bands only, i.e., within shaded regions in Fig. 3.8. These allowed energy bands are separated by ranges of αa corresponding to cos ka being either greater than +1 or less than –1, i.e., corresponding to forbidden energy ranges. Thus, to summarize the main observations, (i) there are allowed and forbidden bands of energy for electrons moving through a periodic potential, (ii) the size of these bands varies as a function of P, i.e., $V_0 b$, and (iii) with increasing αa, i.e., with increasing E, the forbidden bands become narrower. Also, for larger P, the curve is steeper, which results in narrower allowed bands and wider forbidden bands. It is important to note that at the boundary of an allowed band $\cos ka = \pm 1$, and hence $k = n\pi/a$, indicating the discontinuities in energy occurring at these values of k. These conditions also correspond to the Bragg reflection rule, suggesting that the electron states with $k = n\pi/a$ can be described as standing waves, i.e., these electrons cannot propagate through the lattice, indicating the presence of the energy gap for that given k.

Two extreme cases for the energy variation as a function of wave vector are (i) the case of free electrons (for vanishing small product of $V_0 b$) and (ii) the case of bound electrons in a potential well (for $P \to \infty$). In the former case ($P \to 0$), $\cos \alpha a = \cos ka$, and hence, using Eq. (3.3.30)

$$E = \frac{\hbar^2 k^2}{2m_e} \tag{3.3.32}$$

i.e., the case of free electrons. In the case of $P \to \infty$, $\sin \alpha a = 0$, which is possible for $\alpha a = n\pi$, or $\alpha^2 = n^2\pi^2/a^2$ ($n = 1, 2, 3, \ldots$). Using Eq. (3.3.30) ($\alpha^2 = 2m_e E/\hbar^2$), we can write

$$E_n = \frac{\hbar^2 \pi^2}{2m_e a^2} n^2 \quad (n = 1, 2, 3, \cdots) \tag{3.3.33}$$

i.e., the case of electrons bound between two infinitely high potential barriers (for sufficiently thick barrier b, so that electrons with energies less than the barrier height V_0 cannot tunnel from one atomic site to the next). The analysis of these two limits indicates that by varying P from 0 to ∞, one can obtain various cases from free electrons to the completely bound electrons. For intermediate values of P, energy bands (i.e., ranges of allowed energies) are formed and these are separated by forbidden gaps (or energy gaps).

3.4. ENERGY BANDS IN CRYSTALS

In general, there are two ways to describe the physical properties of solids. One requires no need of invoking the periodic potential; instead, the description is based on the chemical bonds in the material. The other method involves the description

of the properties of electrons in the long-range periodic potential. The existence of the energy gaps, and of other related properties, can be considered on the basis of the energy required to remove an electron from a chemical bond in the material, and to allow it to freely move through the material (under the applied field). A more rigorous description of the energy bands in crystals requires the derivation of the $E(k)$ relationship, which facilitates the elucidation of important characteristics of the electronic properties of semiconductors. For free electrons, this relationship is

$$E = \frac{\hbar^2 k^2}{2m_e} \tag{3.4.1}$$

where

$$k = (2m_e E/\hbar^2)^{1/2} \tag{3.4.2}$$

which for one-dimensional case gives

$$k_x = \left(2m_e/\hbar^2\right)^{1/2} E^{1/2} \tag{3.4.3}$$

which is a parabolic function (see Fig. 3.1).

From the Kronig–Penney model

$$\frac{P \sin \alpha a}{\alpha a} + \cos \alpha a = \cos ka \tag{3.4.4}$$

and for free electrons (i.e., $P = 0$),

$$\cos \alpha a = \cos ka \tag{3.4.5}$$

The cosine function is periodic in 2π, so

$$\cos \alpha a = \cos ka \equiv \cos(ka + n2\pi) \tag{3.4.6}$$

where $n = 0, \pm 1, \pm 2, \pm 3, \ldots$ and

$$\alpha a = ka + n2\pi \tag{3.4.7}$$

But $\alpha = (2m_e E/\hbar^2)^{1/2}$, and thus

$$(2m_e/\hbar^2)^{1/2} E^{1/2} = k + n2\pi/a \tag{3.4.8}$$

This indicates that the parabola is repeated periodically with $n\,2\pi/a$. In other words, the energy is a periodic function of k with the periodicity $2\pi/a$. As mentioned earlier (see Fig. 3.8), discontinuities in energy occur at the boundary of an allowed band when $\cos ka = \pm 1$, i.e., for $k = n\pi/a$ $(n = \pm 1, \pm 2, \pm 3, \ldots)$.

At these values of k, deviation from the parabolic $E(k)$ curve occurs, indicating that, in a periodic lattice, the electrons behave similar to free particles, except for $k = n\pi/a$. This is shown in Fig. 3.9, which shows the *extended zone representation* of $E(k)$ dependence. This figure also shows the *Brillouin zones,* i.e. the values of k associated with a specific energy band (in the figure, these are identified as 1st zone, 2nd zone, 3rd zone, etc.). Another, more convenient representation is the *reduced-zone representation* (see Fig. 3.10), which can be derived by folding the bands back into the 1st (Brillouin) zone. These $E(k)$ diagrams and other electronic-band-theory terminology, introduced later, are of great importance for describing various properties of semiconductors. As in the description of crystalline structures (see Chapter 2), which distinguishes between the reciprocal (i.e., **k**-space) and real lattices, the electronic energy bands can be also represented in (i) **k**-space (the reduced-zone representation) and (ii) in real space as a function of space coordinate **r** (see Fig. 3.11).

Thus, to summarize briefly, the $E(k)$ relationship according to the periodic potential model is no longer described by a parabolic function (corresponding to free electron model), but it is a function that reveals the presence of various allowed energy bands separated with energy bands that are forbidden, i.e., bands where no permitted k-states can exist. This model is often referred to as *nearly-free electron model.*

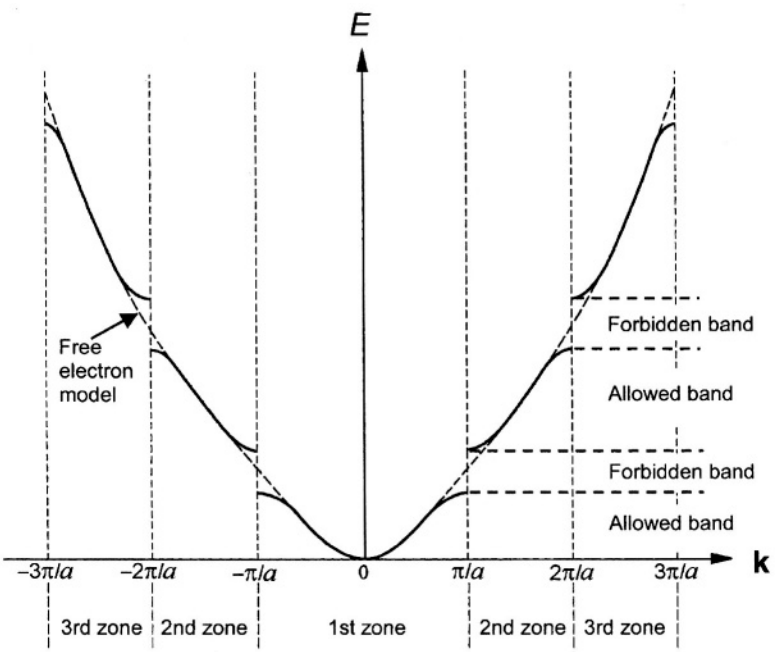

FIGURE 3.9. The extended-zone representation of the nearly-free electron model, showing the modification of the parabolic $E(k)$ dependence for free electrons at the band edges corresponding to $k = n\pi/a$. First three Brillouin zones are also indicated.

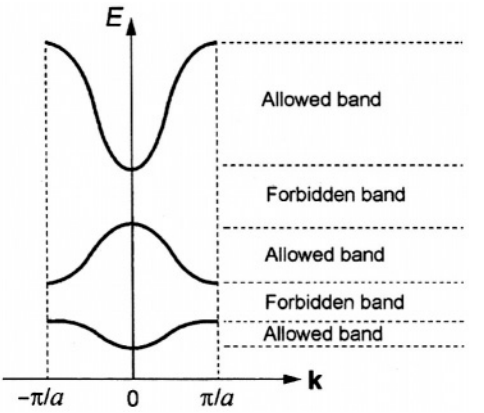

FIGURE 3.10. The reduced-zone representation.

As mentioned earlier, there is also a method that does not require the need of invoking the periodic potential, but instead, the description is based on the analysis of atomic orbitals. In this case, the electrons are considered to be tightly bound to the nuclei, and with the formation of a solid, the electronic wave functions of constituent atoms overlap. These electronic wave functions can be approximated by a linear combination of the atomic wave functions. Such a method is referred to as the *tight-binding approximation* or *linear combination of atomic orbitals* (LCAO). In this method, electron positions in molecular orbitals are approximated by a linear combination of atomic orbitals of the form $\psi = c_1\phi_1 + c_2\phi_2 + c_3\phi_3 + \cdots$, where ψ and ϕ correspond to molecular orbital

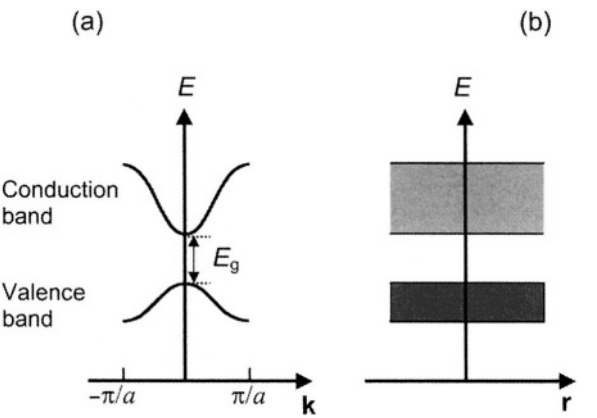

FIGURE 3.11. Schematic diagram of the electronic energy bands in (a) **k**-space (the reduced-zone representation) and (b) in real space as a function of space coordinate **r**.

wave function and atomic orbital wave function, respectively. Thus, the problem of determining the appropriate function for the molecular orbitals is basically reduced to optimizing coefficients in this linear equation. Thus, to summarize briefly, LCAO is essentially an approximation for determining a molecular orbital by using a superposition of atomic orbitals. In employing such an approximation, although the specific atomic orbitals and their number are not initially identified, inclusion of additional atomic orbitals in the linear combination results in the refinement of the approximation. (Note that the main motivation for employing approximate solutions is related to the difficulties in finding the exact solutions to the Schrödinger equation for increasingly complex systems.)

Based on this brief outline of these methods, it follows that the nearly-free electron model is more suited for the calculation of the conduction band states (since electrons in this case are nearly free), whereas the LCAO method is a suitable approximation for the calculation of the valence band states (since the wave functions of valence electrons are comparable with bonding orbitals).

3.5. BRILLOUIN ZONES AND EXAMPLES OF THE ENERGY BAND STRUCTURE FOR SEMICONDUCTORS

The information on the electronic band structure can be illustrated by employing the first Brillouin zone, or as often referred to as Brillouin zone, which is expressed for one period of the reciprocal lattice centered about the origin of the **k**-space (see Fig. 3.12). In this case, the edges of the bands are described using the boundaries of a three-dimensional figure in **k**-space. In other words, this representation provides a three-dimensional description of the band theory by

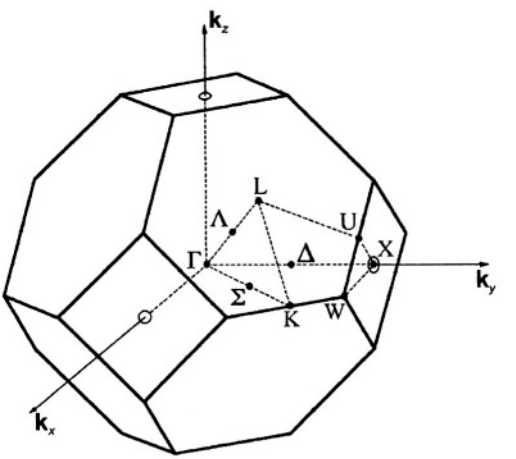

FIGURE 3.12. The first Brillouin zone of the diamond and zincblende-type structures. The important symmetry points and lines are indicated.

using Brillouin zones (i.e., bands of permitted energy). As noted previously, at $k = n\pi/a$, the discontinuities in energy occur at these values of k, and these conditions also correspond to the Bragg reflection rule, suggesting that the electron states with $k = n\pi/a$ can be described as standing waves, i.e., these electrons cannot propagate through the lattice, indicating the presence of the energy gap for that given k. Thus, this essentially also implies that at the boundaries of the Brillouin zones (i.e., $k = n\pi/a$) the electron waves experience reflection from the crystal planes according to the Bragg law, and they cannot propagate through the lattice.

The shape of the Brillouin zone is determined by the crystal lattice geometry and its size depends on the lattice constant. The surface that encloses the occupied states is referred to as the *Fermi surface.* In the description of the Brillouin zone, the standard notation of the points of symmetry inside the zone is in terms of Greek letters (see Fig. 3.12), and it is in terms of Roman letters for the surface. In such a representation, the energy band structure is described along selected crystallographic orientations. Thus, e.g., Γ represents the origin of the reciprocal space (i.e., $k = 0$); Λ represents a direction such as [111] and L denotes the zone end along that direction; Δ represents a direction such as [100] and X denotes the zone end along that direction; and Σ represents a direction such as [110] and K denotes the zone end along that direction. A convenient way of describing the band structure is in terms of $E(\mathbf{k})$ relationship plotted along selected directions in the reduced-zone representation. Such a description of the detailed band structure, derived from theoretical calculations and experimental observations, is of great importance in elucidating the optical and electrical properties of semiconductor

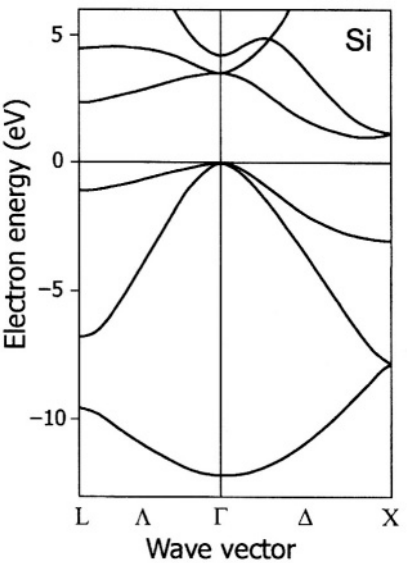

FIGURE 3.13. Electronic energy band structure of Si.

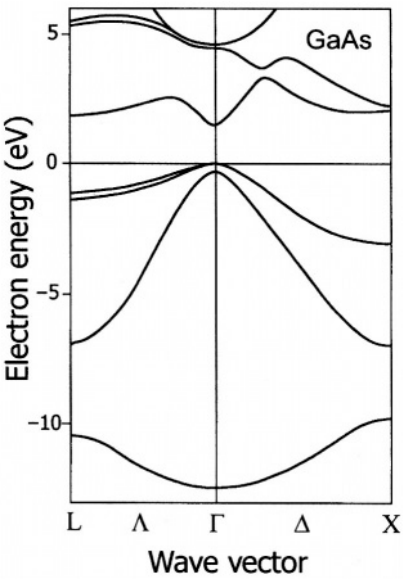

FIGURE 3.14. Electronic energy band structure of GaAs.

and for designing semiconductor devices. The band structures of two most important semiconductors (i.e., Si and GaAs) are shown in Figs. 3.13 and 3.14. Close examination of these diagrams reveals some important facts. As expected from the previous sections, there is a forbidden energy region, which is referred to as the *energy gap,* which separates the *conduction band* from the *valence band.* More specifically, the energy gap (E_g) is defined as the energy separation between the highest valence band maximum and the lowest conduction band minimum. What distinguishes these two important semiconductors (i.e., Si and GaAs) is the location of the lowest conduction band minima in relation to the highest valence band maxima. In the case of GaAs (see Fig. 3.14), both the highest valence band maximum and the lowest conduction band minimum are at the same Γ point; such a material is referred to as *direct energy-gap* semiconductor (this fact is of great importance, e.g., in optical processes such as electronic transitions between the valence and conduction bands, see Chapter 4). In the case of Si (see Fig. 3.13), the highest valence band maximum is at Γ point, but the lowest conduction band minimum is in the Δ or [100] direction near the first Brillouin zone boundary (i.e., X point). In such a case, when the highest valence band maximum and the lowest conduction band minimum are not at the same point in $\mathbf{k}$-space, a semiconductor is referred to as *indirect energy-gap* material (again, this fact is of great importance in optical processes, see Chapter 4). An additional important feature observed in the band structure of GaAs is the presence of the conduction band minima at L and X points, which are referred to as L-valley and X-valley and which have a major effect on high-field transport properties.

3.6. THE EFFECTIVE MASS

The concept of effective mass is an important consideration for the description of the dynamics of the movement of electrons in a crystal. In semiconductors, the presence of the periodic crystal potential modifies the electron properties, resulting in the electron mass that is different from the free electron mass. The *effective mass* (denoted as m^*) is typically referred to the experimentally determined value of the electron mass, and it is usually given in terms of the free electron mass, i.e., m^*/m_0. For different semiconductors, this ratio can be slightly greater or less than unity. In order to derive an expression for the effective mass in terms of the electronic band structure parameters, we recall that, for a wave packet, the group velocity ($v_g = d\omega/dk$) can be expressed as (note that $\omega = E/\hbar$)

$$v_g = \frac{1}{\hbar}\frac{dE}{dk} \tag{3.6.1}$$

and the acceleration a can be expressed as

$$a = \frac{dv_g}{dt} = \frac{1}{\hbar}\frac{d^2E}{dk^2}\frac{dk}{dt} \tag{3.6.2}$$

dE/dk is known, and dk/dt can be evaluated from the expression $p = \hbar k$

$$\frac{dp}{dt} = \hbar\frac{dk}{dt} \tag{3.6.3}$$

Thus,

$$a = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}\frac{dp}{dt} = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}F \tag{3.6.4}$$

where F is the force acting on an electron. By comparing Eq. (3.6.4) with an equation for acceleration, i.e., $a = F/m$, we can write

$$\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{d^2E}{dk^2} \tag{3.6.5}$$

or

$$m^* = \hbar^2\left(\frac{d^2E}{dk^2}\right)^{-1} \tag{3.6.6}$$

In other words, the electron effective mass is inversely proportional to the curvature of an electron band $E(k)$. This implies that the carrier effective mass can

be determined from the electronic band structure, or $E(k)$ relationship. For example, the examination of the extended zone representation of $E(k)$ diagram (see Fig. 3.9) indicates that (i) the effective mass is positive near the bottom of the bands, and it is negative near the top of the bands and (ii) the effective mass near both the top and bottom of an energy band is energy-independent, i.e., constant. Electrons having the negative mass near the top of the band are referred to as an "electron hole". However, in the presence of an electric field, a negative mass with a negative charge is equivalent to a positive mass with a positive charge (i.e., $-ma = -e\mathcal{E}$ is equivalent to $ma = e\mathcal{E}$), and thus, the *holes* are described as carriers having a positive mass and a positive charge. This implies that in the presence of an electric field, electrons at the bottom of the conduction band and holes at the top of the valence band travel in the opposite directions in real space. As it is discussed in the subsequent chapters, the electrical transport in a nearly filled band (i.e., valence band) in terms of the motion of holes is of great importance in describing semiconductor properties.

The inverse dependence of the effective mass on the $E(k)$ curvature indicates that the greater curvature implies smaller effective mass, and vice versa. In addition, since properties of the material depend on crystallographic directions (i.e., anisotropy), the effective mass may differ in each direction. Considering the case of GaAs (see Fig. 3.14), effective masses for electrons in the conduction band differ, e.g., between the central Γ-valley (with greater curvature) and L-valley (with smaller curvature). These masses are $0.067m_0$ and $0.35m_0$ for Γ-valley and L-valley, respectively. The energy at the lowest minimum at L is only about 0.3 eV higher than that at Γ. This difference in the effective mass has an important influence on transport properties of the semiconductors at sufficiently high fields

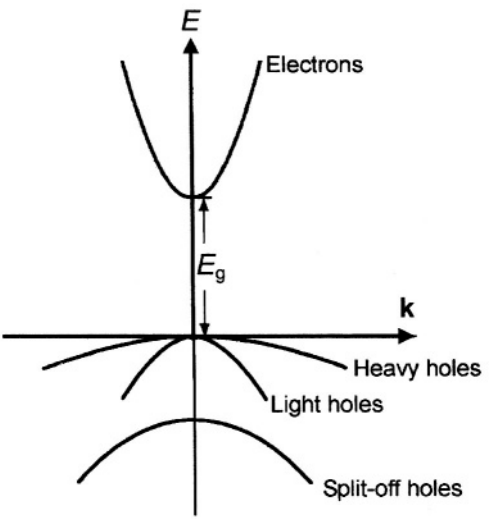

FIGURE 3.15. $E(k)$ diagram illustrating various hole masses, i.e., light holes, heavy holes, and split-off holes.

TABLE 3.1. Effective masses of electrons, m_e^*, holes (light, m_{lh}^*, and heavy, m_{hh}^*) and split-off holes m_{soh}^* in units of m_0 (i.e., the free electron rest mass) for selected direct energy-gap semiconductors

	m_e^*/m_0	m_{lh}^*/m_0	m_{hh}^*/m_0	m_{soh}^*/m_0
InAs	0.026	0.025	0.41	0.08
InP	0.073	0.078	0.40	0.15
GaAs	0.067	0.082	0.50	0.15

(greater than about 3 kV cm^{-1}), since some of the conduction electrons can be transferred into the L-valley with its greater effective mass and can subsequently affect the carrier mobility and result in reduced drift velocity. In the valence band (see Fig. 3.15), similar considerations of the inverse dependence between the effective mass and energy band curvature indicate that there are the *light-hole* band (larger curvature) and *heavy-hole* band (smaller curvature). In addition, in the valence band, the *split-off* band at a slightly lower energy (due to spin–orbit interactions) is also present (Fig. 3.15). Tables 3.1 and 3.2 list the effective masses of both the electrons and holes in selected direct and indirect energy-gap semiconductors. In Chapter 4, the concept of the density of states in the conduction and valence bands will be introduced. The calculations of these quantities involve the appropriate choice of the effective mass, since the band-structure representation of a given band typically invokes different effective masses; thus, this requires employing a combination of the effective masses. Since holes typically occupy the light-hole and heavy-hole bands only, for calculations of the density of states in the valence band, the effective mass of the valence band is expressed as

$$m_{hd}^* = \left[(m_{lh}^*)^{3/2} + (m_{hh}^*)^{3/2}\right]^{2/3} \tag{3.6.7}$$

For the conduction band, the effective mass for density of states calculations is

$$m_{ed}^* = N_{bm}^{2/3}(m_x^* m_y^* m_z^*)^{1/3} \tag{3.6.8}$$

where N_{bm} is the number of equivalent band minima in the conduction band, and m_x^*, m_y^*, and m_z^* are the effective masses along the principal axes of the ellipsoidal energy surface. (Note that in semiconductors, such as Si, constant-energy surfaces

TABLE 3.2. Effective masses of electrons (longitudinal, m_l^*, and transverse, m_t^*), holes (light, m_{lh}^*, and heavy, m_{hh}^*) and split-off holes m_{soh}^* in units of m_0 (i.e., the free electron rest mass) for indirect energy-gap semiconductors Ge and Si

	m_l^*/m_0	m_t^*/m_0	m_{lh}^*/m_0	m_{hh}^*/m_0	m_{soh}^*/m_0
Ge	1.64	0.082	0.04	0.35	0.08
Si	0.98	0.19	0.16	0.54	0.23

in the conduction band are ellipsoids, and thus, the effective masses are characterized by the curvature along the long axis as longitudinal m_l^* and transverse m_t^* masses.) Thus, for Si, for calculations of the density of states in the conduction band, the effective mass of the electron is expressed as

$$m_{ed}^* = 6^{2/3}(m_l^* m_t^{*2})^{1/3} \tag{3.6.9}$$

The effective mass for conductivity (or mobility) calculations is derived from their inverse relationship, i.e., conductivity is proportional to the sum of the inverse of the individual masses, and thus the conductivity effective mass is

$$m_{ec}^* = 3\left(\frac{1}{m_l} + \frac{2}{m_t}\right)^{-1} \tag{3.6.10}$$

One of the methods that is often used for measuring effective mass is the *cyclotron resonance* (i.e., resonant absorption of electromagnetic waves) experiment. In this case, an interaction of electromagnetic waves with charge carriers results in the resonant absorption of the electromagnetic wave in the presence of an applied magnetic field that causes the carrier to vibrate at the same frequency as the frequency of the applied electric field. The effective mass can be derived using equation $m^* = eB/\omega_c$, where B is the magnetic field corresponding to the maximum absorption of the electromagnetic wave having frequency equal to cyclotron frequency ω_c.

3.7. CLASSIFICATION OF SOLIDS ACCORDING TO THE BAND THEORY

The electronic band structure of solids can explain the distinction between metals, semiconductors, and insulators (see Fig. 3.16). First, it should be emphasized that empty energy bands, which do not contain electrons, do not contribute to the electrical conductivity of a material; whereas, although completely filled bands contain electrons, they do not contribute to the electrical conductivity, since the carriers are unable to gain energy (when an electric field is applied) due to the fact that all the energy levels are occupied. In comparison, the partially filled bands contain both the electrons and unoccupied energy levels at higher energies. The latter allow carriers to gain energy in the presence of an applied electric field, and thus, carriers in a partially filled band can contribute to the electrical conductivity of the material. In the case of metals, the electronic band structure results in incomplete filling of the highest occupied energy band. (Note that the metallic behavior may also be a result of an energy overlap between filled and empty bands, not shown in the figure.) The magnitude of the energy gap separating the highest filled (or valence) band and the lowest empty (or conduction) band distinguishes a semiconductor from an insulator. In the case of an intrinsic semiconductor, the highest filled (valence) band is separated

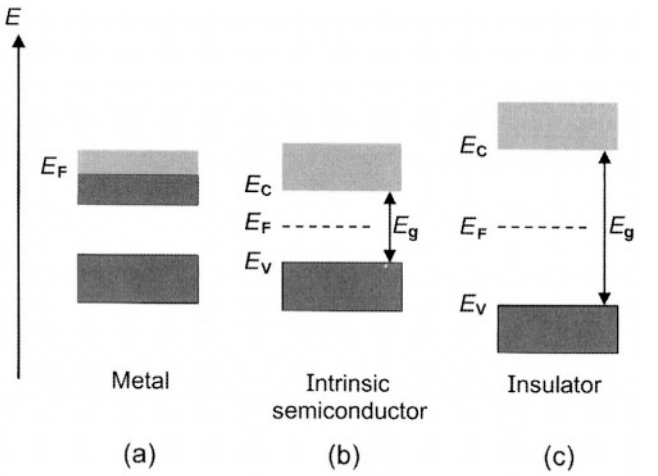

FIGURE 3.16. Schematic illustration of typical band diagrams for (a) a metal, (b) an intrinsic semiconductor ($T = 0$ K), and (c) an insulator. Dashed lines represent the Fermi level (for details, see Chapter 4). Darker regions represent filled bands, whereas lighter regions correspond to the empty bands.

from the lowest empty (conduction) band by a relatively narrow forbidden energy gap, and at $T = 0$ K there are no electrons in the conduction band. However, in semiconductors, the energy gap (E_g) is sufficiently small, so that at room temperature the electrons from the top of the valence band are thermally excited to the conduction band, where they can contribute to the carrier transport in a material. In insulators, the energy gap is so much greater (as compared to semiconductors) that at room temperature the probability of thermal excitation of an electron from the valence band to the conduction band is very low. In Fig. 3.16, the *Fermi energy,* or *Fermi level* (E_F) defines the reference energy for the probability of occupation of electron states (for more details, see Chapter 4). Thus, in metals E_F is located within a partially filled allowed band, as shown in Fig. 3.16, whereas in semiconductors and insulators E_F is positioned within the forbidden band.

At this juncture, it is also important to define an additional parameter that is useful in the description of semiconductors (see Fig. 3.17). This is the *electron affinity* ($e\chi$), defined as the energy difference between the vacuum level (i.e., the energy of a free electron) and the bottom of the conduction band. Another important parameter is the *work function* (often denoted as $e\Phi$), which is the energy difference between the vacuum level and the Fermi level. It should be noted that the electron affinity ($e\chi$) is a constant for a given semiconductor, whereas the work function $e\Phi$ depends on the *doping,* which affects the Fermi level position. (Doping is the process of putting impurities into the lattice to produce materials and devices with desired properties, see Chapter 4.)

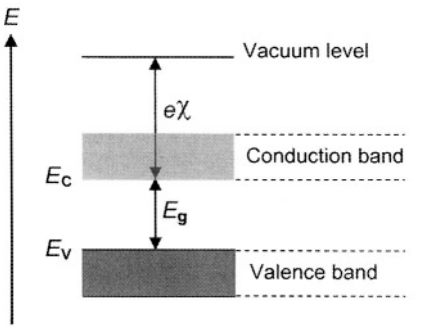

FIGURE 3.17. Schematic illustration of a typical band diagram indicating some important definitions for the description of a semiconductor; this includes the electron affinity ($e\chi$), i.e., the energy difference between the vacuum level and the bottom of the conduction band.

3.8. SUMMARY

The electronic properties in semiconductors can be described by the quantum theory of the electronic energy band structure of solids. In one of the approaches for the description of electronic band structure, with the formation of a solid, the electronic wave functions of constituent atoms overlap, and the application of the Pauli exclusion principle leads to the splitting of the discrete energy levels of the isolated atoms into bands of allowed electron levels separated by forbidden energy gaps.

According to the *Kronig–Penney model* (i.e., the electron in a periodic crystal potential), the electrons, moving in a periodically varying potential field, may possess energies within certain energy bands only, i.e., resulting in the presence of allowed and forbidden bands of energy for electrons moving through a periodic potential. The presence of the periodic crystal potential also results in the electron mass (i.e., the *effective mass*) that is different from the free electron mass. The electron effective mass is inversely proportional to the curvature of an electron band $E(k)$. Electrons having the negative mass near the top of the band are referred to as an "electron hole," or equivalently as the *holes,* i.e., carriers having a positive mass and a positive charge. The inverse dependence of the effective mass on the energy band curvature indicates that there are light-hole band (larger curvature) and heavy-hole band (smaller curvature) in the valence band. Also, since properties of the material depend on crystallographic directions (i.e., anisotropy), the effective mass may differ in each direction.

The electronic properties of a solid are determined by the electron occupation of the highest energy bands, i.e., the valence and conduction bands, which are separated by a *fundamental energy gap,* E_g. Thus, the quantum-mechanical description of the electronic band structure provides a clear distinction between conductors, semiconductors and insulators. In conductors, the electronic band structure results in incomplete filling of the highest occupied energy band,

whereas in semiconductor and insulators, the energy gap separates the highest filled (or valence) band and the lowest empty (or conduction) band. The energy gap in the case of a semiconductor is relatively narrow, so that at room temperature the electrons from the top of the valence band are thermally excited to the conduction band, where they can contribute to the carrier transport in a material. In insulators, the energy gap is so much greater that at room temperature the probability of thermal excitation of an electron from the valence band to the conduction band is very low.

PROBLEMS

3.1. Consider electrons in a cathode-ray tube monitor moving from the cathode (electron gun) to the phosphor screen with the speed of about 7×10^7 m s^{-1}. Is it necessary to use quantum mechanical description for motion of such electrons?

3.2. For a one-dimensional box, $\psi_n(x) = A \sin \frac{n\pi x}{L}$. Show that the normalization constant $A = (2/L)^{1/2}$.

3.3. Derive equation $\frac{P \sin \alpha a}{\alpha a} + \cos \alpha a = \cos ka$, according to the Kronig–Penney model.

4

Basic Properties of Semiconductors

4.1. INTRODUCTION

In general, the main factors that determine basic properties (e.g., optical and electrical properties) of semiconductors are related to the chemical composition and the (crystallographic) structure, the presence of various defects and impurities (both intentional and unintentional), and the dimensions of the semiconductor or semiconductor structure (i.e., related to the quantum confinement regime). The chemical composition and the (crystallographic) structure determine the electronic band structure (e.g., the magnitude and the type of energy gap, and the carrier effective mass), which has the major influence on the semiconductor properties. The presence of various defects and impurities results in the introduction (in the energy gap of the semiconductor) of various electronic states (both shallow and deep) that affect strongly the optical and electrical properties. Finally, as will be discussed in the subsequent chapters, for the semiconductor dimensions commensurate with the de Broglie wavelength of carriers (i.e., of the order of 10 nm), quantum size effects dominate the semiconductor properties.

As mentioned in Chapter 1, the electrical conductivity of semiconductors can be varied (in both sign and magnitude) widely as a function of (i) impurity content (e.g., doping), (ii) temperature (i.e., thermal excitation), (iii) optical excitation (i.e., excitation with photons having energies greater than the energy gap E_g), and (iv) excess charge carrier injection (e.g., in semiconductor devices). It is this capability of controlling the electrical conductivity in semiconductors that offers myriad applications of these materials in a wide variety of electronic and optoelectronic devices.

4.2. ELECTRONS AND HOLES IN SEMICONDUCTORS

In a semiconductor containing no impurities or defects, i.e., an *intrinsic semiconductor,* at temperatures above 0 K some thermally excited electrons are

promoted from the valence band to the conduction band. An unoccupied state in the valence band is called a *hole,* which may be regarded as a positive charge carrier that can contribute to the conduction process (see Section 3.6). Electronic transitions across the energy gap to the conduction band result in a spontaneous generation of holes in the valence band, and the generated carriers are described as *electron–hole pairs.* After a random motion through the lattice, the electron in the conduction band encounters a hole and undergoes a recombination transition. The generation of electron–hole pairs and their subsequent recombination is a continuous process, and the average time that carriers exist between generation and recombination is called the *lifetime* of the carrier. During this process of generation of electron–hole pairs, the concentration of electrons (denoted as n) in the conduction band is equal to the concentration of holes (denoted as p) in the valence band. This can be expressed as

$$n = p = n_i \tag{4.2.1}$$

where n_i is the intrinsic carrier concentration.

Electrons and holes in the conduction and valence bands, respectively (carrying negative and positive electronic charges, respectively), are referred to as *free charge carriers.* In the presence of an electric field $\mathcal{E}$, the free charge carriers attain the *drift velocity,* $\mathbf{v}$, and a net current density, $\mathbf{J}$. (Note that the electron and hole will have the drift velocities in opposite directions.) In the case of Ohm's law (i.e., $\mathbf{J} = \sigma\mathcal{E}$, where σ is the *conductivity*), the drift velocity is proportional to the applied electric field (i.e., for electrons $\mathbf{v}_n = -\mu_e\mathcal{E}$), and the proportionality constant μ_e is referred to as the *carrier mobility* of electrons (the mobility is a measure of the frequency of scattering events and is related to the *scattering relaxation time* τ as $\mu = e\tau/m^*$); note that in this case the direction of drift velocity is opposite to the direction of the electric field, since the charge carriers, i.e., electrons, are negatively charged. The current density for electrons and holes can be expressed as $\mathbf{J}_n = -ne\mathbf{v}_n$ and $\mathbf{J}_p = pe\mathbf{v}_p$, respectively. Thus, the current densities for electrons and holes are in the same direction, since their corresponding drift velocities are in the opposite directions. The conductivity σ for electrons and holes can be written as $\sigma_n = ne\mu_e$ and $\sigma_p = pe\mu_h$ (where $e = 1.602 \times 10^{-19}$ C is the electron charge). Since both electrons and holes contribute to the current in intrinsic semiconductors (i.e., $n=p$), the bulk conductivity σ must in principle be expressed as

$$\sigma = ne\mu_e + pe\mu_h \tag{4.2.2}$$

where μ_e and μ_h are the mobilities of electrons and holes, respectively. However, the electrons have the major contribution to the current, since typically $\mu_e > \mu_h$ (for values of μ_e and μ_h see Table B1 in Appendix B); this is due to the inverse relationship between the carrier mobility and the carrier effective mass.

The carrier mobility is determined by random scattering processes, among which the main mechanisms in semiconductors include *impurity scattering* and the intrinsic *phonon* (*lattice*) *scattering.* Impurity scattering sources include both the ionized and neutral donor and acceptor atoms that are used to dope the

semiconductor to the desired conductivity level. In general, impurity scattering may also arise due to unintentionally introduced impurities and native defects and their complexes. The ionized donor and/or acceptor scattering is due to the Coulomb attraction or repulsion between the charge carriers and these ionized impurities. The scattering by neutral impurity atoms becomes considerable at sufficiently low temperatures at which a substantial number of the impurity atoms becomes neutral. The intrinsic scattering mechanisms may also include, besides phonon scattering, carrier–carrier scattering and inter-valley scattering. Typically, carrier–carrier scattering may have a significant contribution only in heavily doped semiconductors under degenerate conditions or under high-field conditions, and inter-valley scattering may occur in materials having several minima in the conduction band. The phonon (lattice) scattering mechanism usually involves two types of phonon scatterings, i.e., acoustical and optical phonon scattering mechanisms. In general, ionized impurity scattering (μ_I) and phonon scattering (μ_{ph}) are dominant electron scattering mechanisms in practical semiconductors. (In such a case, the combined mobility can be expressed as $1/\mu = 1/\mu_I + 1/\mu_{ph}$.) It should be noted that higher doping concentrations result in increased scattering (and lower mobilities) of the carriers by the ionized impurities.

The temperature dependence of the mobility, as described by theoretical considerations, reveals opposite temperature dependencies for the ionized impurity scattering and the phonon scattering. These are, for ionized impurity scattering, $\mu_I \propto T^{3/2}/N_I$ (where N_I is the total impurity concentration, i.e., $N_d + N_a$), and for phonon scattering, $\mu_{ph} \propto T^{-3/2}$. At lower temperature regime, the carrier velocity increases with increasing temperature, resulting in the reduced time that the carriers spend in the vicinity of the scattering centers. Thus, the mobility increases with temperature. However, with further increase in temperature, the lattice vibrations become dominant, resulting in increasing probability of scattering by the lattice and in decreasing mobility. Thus, to summarize briefly, because of the opposite temperature dependences for the ionized impurity scattering and the lattice scattering, the mobility reaches a maximum at a specific temperature, which depends on the concentration of ionized impurities (i.e., with increasing impurity concentrations, the position of the maximum is shifted to increasingly higher temperatures).

4.3. THE FERMI–DIRAC DISTRIBUTION FUNCTION AND THE DENSITY OF STATES

One of the most important objectives in describing a semiconductor in relation to its electrical and optical properties is to determine both the carrier concentrations and the energy distributions. These require knowledge of (i) the *probability of carrier occupancy* of a state at energy E and (ii) the density of available states, or *density of states* (DOS). The occupation statistics of energy levels by carriers is described by the *Fermi–Dirac occupation statistics.* In other words, the carrier energy distribution is described by the Fermi–Dirac statistics, and, specifically, the probability of an electron occupying a state at energy E is

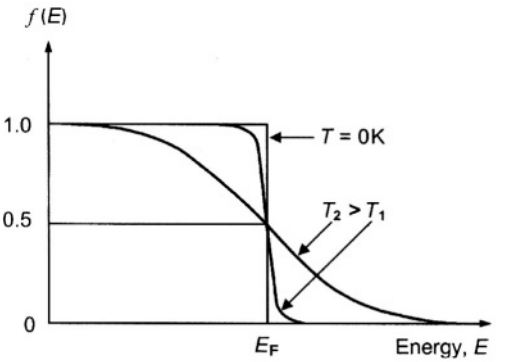

FIGURE 4.1. The Fermi–Dirac distribution function $f(E)$ vs. energy at different temperatures.

described by the *Fermi–Dirac distribution function* $f(E)$, which for electrons is (see Fig. 4.1)

$$f_n(E) = \frac{1}{\exp[(E - E_F)/k_B T] + 1} \tag{4.3.1}$$

where the energy E_F is called the *Fermi energy* (also referred to as *Fermi level*), which is defined as the energy at which $f(E) = 1/2$, i.e., the energy for which the probability of occupation is 1/2.

Correspondingly, the hole distribution function is

$$f_p(E) = 1 - f_n(E) = \frac{1}{\exp[(E_F - E)/k_B T] + 1} \tag{4.3.2}$$

In a semiconductor, the mobile electrons are those occupying the energy state E greater than E_c. In the analysis and description of various processes in semiconductors, it is essential to calculate the electron concentration in the conduction band. In this case, the total concentration of electrons is proportional to (i) the number of states per unit volume and per unit energy [i.e., the DOS, $g(E)$], over the energy interval between E and $E + dE$ and (ii) the probability of an electron occupying a state at energy E, *i.e.*, $f(E)$. The total number of electrons in the conduction band can be determined by integrating this function over energy from the bottom energy E_c to the top energy in the conduction band. The difficulty involved in evaluating the integral in such a case is due to the fact that it would require introducing an additional parameter related to the width of the conduction band. However, since the Fermi level is much below the top of the conduction band, the upper limit for the integration can be replaced by infinity, and thus, the total number of electrons in the conduction band is

$$n = \int_{E_c}^{\infty} g_n(E) f_n(E) dE \tag{4.3.3}$$

Similarly, for holes in the valence band, the total number of holes can be expressed as

$$p = \int_{-\infty}^{E_v} g_p(E) f_p(E) dE \tag{4.3.4}$$

Thus, the evaluation of the total number of electrons and holes in their respective bands requires knowledge of the DOS $g(E)$. From the analysis of the case of an electron in a three-dimensional box (e.g., a cube of side L), described in Chapter 3, the expression for its energy is

$$E_n = \frac{\hbar^2 \pi^2}{2m_e L^2}(n_x^2 + n_y^2 + n_z^2) \tag{4.3.5}$$

where the point (n_x, n_y, n_z) in the three-dimensional (k_x, k_y, k_z) space corresponds to a quantum energy state. (Note that the consideration of the electron spin actually implies two quantum states.) Thus, by counting the number of (n_x, n_y, n_z) states per unit volume over the energy range between E and $E+dE$, the expressions for the DOS can be determined. The volume of a unit cell in **k**-space, occupied by one state with a specific k, is $(2\pi/L)^3$, or $(2\pi)^3/V$. In order to evaluate the number of electronic states (dN) over the range between k and $k + dk$, the spherical volume between k and dk, i.e., $4\pi k^2 dk$, is divided by $(2\pi)^3/V$. (The result should also be multiplied by a factor of two in order to account for the fact that each state with a specific value of k can be occupied by two electrons with opposite spins.) Thus, the expression for dN can be written as

$$dN = V \frac{k^2 dk}{\pi^2} \tag{4.3.6}$$

For the parabolic bands, we can write $k = [2m_e^*(E - E_c)/\hbar^2]^{1/2}$, and $kdk = m_e^* dE/\hbar^2$, and thus, the DOS can be written as

$$g_n(E) = \frac{1}{V}\frac{dN}{dE} = 4\pi(2m_e^*/h^2)^{3/2}(E - E_c)^{1/2} \tag{4.3.7}$$

Thus, the expressions for the DOS for the conduction and valence bands, respectively, are

$$g_n(E) = 4\pi(2m_e^*/h^2)^{3/2}(E - E_c)^{1/2} \tag{4.3.8}$$

$$g_p(E) = 4\pi(2m_h^*/h^2)^{3/2}(E_v - E)^{1/2} \tag{4.3.9}$$

The Fermi–Dirac function can be expressed in a more simplified form; since $k_B T$ at room temperature is about 0.026 eV, $E - E_F \gg k_B T$, and thus, for sufficiently large energies, the Fermi–Dirac function (see Eq. 4.3.1) is reduced to

the classical *Maxwell–Boltzmann function,* i.e., Eq. (4.3.1) can be approximated by

$$f_n(E) = \frac{1}{\exp[(E - E_F)/k_B T] + 1} \cong \exp[-(E - E_F)/k_B T] \tag{4.3.10}$$

By substituting Eq. (4.3.8) into Eq. (4.3.3), the electron concentration in the conduction band can be expressed as

$$n = \int_{E_c}^{\infty} g_n(E) f_n(E) dE = 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2} F_{1/2}(\eta_n) \tag{4.3.11}$$

where $F_{1/2}(\eta_n)$ is the Fermi integral:

$$F_{1/2}(\eta_n) = \frac{2}{\pi^{1/2}} \int_0^{\infty} \frac{x^{1/2} dx}{1 + \exp(x - \eta_n)} \tag{4.3.12}$$

and

$$\eta_n = (E_F - E_c)/k_B T \tag{4.3.13}$$

The term

$$N_c = 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2} \tag{4.3.14}$$

is referred to as the *effective DOS* for the conduction band. Note that in this expression, for materials such as Si (with several equivalent minima with anisotropic effective masses), the effective mass for DOS calculations (i.e., m_{ed}^*) should be used (see Eq. 3.6.9, Chapter 3).

Analogously, by substituting Eq. (4.3.9) into Eq. (4.3.4), the hole concentration in the valence band can be expressed as

$$p = \int_{-\infty}^{E_v} g_p(E) f_p(E) dE = 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2} F_{1/2}(\eta_p) \tag{4.3.15}$$

where

$$\eta_p = (E_v - E_F)/k_B T \tag{4.3.16}$$

and

$$N_v = 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2} \tag{4.3.17}$$

TABLE 4.1. Effective DOS in the conduction band (N_c) and valence band (N_v) for selected semiconductors (at 300 K)

	Ge	Si	GaAs
N_c (cm^{-3})	1.0×10^{19}	2.8×10^{19}	4.7×10^{17}
N_v (cm^{-3})	6.0×10^{18}	1.0×10^{19}	7.0×10^{18}

is the effective DOS for the valence band. Note again that in this expression, the effective mass for DOS calculations should be used (see Eq. 3.6.7, Chapter 3). Effective DOS in the conduction band (N_c) and valence band (N_v) for selected semiconductors at 300 K are listed in Table 4.1.

For $\eta \leqslant -3$, $F_{1/2}(\eta)$ is approximated by $\exp(\eta)$, and thus, in this case, the concentration of electrons in the conduction band can be expressed as [for $E_c - E_F \geqslant 3k_BT(\eta_n \leqslant -3)$]

$$n = N_c \exp[(E_F - E_c)/k_BT] \tag{4.3.18}$$

and the concentration of holes in the valence band can be expressed as [for $E_F - E_v \geqslant 3k_BT(\eta_p \leqslant -3)$]

$$p = N_v \exp[(E_v - E_F)/k_BT] \tag{4.3.19}$$

or by substituting for N_c and N_v, using Eqs. (4.3.14) and (4.3.17)

$$n = 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2} \exp[(E_F - E_c)/k_BT] \tag{4.3.20}$$

$$p = 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2} \exp[(E_v - E_F)/k_BT] \tag{4.3.21}$$

To summarize briefly, from the known density of available states and the probability of occupation of these states (i.e., the Fermi–Dirac distribution function), the resulting distribution of carriers as a function of energy for an intrinsic semiconductor can be calculated. These results are shown in Fig. 4.2. In the case of the conduction band, it can be inferred from this figure that although the DOS increases with energy, the Fermi–Dirac distribution function becomes very small for higher energies; thus, the resulting distribution of carriers as a function of energy, i.e., the product $g_n(E)f(E)$, decreases rapidly above the conduction band edge. Similar considerations apply to the valence band case. (Note that in this case increasing energy is in opposite direction to that of electron energy.)

It is important at this juncture to distinguish between the *degenerate* and *nondegenerate* semiconductors. The conditions of $E_c - E_F \geqslant 3k_BT$ and

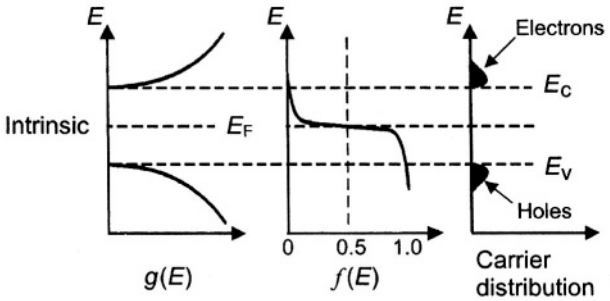

FIGURE 4.2. Schematic diagram of the density of available states, the Fermi–Dirac distribution function (i.e., probability of occupation of states), and the resulting distribution of carriers as a function of energy for an intrinsic semiconductor. The carrier concentration for electrons is a product of $g_n(E)f(E)$, and for holes it is $g_p(E)[1-f(E)]$.

$E_F - E_v \geq 3k_BT$ used earlier imply that the Fermi level position in the energy gap must be greater than $3k_BT$ from the band edges (either conduction or valence); for such a case, the semiconductor is referred to as nondegenerate. For the case of the Fermi level being positioned within $3k_BT$ of the band edges, or positioned inside either band, the semiconductor is referred to as degenerate.

4.4. INTRINSIC AND EXTRINSIC SEMICONDUCTORS

4.4.1. Intrinsic Semiconductors

As mentioned earlier, intrinsic semiconductors are those that do not contain impurities or defects. In such a case, the thermal activation of an electron from the valence band to the conduction band produces (i) a free electron in the conduction band and (ii) a free hole in the valence band, and the densities of electrons and holes are equal.

Typically, intrinsic semiconductors are nondegenerate, thus the equations for carrier concentrations, derived in the previous section, can be employed in this case. Thus, we can write

$$n_i = N_c \exp[(E_i - E_c)/k_BT] \tag{4.4.1}$$

$$n_i = N_v \exp[(E_v - E_i)/k_BT] \tag{4.4.2}$$

where, in this case, $E_F = E_i$. Using these equations, the effective densities for the conduction and valence bands, respectively, can be written as

$$N_c = n_i \exp[(E_c - E_i)/k_BT] \tag{4.4.3}$$

$$N_v = n_i \exp[(E_i - E_v)/k_BT] \tag{4.4.4}$$

Substituting these expressions for N_c and N_v in Eqs. (4.3.18) and (4.3.19) yields

$$n = n_i \exp[(E_F - E_i)/k_B T] \tag{4.4.5}$$

$$p = n_i \exp[(E_i - E_F)/k_B T] \tag{4.4.6}$$

The intrinsic Fermi energy can be derived from Eqs. (4.4.1) and (4.4.2) as

$$E_i = \frac{E_c + E_v}{2} + \frac{k_B T}{2} \ln\left(\frac{N_v}{N_c}\right) \tag{4.4.7}$$

and since $(N_v/N_c) = (m_h^*/m_e^*)^{3/2}$, we can also write

$$E_i = \frac{E_c + E_v}{2} + \frac{3k_B T}{4} \ln\left(\frac{m_h^*}{m_e^*}\right) \tag{4.4.8}$$

which indicates that, if $m_h^* = m_e^*$ or $T=0$ K, the Fermi level in an intrinsic semiconductor is positioned at mid-gap. In real cases, $m_h^* \neq m_e^*$, resulting in a small deviation of the Fermi level from mid-gap.

The equilibrium density of electrons and holes in a nondegenerate semiconductor is constant at a given temperature. The product of the electron and hole density, in a nondegenerate semiconductor at equilibrium, is always equal to the square of the intrinsic carrier density, i.e.,

$$np = n_i^2 \tag{4.4.9}$$

The intrinsic carrier density (which is specific to a given semiconductor) is related to the effective conduction band (denoted as N_c) and valence band (denoted as N_v) densities of states, i.e.,

$$np = n_i^2 = N_c N_v \exp[(E_v - E_c)/k_B T] = N_c N_v \exp(-E_g/k_B T) \tag{4.4.10}$$

This relationship, referred to as the *mass action law,* allows (at thermal equilibrium) to determine the electron density if the hole density is known or vice versa. It should be noted that this equation signifies that, although the electron–hole pairs may be continuously generated and recombined, the product of the concentration (averaged in time) stays constant. This equation also indicates that, for a nondegenerate material in equilibrium, the *np* product depends on the effective conduction band and valence band densities of states, the energy gap of a semiconductor, and the temperature, and it is independent of the Fermi level position and of the individual electron and hole densities. In other words, regardless of doping, the *np* product is a constant at a given temperature. From this equation, the intrinsic carrier density can be written as

TABLE 4.2. The energy gap and intrinsic carrier concentration (n_i) for selected semiconductors (at 300 K)

	Ge	Si	GaAs
E_g (eV)	0.67	1.12	1.42
n_i (cm^{-3})	2.5×10^{13}	1.0×10^{10}	2.0×10^{6}

$$n_i = (N_c N_v)^{1/2} \exp(-E_g/2k_B T) \tag{4.4.11}$$

The intrinsic carrier concentration (n_i) for selected semiconductors (at 300 K) is listed in Table 4.2.

As can be seen from Eq. (4.4.11), the temperature dependence of the intrinsic carrier concentration is dominated by the exponential factor containing the energy gap, which itself depends on the temperature (see Section 4.10). The temperature dependence of the effective densities of states are according to Eqs. (4.3.14) and (4.3.17) that also contain the effective masses in which temperature dependences can be excluded, as these are small in comparison. The dependence of the intrinsic carrier concentration in Si and GaAs on temperature is shown in Fig. 4.3.

As can be seen from Eq. (4.4.11) and Table 4.2 (showing the exponential dependence of the carrier concentration on the energy gap), at room temperature, for the change in energy gap from 0.67 (for Ge) to 1.42 eV (for GaAs), the intrinsic carrier concentration varies by seven orders of magnitude; so does the number of carriers (both electrons and holes) available for charge transport.

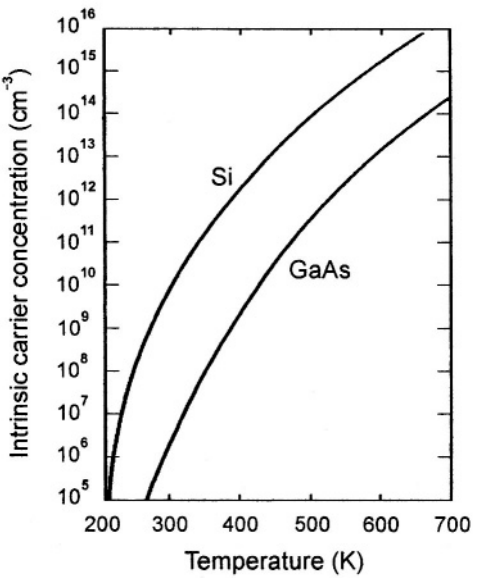

FIGURE 4.3. Dependence of the intrinsic carrier concentration on temperature in Si and GaAs.

As mentioned earlier (see Section 4.2), in semiconductors, both electrons and holes contribute to the current; thus in general, the bulk conductivity σ of an intrinsic semiconductor is expressed as

$$\sigma = ne\mu_e + pe\mu_h \tag{4.4.12}$$

where n and μ_e are the concentration and mobility of electrons, respectively, and p and μ_h are the concentration and mobility of holes, respectively, and e is the carrier charge. Thus, in the case of an intrinsic semiconductor (i.e., $n = p = n_i$), from Eq. (4.4.11) and equations for the effective densities of states for the conduction and valence bands (see Eqs. 4.3.14 and 4.3.17), we can write

$$n = p = \text{constant} \times T^{3/2} \exp\left(-E_g/2k_B T\right) \tag{4.4.13}$$

and

$$\sigma = \text{constant} \times e(\mu_e + \mu_h) T^{3/2} \exp\left(-E_g/2k_B T\right) = \sigma_0 \exp\left(-E_g/2k_B T\right) \tag{4.4.14}$$

Thus, by plotting $\ln \sigma$ as a function of $1/T$ (which yields a straight line), the energy gap E_g can be derived from the slope $(-E_g/2k_B)$. (In this case the $T^{3/2}$ variation and the temperature variation of E_g are disregarded as negligible in comparison with the exponential temperature variation term.) Note that the electron and hole mobilities are also dependent on temperature; however, these dependences, which are in theory dependent on T as $\mu \propto T^{-3/2}$, would largely cancel out the $T^{3/2}$ dependence.

4.4.2. Extrinsic Semiconductors

The availability of charge carriers in the valence and conduction bands is greatly affected by the presence of intentionally (or unintentionally) introduced impurities (i.e., foreign atoms incorporated into the crystal structure of a semiconductor). In semiconductors, some impurities are deliberately introduced to produce materials and devices with desired properties. In this case, the material is referred to as *extrinsic,* and the process of putting impurities into the lattice is called *doping.* The contribution of free carriers by dopants requires them being *ionized* (i.e., the dopants have donated or accepted an electron). The ionization of the dopants depends on the thermal energy and the position of the impurity level in the energy gap of a semiconductor.

Impurities that contribute electrons to the conduction band are called *donors,* and those that supply holes to the valence band are *acceptors.* (In this context, it should be reiterated again that impurity atoms in crystals are considered as point defects if they are detrimental in the utilization of the material or device, but if they are deliberately incorporated in the material in order to control conductivity, they are referred to as donors or acceptors.) These properties of semiconductors are discussed in the following section.

4.5. DONORS AND ACCEPTORS IN SEMICONDUCTORS

As mentioned earlier, in extrinsic semiconductors, some impurities are deliberately introduced to produce materials and devices with desired properties. Such a process of putting impurities into the lattice is called doping, and those impurities that contribute ("donate") electrons to the conduction band are called donors, and those that supply holes to the valence band (i.e., "accept" electrons) are acceptors.

Donors are substitutional impurities that have a higher valence than the atoms of the host material; when a donor impurity is ionized, an electron is donated to the conduction band, which leads to an excess of mobile electrons, and the material is referred to as *n*-type. The electrons donated to the conduction band can participate in the conduction process, whereas the donor centers become positively charged. At sufficiently low temperatures, electrons can be captured by these positively charged donor centers, which become neutral.

Acceptor impurities have a lower valence than the host, which leads to incomplete atomic bonding in the lattice; thus they capture electrons, i.e., supply holes to the valence band (the acceptor centers become negatively charged), and the semiconductor is referred to as *p*-type. At sufficiently low temperatures, holes become localized at acceptor centers, which become neutral.

Both the donor and acceptor levels are located in the forbidden energy gap (see the energy band diagram of a semiconductor in Fig. 4.4). In general, the energy levels in the gap of a semiconductor can be categorized as *shallow* and *deep levels* according to their depth from the nearest band edges. The donors and acceptors are called shallow when their levels are close to the bottom of the conduction band and the top of the valence band, respectively. *Shallow impurities* are those that require typically energies corresponding to about the thermal energy to be ionized. On the other hand, *deep impurities* require higher energies in order to be ionized,

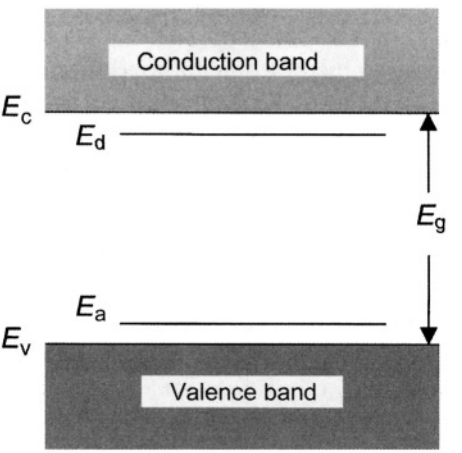

FIGURE 4.4. Schematic diagram of donor and acceptor levels located in the forbidden energy gap.

thus they are typically not expected to contribute free carriers. Such deep impurities can be effective recombination centers, i.e., centers in which electrons and holes drop and annihilate each other.

Typically, for shallow impurities, at room temperature almost the entire donor or acceptor sites are ionized, and the free carrier density corresponds to the impurity concentration. In other words, in the case of donors, the electron density n equals the concentration of donors (note that in this case $N_d^+ \cong N_d$), i.e., $n \cong N_d$; and in the case of acceptors, the density of holes p equals the concentration of acceptors ($N_a \cong N_a^-$), i.e., $p \cong N_a$. In these cases, N_d and N_d^+ are total concentration of donors and total concentration of ionized donors (positively charged), respectively; and N_a and N_a^- are total concentration of acceptors and total concentration of ionized acceptors (negatively charged), respectively.

In an extrinsic semiconductor, electrons are *majority carriers* and holes are *minority carriers* in n-type semiconductor. In p-type material, electrons are minority carriers and holes are majority carriers.

When similar concentrations of shallow donors and acceptors are present in the semiconductor, one type of impurity will cancel out the effect of the other, and the semiconductor is referred to as *compensated.* In such a case, if the donor concentration is larger and if $N_d^+ - N_a^- \gg n_i$, the carrier concentration is about equal to the difference between the donor and acceptor concentration, giving n-type material, i.e., $n \cong N_d^+ - N_a^-$. Similar considerations can be applied if the acceptor concentration is larger than the donor concentration.

In compound semiconductors, such as the III–V binary compound GaAs, an excess of one of the components may also generate donor or acceptor states. In this case, an excess of Ga atoms would lead to a p-type material, whereas an excess of As atoms would lead to an n-type material. Also, in compound semiconductors such as GaAs, e.g., group IV element Si may occupy lattice sites either in the (group III) Ga sublattice or in the (group V) As sublattice; thus, the former case (i.e., Si_{Ga}) results in an n-type doping, whereas the Si_{As} case results in p-type doping. Because of the preferential occupation by the Si atoms of the Ga sublattice sites, Si is an n-type dopant in GaAs; when the concentrations of Si_{Ga} and Si_{As} are comparable, the conductivity of GaAs decreases and the material is referred to as compensated.

In the case of shallow donors, the weakly-bound excess electron is, on the average, located sufficiently far from the donor center, indicating that the specific atomic structure of the impurity can be considered as a positive point charge. Thus, the *binding energy,* E_D, of the electron to the donor impurity, or the *impurity ionization energy* (i.e., the energy required to transfer an electron from the donor level to the bottom of the conduction band), can be determined by considering an extra electron in the donor atom as a particle with an effective mass m^* moving in the presence of a positive net charge. This is analogous to a hydrogen atom embedded in the dielectric medium of the crystal. Thus, the ground-state binding energy of the extra electron becomes

$$E_D = \frac{13.6 m_e^*}{m_0 \varepsilon^2} \text{ (eV)} \tag{4.5.1}$$

where 13.6 eV is the ionization energy of the hydrogen atom and ε is the dielectric constant of the solid, and m_0 is the free electron mass. (Note that in this case, a static dielectric constant, ε_{st}, should be used, see Table B1 in Appendix B.) The binding energy is measured relative to the conduction band level. In terms of the energy-band diagram, donor impurities introduce levels at energies E_d below the bottom of the conduction band E_c, thus $E_D = E_c - E_d$ (see Fig. 4.4).

The electron orbit around the impurity atom, i.e., the spatial extent of the wave function, can also be estimated. In this case, the radius of the first Bohr orbit becomes $r = (\varepsilon m_0/m^*)a_0$, where a_0 is the radius of the first Bohr orbit in the hydrogen atom. These results indicate that for typical cases, the ionization energies are less than the room temperature energy $k_B T$ of about 0.026 eV, and values of r are much larger than the atomic diameter. Thus, most shallow donors are expected to be ionized at room temperature, and the wave functions are expected to extent over many atomic diameters; in other words, the electron is not localized at the impurity.

Similar analysis can be considered in the case of acceptor impurities. In this case it is convenient to describe a positive hole orbiting a negatively ionized impurity atom. The donor and acceptor binding energies are expected to be somewhat different because of the difference in the effective masses of the electrons and holes (see Section 3.6). In terms of the energy-band diagram, acceptor impurities introduce levels at energies E_a above the top of the valence band E_v, thus the acceptor binding energy (i.e., the energy required to excite an electron from the valence band to the acceptor energy levels) $E_A = E_a - E_v$ (see Fig. 4.4).

Following the arguments presented earlier for shallow impurities, in elemental group IV semiconductors, such as Si or Ge, group V elements (e.g., P, As) placed in the tetrahedral host structure can easily donate an extra electron and become donors, whereas group III elements (e.g., B, Al) can easily capture an electron from the host structure and become acceptors. As mentioned earlier, in compound semiconductors, such as GaAs, however, the same element (e.g., Si, Ge) can become a donor or an acceptor depending on whether it substitutes for Ga or As. Ionization energies of common donors and acceptors in Si and GaAs are given in Tables 4.3 and 4.4.

In the analysis of extrinsic semiconductors, it is of great importance to determine the carrier concentrations in terms of the Fermi level and other semiconductor parameters. This will allow to determine the dependence of the Fermi level on the carrier concentration in an extrinsic semiconductor.

TABLE 4.3. Ionization energies of common shallow donors and acceptors in Si

Donors	Ionization energy, E_D (eV)	Acceptors	Ionization energy, E_A (eV)
P	0.045	B	0.045
As	0.049	Al	0.057
Sb	0.039	Ga	0.065
Bi	0.069	In	0.16

TABLE 4.4. Ionization energies of common donors and acceptors in GaAs

Donors	Ionization energy, E_D (eV)	Acceptors	Ionization energy E_A (eV)
C	0.006	C	0.026
Si	0.006	Si	0.035
S	0.006	Be	0.028
Se	0.006	Zn	0.031

The distribution of electrons and holes in their corresponding bands as a function of energy for both donor-doped and acceptor-doped semiconductor can be derived from the known density of available states and the probability of occupation of these states (i.e., the Fermi–Dirac distribution function). These results are shown in Fig. 4.5. Similar to the case of an intrinsic semiconductor (see Fig. 4.2), most of the carriers in both the conduction and valence bands are distributed (as a function of energy) near the corresponding band edges. The difference in the carrier distribution in the valence and conduction bands (in both *n*- and *p*-type cases) is due to the positioning of the Fermi level. In other words, for the Fermi level located in the upper half of the energy gap, the electron population

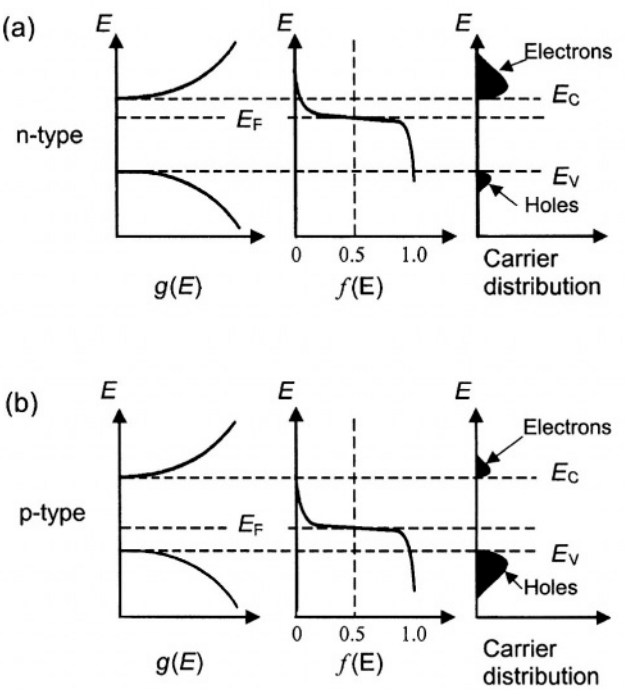

FIGURE 4.5. Schematic diagram of the DOS, the Fermi–Dirac function, and the resulting distribution of carriers as a function of energy for (a) an *n*-type semiconductor and (b) for a *p*-type semiconductor. The carrier concentration for electrons is a product of $g_n(E)f(E)$, and for holes it is $g_p(E)[1 - f(E)]$.

is higher than the hole population; whereas, for the Fermi level located in the lower half of the energy gap, the hole population is higher than the electron population.

As described earlier (see Eq. 4.4.9), for a given nondegenerate semiconductor in a thermal equilibrium, the product of the electron and hole density always equals the square of the intrinsic carrier density, i.e., $np = n_i^2$. In general, a semiconductor can be considered as a reservoir of electrical charge, which includes both fixed charge (due to ionized dopant atoms that are immobile) and mobile charge (due to electrons and holes). In equilibrium, the semiconductor is neutral, and thus the condition of charge balance (or charge neutrality) between the negative and positive charges can be expressed as

$$n + N_a^- = p + N_d^+ \tag{4.5.2}$$

where n and p are the densities of mobile conduction band electrons (–) and mobile valence band holes (+), respectively; and N_a^- and N_d^+ are the densities of fixed ionized acceptor atoms (–) and fixed ionized donor atoms (+), respectively. As mentioned earlier, for shallow impurities, at room temperature almost the entire donor or acceptor sites are ionized, and in this case $N_d^+ \cong N_d$, and $N_a^- \cong N_a$, and thus the charge neutrality relationship is often written as

$$n + N_a = p + N_d \tag{4.5.3}$$

Equations (4.4.9) (i.e., $np = n_i^2$) and (4.5.3) can be used to derive n and p if N_a and N_d are known (assuming total ionization and nondegenerate semiconductors). For the case of a semiconductor with $|N_a - N_d| \gg n_i$ and $N_a > N_d$, the material is p-type, and in this case

$$p = N_a - N_d \tag{4.5.4}$$

$$n = \frac{n_i^2}{N_a - N_d} \tag{4.5.5}$$

and if $N_d > N_a$, the material is n-type, and in this case

$$n = N_d - N_a \tag{4.5.6}$$

$$p = \frac{n_i^2}{N_d - N_a} \tag{4.5.7}$$

The general expressions for carrier concentrations can be derived by eliminating n or p in Eq. (4.5.3) and using Eq. (4.4.9); thus, quadratic equations for n or for p can be written as

$$n^2 - n(N_d - N_a) - n_i^2 = 0 \tag{4.5.8}$$

$$p^2 - p(N_a - N_d) - n_i^2 = 0 \tag{4.5.9}$$

The solutions for n and p are

$$n = \frac{N_d - N_a}{2} + \left[\left(\frac{N_d - N_a}{2}\right)^2 + n_i^2\right]^{1/2} \tag{4.5.10}$$

$$p = \frac{N_a - N_d}{2} + \left[\left(\frac{N_a - N_d}{2}\right)^2 + n_i^2\right]^{1/2} \tag{4.5.11}$$

Note that since n, or p, must be $\geqslant 0$, only the positive roots are considered.

These equations, which assume total ionization of dopant sites, provide general expressions for deriving carrier concentrations in compensated semiconductors. For specific cases, these equations can be simplified. For example, considering the cases described above (see Eqs. 4.5.4 – 4.5.7), for specific cases of either $N_a \gg N_d$ (i.e.,, p-type semiconductor), or $N_d \gg N_a$ (i.e., n-type material), which are of practical importance in device applications, we obtain $p \cong N_a$ and $n \cong n_i^2/N_a$ and $n \cong N_d$ and $p \cong n_i^2/N_d$, respectively.

For donor-doped (i.e., n-type) and acceptor-doped (i.e., p-type) nondegenerate semiconductors at equilibrium, the Fermi level can be related to the carrier concentration by using Eqs. (4.4.5) and (4.4.6), which give (assuming that all the dopants are fully ionized and $n \cong N_d$ and $p \cong N_a$, for n- and p-type materials, respectively)

$$E_F = E_i + k_B T \ln (N_d/n_i) \tag{4.5.12}$$

$$E_F = E_i - k_B T \ln (N_a/n_i) \tag{4.5.13}$$

From these equations it follows that with increasing doping the Fermi level approaches the conduction band in the case of donor-doped semiconductors, and it approaches the valence band in the case of acceptor-doped materials (see Fig. 4.6). This figure, which represents the variation of the Fermi energies with dopant concentration, shows the dependence of the Fermi energy (at room temperature) of n- and p-type semiconductors as a function of dopant concentration (in log scale). The dotted segments indicate the zones corresponding to degeneracy, i.e., the energy difference between the band edge and the onset of dotted curve corresponds to $3k_BT$. For example, in Si such an onset of degeneracy corresponds to a donor concentration of about 1.5×10^{18} cm^{-3}, and acceptor concentration of about 0.9×10^{18} cm^{-3}. (Note that such semiconductors are also referred to as *highly doped* and denoted as n^+, or p^+-semiconductors.) Thus, to summarize briefly, for the case of the Fermi level being positioned within $3k_BT$ of the band edges, or positioned inside either (conduction or valence) band, the semiconductor is referred to as degenerate, and according to the previous discussion (see Section 4.2), the exponential relationships for the carrier concentrations, i.e., Eqs. (4.3.18), (4.3.19), (4.4.5) and (4.4.6) are no longer valid.

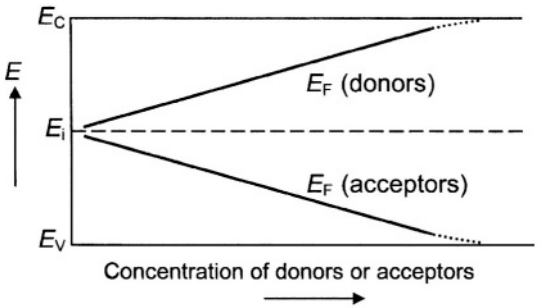

FIGURE 4.6. Dependence of Fermi energy (at room temperature) of *n*- and *p*-type semiconductors as a function of dopant concentration (in log scale). The dotted segments indicate the zones corresponding to degeneracy (i.e., the energy difference between the band edge and the onset of dotted curve corresponds to $3k_BT$).

It should be noted that, for increasingly high doping concentrations, the wave functions of the electrons bound to the dopant atoms begin to overlap as the average distance between the nearest dopant atoms is reduced; thus, discrete dopant (donor or acceptor) levels broaden out forming an energy band, which eventually merges with the conduction or valence band (depending on the type of doping); this is also accompanied by the narrowing of the energy gap (not shown in Fig. 4.6), which can be usually disregarded for doping concentrations below about 10^{18} cm $^{-3}$. It should be noted that in heavily doped semiconductors, the filling of the states near the band edges may result in the *Burstein–Moss shift* of the absorption edge (see Section 4.7). This occurs due to the fact that with an increasing doping concentration in, e.g., an *n*-type semiconductor, the Fermi level also gradually increases, and in the degenerated case, it moves above the conduction band edge. In such a case, the states for electron transitions near the conduction band edge are now occupied, resulting in the electron transitions to higher energy states, and consequently the onset of optical absorption is shifted to higher energies with increasing doping.

From Eqs. (4.5.12) and (4.5.13), it is also possible to determine the dependence of the Fermi energy on temperature for a given doping concentration. In the previous discussions, the total ionization of the dopant sites was assumed at room temperature. However, in the lower temperature cases, it is essential to consider the degree of ionization of dopant sites. The concentration of ionized donors N_d^+ in relation to the total concentration of donor atoms N_d can be expressed as

$$\frac{N_d^+}{N_d} = \frac{1}{1 + g_d \exp[(E_F - E_d)/k_BT]} \tag{4.5.14}$$

where $g_d = 2$ is the donor degeneracy factor, representing the fact that the donor state can have an electron with the spin up or down.

The concentration of ionized acceptors N_a^- in relation to the total concentration of acceptor atoms N_a can be expressed as

$$\frac{N_a^-}{N_a} = \frac{1}{1 + g_a \exp[(E_a - E_F)/k_B T]} \tag{4.5.15}$$

where $g_a = 4$ is the acceptor degeneracy factor. In this case, in addition to spin up or down, light and heavy holes must be accounted for.

The dependence of the Fermi energy on temperature for *n*- and *p*-type semiconductors with various doping concentrations is given in Fig. 4.7. As shown in this figure, the Fermi energy approaches at higher temperatures the Fermi energy corresponding to the intrinsic material; whereas at lower temperatures, the Fermi energy in *n*- and *p*-type materials is closer to the bottom of the conduction band and the top of the valence band, respectively. The curves corresponding to E_c and E_v also show the dependence of the energy gap on temperature (see Section 4.10). Note that the intrinsic Fermi level E_i, given in Fig. 4.7, is plotted according to Eq. (4.4.8), which indicates that, for $m_h^* = m_e^*$ (or for T=0 K), E_i is positioned at mid-gap. In real cases, however, typically $m_h^* > m_e^*$, resulting in a small deviation of the Fermi level from mid-gap, which would result (especially at higher temperatures) to upward bending of E_i (not shown in Fig. 4.7).

As mentioned earlier, one of the most important objectives in describing a semiconductor in relation to its properties is the determination of the carrier concentrations in the material. For understanding the semiconductor device operation at different temperatures, the dependence of the carrier concentration on temperature is required. Such dependence is presented in Fig. 4.8, which shows a plot of ln *n* as a function of $1/T$ for a donor-doped semiconductor. Recalling some of the observations related to the Fermi level, (i) in an intrinsic semiconductor, the Fermi level is positioned at mid-gap and (ii) in a doped semiconductor, the Fermi energy approaches at higher temperatures the Fermi energy corresponding to

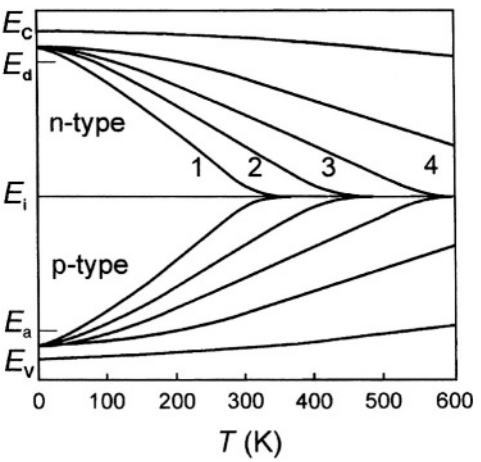

FIGURE 4.7. The dependence of the Fermi energy on temperature for *n*- and *p*-type semiconductors; the curves 1, 2, 3, and 4 correspond to increasing doping concentrations. The curves corresponding to E_c and E_v also show the dependence of the energy gap on temperature.

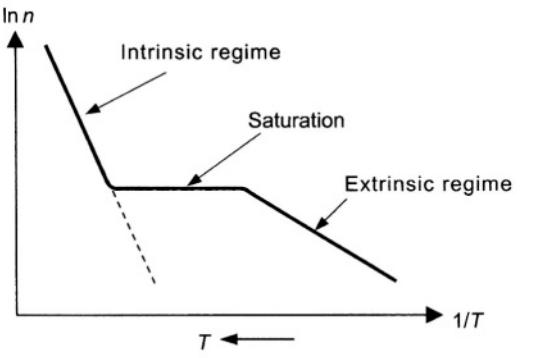

FIGURE 4.8. Dependence of the carrier concentration of a donor-doped semiconductor on temperature.

the intrinsic material; whereas at lower temperatures, the Fermi energy in *n*- and *p*-type materials is closer to the bottom of the conduction band and the top of the valence band, respectively. Three distinct regions can be distinguished in Fig. 4.8. At lower temperatures (e.g., below about 100 K for Si), carriers are excited from the dopant levels, and the carrier concentration is determined by the degree of ionization of dopant sites (see Eq. 4.5.14). This region is referred to as an *extrinsic region.* In the intermediate range (e.g., between about 100 and 400 K in Si), between the low and high temperature regimes, since all the donors are ionized, no further contribution to the carrier concentration in the conduction band with temperature is expected, but these temperatures are insufficient for the noticeable carrier excitations from the valence band to the conduction band. Thus, the carrier concentration does not change much with temperature, and the *saturation range* (or as often referred to as *exhaustion range*) is established. At higher temperatures, since all the donors are ionized and the carrier concentration in the conduction band is dominated by the thermal excitations from the valence band, the material exhibits the behavior of an intrinsic semiconductor. It should be noted that, in general, semiconductor devices operate in the saturation range, where the carrier concentrations are constant, and any changes in conductivity are predominantly due to temperature variations in mobility.

As discussed earlier, in the case of an intrinsic semiconductor, the conductivity can be expressed by Eq. (4.4.14) as

$$\sigma = \text{constant} \times e(\mu_e + \mu_h) T^{3/2} \exp\left(-E_g/2k_B T\right)$$

$$= \sigma_0 \exp\left(-E_g/2k_B T\right) \tag{4.5.16}$$

In the case of an extrinsic (or doped) material, for $n \gg p$, the material is a *n*-type semiconductor, and for $p \gg n$, the material is a *p*-type semiconductor. From the temperature dependence of the conductivity it is possible to determine the dopant (or impurity) levels in semiconductors. This is because (i) of the dominant effect

of the carrier density (i.e., n or p) on the σ (T), and (ii) of the fact that the carrier concentration, determined by a balance between the thermal generation of carriers from existing levels and recombination processes, is related to the corresponding dopant energy level in the energy gap of a semiconductor. As mentioned earlier, for the extrinsic semiconductor (e.g., n-type material), three possible temperature regimes have to be considered. At high temperatures, since all the donors are ionized and the carrier concentration in the conduction band is dominated by the excitations from the valence band, the material exhibits the behavior of an intrinsic semiconductor; thus, the plot of $\ln \sigma$ as a function of $1/T$ will yield a straight line with the slope $s = -E_g/2k_B$. At low temperatures, on the other hand, the concentration of electrons (in the conduction band) originating from the donor levels dominates, and one can disregard the contribution of the holes to the conductivity process. In such a case, the conductivity can be expressed as

$$\sigma = \sigma_0 \exp[-(E_c - E_d)/2k_B T] \tag{4.5.17}$$

Thus, the plot of $\ln \sigma$ as a function of $1/T$ will yield a straight line with the slope $s = -(E_c - E_d)/2k_B$ (see Fig. 4.9). In the intermediate range, between the low and high temperature regimes, as mentioned earlier, the saturation range is established, where the $T^{3/2}$ term and temperature variations in mobility determine the specific shape of the curve. Similar reasoning applies to p-type semiconductors.

Thus, to summarize briefly, the shallow donor and acceptor impurities produce in the crystal relatively small perturbations that result in the formation, in the energy gap, of bound states in very close proximity to the boundaries of the valence and conduction bands. These states can contribute carriers to the respective bands, and thus control the electrical properties of a semiconductor. However, the simple hydrogenic model for impurity levels is not always applicable. In some cases impurities introduce deep levels in the fundamental energy gap of a semiconductor. In addition to the impurity-induced deep levels, a variety of defects may also give rise to bound states in the energy gap of the

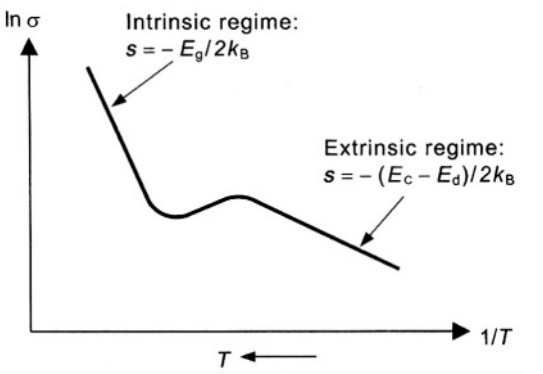

FIGURE 4.9. Variation of electrical conductivity of donor-doped semiconductor on temperature.

material. These defects can be vacancies, interstitials, antisite defects and their complexes, dislocations, stacking faults, grain boundaries, or precipitates. These states are usually located deeper in the energy gap and are more localized. The information about these deep centers is important in both the understanding and applications of semiconductors, since deep centers typically act as efficient traps and control the carrier lifetime. However, no universal theory is available to account for all these centers.

4.6. NONEQUILIBRIUM PROPERTIES OF CARRIERS

Important processes related to excess minority carriers in a semiconductor include (i) *generation* of excess carriers, (ii) *diffusion* (due to the concentration gradient), (iii) *drift* (due to the applied electric filed), and (iv) *recombination.* A semiconductor is considered to be in a nonequilibrium state if the equation $np = n_i^2$ is no longer satisfied. The condition $np > n_i^2$ implies the injection of an additional charge into a semiconductor (i.e., *carrier injection*), and it can be caused by optical excitation with photon energy greater than the energy gap, or by forward bias of a semiconductor junction device. For the case of carrier injection, two regimes are considered: *low-level injection* (the increase in carrier concentration is much smaller than the doping concentration) and the *high-level injection* (the increase in carrier concentration is commensurate to the doping concentration). The case of $np < n_i^2$ corresponds to *carrier extraction.*

In the case of carrier injection into a semiconductor, the carrier distribution described by the Fermi–Dirac function (see Section 4.3) is no longer valid. In such a nonequilibrium case, under certain assumptions, the carrier occupation can be described by introducing the *quasi-Fermi energies* E_{Fn} and E_{Fp} for electrons and holes, respectively (these are also referred to as *quasi-Fermi levels*). This description indicates that although the electrons and holes are no longer in thermal equilibrium, they can be described by introducing different Fermi energies for electrons and holes. (This is valid under an assumption that the electrons are in thermal equilibrium in the conduction band and holes are in thermal equilibrium in the valence band.) Thus, by introducing distinct Fermi energies for electrons and holes (i.e., quasi-Fermi energies E_{Fn} and E_{Fp}), the excess carrier densities (for a nondegenerate case) can be described by the following equations (see Eqs. 4.3.18 and 4.3.19)

$$n = N_c \exp[(E_{Fn} - E_c)/k_B T] \tag{4.6.1}$$

$$p = N_v \exp[(E_v - E_{Fp})/k_B T] \tag{4.6.2}$$

At equilibrium, $E_{Fn} = E_{Fp} = E_F$. (Note that with excess electron and hole injection into a semiconductor, E_{Fn} and E_{Fp} move towards the conduction and valence bands, respectively.) In a nonequilibrium case, $E_{Fn} \neq E_{Fp}$ and these may depend

on such parameters as the position and time. Combining these equations for n and p and recalling from Eq. (4.4.10) the expression

$$n_i^2 = N_c N_v \exp[(E_v - E_c)/k_B T] \tag{4.6.3}$$

the np product for the nonequilibrium case can be expressed as

$$np = n_i^2 \exp[(E_{Fn} - E_{Fp})/k_B T] \tag{4.6.4}$$

This description of carrier properties in a nonequilibrium case provides a valuable method for analyzing semiconductor materials and devices.

The details of carrier recombination are discussed further.

4.7. INTERBAND ELECTRONIC TRANSITIONS IN SEMICONDUCTORS

During the electronic transitions (see Fig. 4.10), energy and momentum ($\hbar k$) must be conserved. As discussed in Chapter 3, when the minimum of the conduction band and the maximum of the valence band occur at the same value of the wave vector **k**, transitions are direct (or vertical), and the material is referred to as a direct-gap semiconductor. In such semiconductors, e.g., in the case of optical emission, the most likely transitions are across the minimum-energy gap, between the most probably filled states at the minimum of the conduction band and the states most likely to be unoccupied at the maximum of the valence band. If the band extrema do not occur at the same wave vector **k**, transitions are indirect, and the material is referred to as an indirect-gap semiconductor. For momentum conservation in such semiconductors, a participation of an extra particle, i.e., phonon, is required; the probability of such a process is substantially lower compared with direct transitions.

4.7.1. Optical Absorption

In general, there are several optical absorption processes, and each of these contributes to the total absorption coefficient α. These mechanisms include (i) fundamental absorption process, (ii) exciton absorption, (iii) absorption due to dopants and imperfections, (iv) absorption due to intraband transitions, and (v) free carrier absorption. At photon energies greater than the energy gap, the absorption mechanism is due to the transfer of electrons from filled valence band

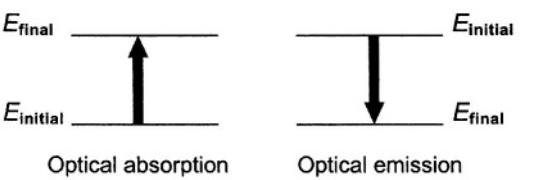

FIGURE 4.10. Schematic illustration of optical absorption and emission processes; the arrows indicate the electronic transitions.

states to the empty conduction band states. At energies slightly below the energy gap, the absorption mechanism is due to the excitons and transitions between impurity and band states (e.g., acceptor to conduction band and valence band to donor). Free carrier absorption due to the transitions within the energy bands results in absorption continuum at lower energies. These absorption mechanisms are outlined further.

In the *fundamental absorption process,* a photon excites an electron from the valence band to the conduction band. Both energy and momentum must be conserved in this process. Since the photon momentum is small compared with the crystal momentum, the absorption process should essentially conserve the electron momentum, i.e., $\hbar k$. As discussed earlier, when the minimum of the conduction band and the maximum of the valence band occur at the same value of the wave vector **k**, transitions are direct, and the material is referred to as a direct-gap semiconductor. If the band extrema do not occur at the same wave vector **k**, transitions are indirect, and the material is referred to as an indirect-gap semiconductor. For momentum conservation in such semiconductors, a participation of an extra particle, i.e., phonon, is required; the probability of such a process is substantially lower compared with direct transitions. Therefore, in general, fundamental absorption in indirect-gap semiconductors is relatively weaker as compared with direct-gap materials.

The optical absorption is described by an absorption coefficient α, which can be derived from transmission measurements. If I_0 is an incident light intensity, I is the transmitted light intensity, and R is the reflectivity, the transmission, $T = I/I_0$, can be written as (neglecting interference)

$$T = \frac{(1 - R)^2 \exp(-\alpha d)}{1 - R^2 \exp(-2\alpha d)} \tag{4.7.1}$$

where d is the thickness of the material. For large αd, this expression can be reduced to

$$T = (1 - R)^2 \exp(-\alpha d) \tag{4.7.2}$$

and, in the absence of reflection, it can be further reduced to

$$I = I_0 \exp(-\alpha d) \tag{4.7.3}$$

The absorption coefficient can be derived from the following simple treatment related to the absorption transitions between direct parabolic bands (see Fig. 4.11). The absorption coefficient can be expressed as (see Pankove, 1971, Bibliography Section B2)

$$\alpha(h\nu) = A \sum P_{if} n_i n_f \tag{4.7.4}$$

where P_{if} is the transition probability, n_i and n_f are the densities of the electrons in the initial state and of empty energy levels in the final state, respectively, and

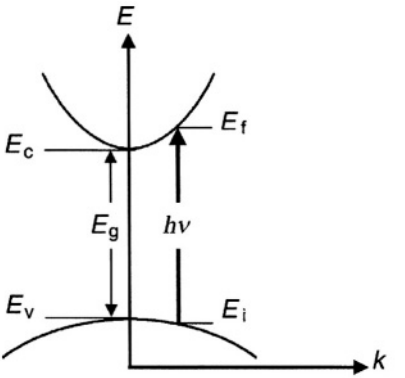

FIGURE 4.11. Schematic diagram of the absorption transitions between direct parabolic bands.

the sum is over all initial and final states. As discussed in Chapter 3, the energy associated with a given state is given by

$$E = \frac{\hbar^2 k^2}{2m} \tag{4.7.5}$$

Thus, the electron energy (relative to E_c), $E_e = (\hbar^2 k^2)/2m_e^*$, and the hole energy (relative to E_v), $E_h = (\hbar^2 k^2)/2m_h^*$. Using these equations, for the case of the transition shown in Fig. 4.11, the transition energy can be expressed as

$$h\nu = \frac{\hbar^2 k^2}{2m_e^*} + \frac{\hbar^2 k^2}{2m_h^*} + E_g \tag{4.7.6}$$

and

$$h\nu - E_g = \frac{\hbar^2 k^2}{2m_e^*} + \frac{\hbar^2 k^2}{2m_h^*} = \frac{\hbar^2 k^2}{2m_r^*} \tag{4.7.7}$$

where $m_r^* = m_e^* m_h^*/(m_e^* + m_h^*)$ is the reduced effective mass. The general expression for the DOS (see Section 4.3) is $N(E)dE = (2\pi^2\hbar^3)^{-1}(2m^*)^{3/2}E^{1/2}dE$, and thus, for the present case, the DOS can be written as

$$N(h\nu)d(h\nu) = (2\pi^2\hbar^3)^{-1}(2m_r^*)^{3/2}(h\nu - E_g)^{1/2}d(h\nu) \tag{4.7.8}$$

Thus, for direct transitions between parabolic valence and conduction bands, the absorption coefficient is

$$\alpha(h\nu) = A(h\nu - E_g)^{1/2} \tag{4.7.9}$$

where A is a constant ($A \approx 10^4$), the energy gap E_g and $h\nu$ are in eV, and α is in cm^{-1}.

It should be noted that in many semiconductors (both crystalline and amorphous), the absorption coefficient, in the absorption edge region, was found empirically to obey Urbach's rule, i.e.,

$$\alpha(h\nu) = \alpha_0 \exp[g(h\nu - h\nu_0)] \tag{4.7.10}$$

where the coefficient g is a temperature-dependent parameter for ionic crystals, whereas for covalent semiconductors g depends on the concentrations and the electrical charges of impurities.

Several models have been proposed to explain this behavior of the absorption edge. These include, e.g., the Dow–Redfield model of internal-electric field-assisted broadening of the lowest excitonic state (the sources of the internal microfields can vary from material to material and may involve, e.g., ionized impurities, phonons, dislocations, and surfaces), the Skettrup thermal fluctuation model, and the Sumi–Toyozawa model of momentary exciton trapping by phonons. None of these holds for all cases, and it is likely that several possible mechanisms have to be invoked in order to explain the Urbach rule in a wide variety of materials. As it will be discussed in Chapter 6 (Section 6.9), in amorphous semiconductors, the shape of the typically observed exponential absorption edge can be explained in terms of the joint DOS of the valence and conduction band tails [see Street, 1991 in Bibliography Section B2].

In indirect-gap materials (see Fig. 4.12), the maximum energy of the valence band and the minimum energy of the conduction band do not occur at the same wave vector **k**. In such cases, in order to conserve both energy and momentum, transitions involve three particles, i.e., photons, electrons, and phonons. The phonons (in contrast to photons) have low energy but relatively high momentum. Thus, in indirect transitions, momentum is conserved via absorption or emission of a phonon with a characteristic energy E_p (see Fig. 4.12). The absorption coefficient for a transition involving phonon absorption is given by (see Pankove, 1971)

$$\alpha_a(h\nu) = \frac{A(h\nu - E_g + E_p)^2}{\exp(E_p/k_B T) - 1} \tag{4.7.11}$$

And the absorption coefficient for a transition involving phonon emission is given by

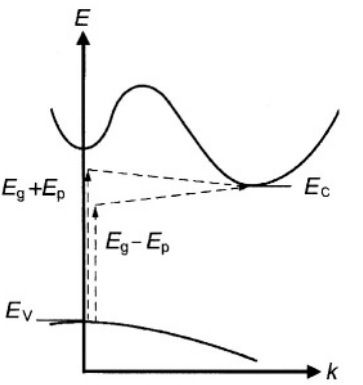

FIGURE 4.12. Schematic diagram of the absorption transitions in indirect-gap semiconductor, showing photon absorption processes involving both phonon absorption and phonon emission.

$$\alpha_e(h\nu) = \frac{A(h\nu - E_g - E_p)^2}{1 - \exp(-E_p/k_B T)} \tag{4.7.12}$$

Note that the term in the denominator in these expressions is related to the number of phonons according to Bose–Einstein statistics, $n_{\text{phonons}} = [\exp(E_p/k_B T) - 1]^{-1}$. Both processes (i.e., phonon absorption and emission) are possible for $h\nu > E_g + E_p$, and thus, the absorption coefficient can be written as

$$\alpha(h\nu) = \alpha_a(h\nu) + \alpha_e(h\nu) \tag{4.7.13}$$

Figure 4.13 shows the dependences in Eqs. (4.7.11) – (4.7.13).

The absorption process in indirect-gap materials requires the involvement of an extra particle (i.e., phonon) as compared with the direct-gap material. Thus, the probability of the absorption of light (and hence the absorption coefficient) in this case is much lower than in direct-gap materials. This implies that light has to travel a large distance into the material before being absorbed. At the absorption edge region, the absorption coefficient α may vary in the range between 10^4 and 10^5 cm^{-1} for direct transitions, and it may have values between 10 and 10^3 cm^{-1} for indirect transitions.

As mentioned in Section 4.5, in heavily doped semiconductors, the filling of the states near the band edges may result in the *Burstein–Moss shift* of the absorption edge. This occurs due to the fact that with increasing doping concentration in, e.g., an *n*-type semiconductor, the Fermi level also gradually increases, and in the degenerated case, it moves above the conduction band edge. In such a case, the states for electron transitions near the conduction band edge are now occupied, resulting in the electron transitions to higher energy states, and thus, the absorption edge shifts to higher energies. The Burstein–Moss shift of the absorption edge can be substantial, especially in the narrow energy-gap semiconductors, such as InAs.

Exciton absorption is observed at energies which are slightly lower than the energy gap of a semiconductor. Excitons (i.e., electron–hole pairs in a bound state)

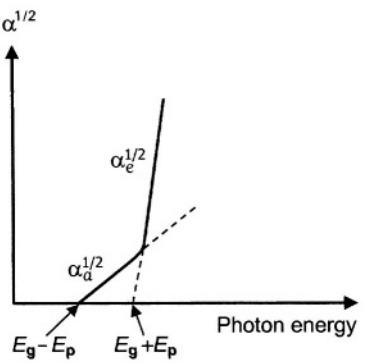

FIGURE 4.13. Schematic illustration of a plot of $\alpha^{1/2}$ as a function of photon energy.

are typically formed at low (cryogenic) temperatures through the Coulombic interaction between the electron–hole pairs. The excitonic states are at energies just below the conduction band, and they can exist in a series of bound states in the gap. Excitonic levels are typically observable at such low temperatures at which $k_B T$ is smaller than the excitonic binding energy (see Table 4.5). The Coulombic interaction between the pairs, modified by the dielectric constant of the semiconductor, brings their energy levels closer together than the width of the energy gap. Excitonic states are analogous to electronic states in the hydrogen atom, i.e., excitons can exist in a series of bound states in the gap (see Section 4.8). The exciton, which can move through the crystal without transporting any net charge, may thermally dissociate (depending on temperature) or may recombine leading to the emission of photons.

Absorption due to dopants (observed as shoulders on the low-energy region of the absorption edge) occurs as a result of transitions between impurity and band states (e.g., acceptor level-to-conduction band and valence band-to-donor level), as well as transitions between donor level-to-conduction band and valence band-to-acceptor level that result in absorption in the far infrared region. Absorption may also occur due to the imperfections that introduce localized energy levels in the energy gap of a semiconductor. (These imperfections include, e.g., point defects, impurities, dislocations, and grain boundaries; see Section 2.4.) In this case, electronic transitions are from occupied imperfection levels to the conduction band, or from the valence band to the unoccupied imperfection levels. Absorption between two imperfection levels may also occur.

Absorption due to intraband transitions is due to the presence, in most semiconductors, of the valence band with separated light-hole (LH) and heavy-hole (HH) bands, as well as the split-off band (see Chapter 3). This intervalence band absorption is due to the electron transitions between the split-off band and LH and HH bands, as well as between LH and HH bands.

Free carrier absorption is due to the transitions to higher energy levels within the same energy band resulting in absorption continuum at lower energies. This case of optical absorption requires a change in both the energy and momentum of the carrier during the transition; thus, it requires participation of a photon and a phonon. The absorption coefficient due to free carriers α_{fc} is proportional to λ^p, where p is between 2 and 3. Additional optical absorption due to free carriers is the *plasma resonance absorption,* related to free carriers acting collectively as a free-electron gas.

TABLE 4.5. Exciton binding energies (in milli-electron-volts, meV) of selected semiconductors

Si	14
InP	4.8
GaAs	4.2
CdTe	12
CdS	28
ZnS	39

4.8. RECOMBINATION PROCESSES

In semiconductors, various excitations (e.g., photon or electron irradiation) may lead to the generation of charge carriers in excess of the thermal equilibrium densities. Recombination of electron–hole pairs restores that equilibrium. Recombination centers with energy levels in the gap of a semiconductor are distinguished as *radiative* or *nonradiative,* depending on whether the recombination results in the emission of a photon or not, respectively.

One of the important applications of semiconductors is related to the process of *luminescence,* where these recombination processes play a crucial role. When a semiconductor is supplied with a certain form of energy, it may emit photons in excess of thermal radiation. Depending on the source of excitation of the luminescent material, the luminescence process can be categorized as *photoluminescence* (photon excitation), *electroluminescence* (excitation by application of an electric field), and *cathodoluminescence* (excitation by cathode rays, or energetic electrons). In semiconductors, luminescence is generally described in terms of radiative recombination of electron–hole pairs, which may involve transitions between states in the conduction or valence bands and those in the energy gap due to, e.g., donors and acceptors. In semiconductors, having appropriate values of energy gap, emission of photons occurs in the visible range of the electromagnetic spectrum (i.e., between about 0.4 and 0.7 μm, corresponding to about 3.1 and 1.8 eV). This makes semiconductors very attractive in a variety of optoelectronic applications, as well as makes the luminescence measurement a powerful tool for the characterization of electronic properties of semiconductors (see Chapter 7).

In semiconductors, luminescence is due to electronic transitions between quantum mechanical states that usually differ in energy by less than 1 eV to more than several electron-volts. Luminescence spectra can be divided between (i) *intrinsic* (fundamental or edge emission) and (ii) *extrinsic* (characteristic or activated) emission. Intrinsic luminescence is due to the recombination of electrons and holes across the energy gap (i.e., it is an "intrinsic" property of the material). At ambient temperatures, intrinsic luminescence appears as a band of energies with its intensity peak at a photon energy $h\nu_p \cong E_g$. Thus, any change in the energy gap (e.g., due to temperature or high doping concentrations) can be monitored by measuring $h\nu_p$.

As discussed earlier, in direct-gap semiconductors (e.g., GaAs and CdS), the most likely (direct) transitions are across the minimum energy gap, between the most probably filled states at the minimum of the conduction band and the states most likely to be unoccupied at the maximum of the valence band. Radiative recombination between electrons and holes is relatively likely in such transitions. In indirect-gap semiconductors (e.g., Si and GaP), since the recombination of electron–hole pairs requires a participation of an extra particle (i.e., phonon), the probability of such a process is significantly lower as compared with direct transitions. Thus, fundamental emission in indirect-gap semiconductors is relatively weak, especially when compared with that due to impurities or defects. In both direct- and indirect-gap semiconductors, the emission spectra that depend on the presence of various impurities are extrinsic in nature.

4.8.1. Radiative Transitions

A simplified schematic diagram of radiative transitions that lead to emission in semiconductors containing impurities is presented in Fig. 4.14. Basic properties of these transitions will now be outlined.

In process 1, which is intraband transition, an electron excited well above the conduction-band edge dribbles down and reaches thermal equilibrium with the lattice. This *thermalization* process may lead to phonon-assisted photon emission or, more likely, phonon emission only.

In process 2, which is interband transition, direct recombination between an electron in the conduction band and a hole in the valence band results in the emission of a photon of energy $h\nu \cong E_g$; this transition produces *intrinsic luminescence.* Although this recombination occurs from states close to the corresponding band edges, the thermal distribution of carriers in these states will lead, in general, to a broad emission spectrum.

Process 3 is the *exciton* decay observable at low (cryogenic) temperatures. As discussed in the preceding section, it is possible at low (cryogenic) temperatures for electron–hole pairs to form a bound state, an exciton. Both free excitons and excitons bound to an impurity may undergo such transitions. In the case of bound excitons, one of the charge carriers is localized at a center that can assist in conserving momentum during the transition. (This is especially important in indirect-gap materials.) These transitions are often denoted with the following symbols: free-exciton recombination is denoted by X, recombination of an exciton bound at a neutral donor is D^0X, at a neutral acceptor is A^0X, and of excitons bound to the corresponding ionized impurities are D^+X and A^-X.

Processes 4–6 arise from transitions that start and/or finish on localized states of impurities (e.g., donors and acceptors) in the gap. Process 4 represents donor-to-free-hole transition labeled D^0h, process 5 represents free-electron-to-acceptor transition labeled eA^0, and process 6 represents the donor–acceptor pair (DAP) recombination model. These processes (i.e., 4–6) account for most of the processes in a wide variety of luminescent semiconductor materials. Similar transitions via deep donor and acceptor levels can also occur. The energy of the transition in such

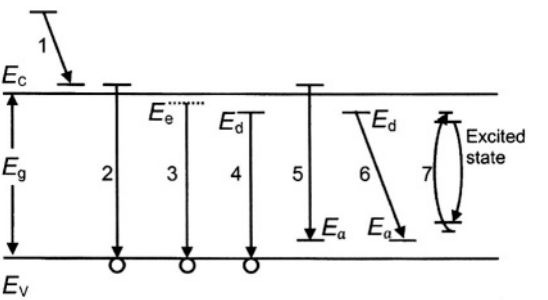

FIGURE 4.14. Schematic diagram of radiative transitions between the conduction band E_c, the valence band E_v, and exciton E_e, donor E_d, and acceptor E_a levels in a semiconductor.

cases is much smaller than that of the band-to-band transition. (It is important to note, however, that in wide energy-gap semiconductors, such transitions associated with deep levels may result in the emission of photons in the visible and near-infrared ranges.) In such impurity states, an electron is highly localized in space; in other words, its wave function extends only to the nearest neighbors. But if the uncertainty in the position Δx is small, the uncertainty in the momentum Δp must be large, as follows from Heisenberg's uncertainty principle ($\Delta x \Delta p \geq \hbar/2$). Therefore, since $p = \hbar k$, we can write $\Delta k \geq (2\Delta x)^{-1}$, indicating that the energy of a deep level extends over a wide range of k-values. Consequently, a direct transition from the impurity level to a wide range of extended states is allowed without the participation of phonons. This is especially relevant to indirect-gap semiconductors.

Transition 7 represents the excitation and radiative de-excitation of an impurity with incomplete inner shells, such as a rare earth ion or a transition metal.

The electron–hole pair recombination may occur via nonradiative processes as well. Nonradiative recombination processes include (i) multiple-phonon emission (i.e., direct conversion of the energy of an electron to heat), (ii) Auger effect (the energy of an electron transition is absorbed by another electron which is raised to a higher-energy state in the conduction band, with subsequent emission of the electron from the material or dissipation of its energy through emission of phonons), and (iii) recombination due to surface states and defects.

Thus, to summarize briefly, we can distinguish between radiative and nonradiative processes in semiconductors. Radiative processes include the processes described earlier, i.e., band-to-band electron–hole recombination, free exciton recombination, DAP recombination, bound exciton recombination at shallow centers and isoelectronic traps. Nonradiative processes include multiple-phonon recombination at deep centers, Auger recombination, surface recombination, and recombination at various defects (e.g., dislocations). At this juncture it should be noted that lattice defects could introduce localized levels in the energy gap. For example, dislocations may introduce both shallow levels due to the elastic strain fields and deep levels associated with dangling bonds. In addition, a wide variety of native point defects (e.g., vacancies and their complexes with impurity atoms) may also be present and may introduce a range of localized levels in the energy gap of binary semiconductors, such as GaAs. It should also be noted that in a degenerately doped semiconductor, in which the dopant concentration exceeds the value for which the wave functions of the shallow states begin to overlap, the energy levels broaden into a band. Such a band near, e.g., the top of the valence band may overlap with valence band states; the energy gap in such cases depends on the doping concentration, and the photon energy of the (previously) intrinsic emission will depend on the concentration of dopants.

As mentioned in Section 2.4, impurities and various defects have a major effect on both semiconductor properties and devices. Shallow impurities are often deliberately introduced to produce materials and devices with desired properties. For example, the electrical conductivity of semiconductors can be varied widely as a function of n- and p-type doping. Thus, the main effect of shallow impurities

(both donors and acceptors) is to control both the sign and magnitude of the electrical conductivity. As mentioned in the preceding sections, we consider impurity atoms in semiconductors as point defects if they are detrimental in the utilization of the material or device, but if the impurities are incorporated in the material to control electrical conductivity or optical properties, we refer to them as donors, acceptors, and recombination centers.

In some cases impurities give rise to deep levels in the energy gap of a semiconductor. In addition to the impurity-induced deep levels, various lattice defects may also give rise to bound states in the energy gap of the material. Such states due to these defects (i.e., vacancies, interstitials, antisite defects and their complexes, dislocations, stacking faults, grain boundaries, or precipitates) are usually located deeper in the energy gap and are more localized. These deep centers typically act as efficient traps and they control the carrier lifetime. In devices, such as solar cells, such deep centers are detrimental. Some deep centers in specific cases, however, are desirable as recombination centers in light-emitting devices. In general, extended defects such as dislocations and grain boundaries are detrimental in device applications, although in some cases these defects may be useful in gettering impurities and other defects from active regions of the device.

In general, the information in broad luminescence bands observed at elevated temperatures (above liquid nitrogen temperatures) is relatively difficult to elucidate. At low temperatures (e.g., liquid helium temperatures), the thermal broadening effects are minimized, and luminescence spectra in general become much sharper and more intense, allowing a more unambiguous identification of luminescence centers. The near-energy-gap emission (i.e., edge emission) at liquid helium temperatures is often resolved into emission lines, which can be due to excitons, free-carrier-to donor (or acceptor) transitions and their phonon replicas, and/or DAP lines.

There are two models of excitons. These are (i) strongly bound, closely localized excitons (i.e., the *Frenkel excitons)* and (ii) weakly bound excitons with a wave function spread over many interatomic distances (i.e., the *Wannier–Mott excitons).* The latter are typically present in materials with high dielectric constant. The energy levels can be described by a hydrogenic-like expression (See Fig. 4.15):

$$E_n = E_g - E_B/n^2 \tag{4.8.1}$$

where $n = 1, 2, 3, \ldots$ is the principal quantum number, and E_B is the exciton binding energy (Table 4.5), $E_B = m_r^* e^4/2\hbar^2\varepsilon^2$, where $m_r^* = m_e^* m_h^*/(m_e^* + m_h^*)$ is the reduced effective mass and ε is the dielectric constant. Frenkel and Wannier–Mott excitons are two limiting models differing in the degree of pair separation, and intermediate separations of electrons and holes are also possible. Note that at and below liquid nitrogen temperatures, the excitonic lines can be resolved from the lower-energy side of the energy-gap emission.

Phonon replicas are series of lines separated by a phonon energy $\hbar\omega$. Any mechanism giving a sharp emission line may be accompanied by these replicas.

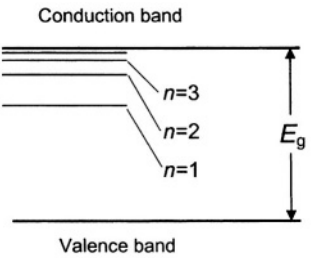

FIGURE 4.15. Schematic diagram of the energy levels of an exciton in a semiconductor.

As discussed in Chapter 2 (Section 2.5), in the phonon spectrum, there are transverse and longitudinal phonons in both the acoustic and optical branches. These phonons are denoted as TA (transverse acoustic), LA (longitudinal acoustic), TO (transverse optical), and LO (longitudinal optical).

In luminescence spectra, series of emission lines can also arise from DAP. In this case, an electron, captured by a donor, recombines with a hole captured by an acceptor. The energy of the donor–acceptor recombination emission depends on pair separation (see Fig. 4.16):

$$h\nu(r) = E_g - (E_A + E_D) + e^2/\varepsilon r \tag{4.8.2}$$

where E_A and E_D are the binding energies of the acceptor and donor, respectively, and ε is the dielectric constant. The term $e^2/\varepsilon r$ arises from the Coulombic interaction of the carriers and depends on the pair separation r having only values corresponding to integral numbers of interatomic spacing. This results in a fine structure consisting of sharp emission lines. For DAP recombination, two extreme cases are (i) the widely separated (or distant) DAP and (ii) associated (nearest neighbor) DAP. Since for distant DAP the pair separation is large, the term $e^2/\varepsilon r$ is small, resulting in the discrete lines forming a continuum (i.e., broad unresolved DAP bands). It should be noted that for the distant pair case, the static dielectric constant is used, and for the associated pair case, the optical dielectric constant is used. It should also be noted that for small pair separation r, an additional term, the van der Waals term, may become important, and thus must be included in Eq. (4.8.2).

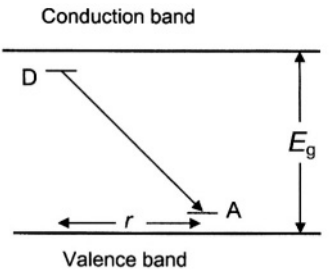

FIGURE 4.16. Schematic diagram of a donor-to-acceptor transition (r is the pair separation).

The donor–acceptor recombination rate depends on the pair separation r. The radiative transition probability in this case can be expressed as

$$P(r) = P(0)\exp(-2r/a) \tag{4.8.3}$$

where a is the Bohr radius of the less tightly bound center, and P(0), i.e., the limiting transition probability as $r \to 0$, is a constant for all pairs. A characteristic feature of DAP recombination is the shift of the peak energy as a function of the excitation intensity. This is a result of the reciprocal dependence of the peak energy on the pair separation r and the reduction in the transition probability with increasing r. At higher excitation intensities, widely separated pairs are saturated due to the lower transition probability, and a larger portion of pairs with smaller r are excited and decay radiatively because of their higher transition probability. Thus, a relative increase in the intensity due to pair transitions with smaller r is expected with increasing excitation intensity, resulting in a shift of the peak to higher energies. Also, a shift of the peak to lower energies can be observed as the delay after excitation in time-resolved measurements is increased. These features of DAP recombination emission allow one to distinguish it from other recombination processes.

4.8.2. Nonradiative Recombination Mechanisms

In semiconductors, the main recombination pathways between the conduction and valence bands involve donor and/or acceptor levels, recombination via deep-level traps, and recombination at the surface. The latter two recombination mechanisms are expected to be nonradiative.

The kinetics of the recombination processes, related to the minority carrier capture at defects having levels in the energy gap of the semiconductor, can be described by the *Shockley–Read–Hall (SRH) recombination model.* (This model is also referred to as trap-assisted recombination model.) According to this model, for a single trapping energy level present in the energy gap, the SRH recombination rate (under steady state nonequilibrium conditions) can be expressed as (see Sze 1981)

$$R_{\mathrm{SRH}} = \frac{N_{\mathrm{t}}\sigma_{\mathrm{n}}\sigma_{\mathrm{p}}v_{\mathrm{th}}(np - n_{\mathrm{i}}^2)}{\sigma_{\mathrm{n}}\{n + n_{\mathrm{i}}\exp[(E_{\mathrm{t}} - E_{\mathrm{i}})/k_{\mathrm{B}}T]\} + \sigma_{\mathrm{p}}\{p + n_{\mathrm{i}}\exp[(E_{\mathrm{i}} - E_{\mathrm{t}})/k_{\mathrm{B}}T]\}} \tag{4.8.4}$$

where σ_{n} and σ_{p} are the electron and hole capture cross-sections, respectively; N_{t} is the volume density of deep levels; v_{th} is the carrier thermal velocity; E_{t} is the energy level of the trap; and R has units of $\mathrm{cm}^{-3}\mathrm{s}^{-1}$. The term $(np - n_{\mathrm{i}}^2)$ in this equation indicates the deviation from the thermal equilibrium conditions, and for $np = n_{\mathrm{i}}^2$ (i.e., for the thermal equilibrium condition) the recombination rate $R_{\mathrm{SRH}} = 0$. For electron traps $\sigma_{\mathrm{n}} \gg \sigma_{\mathrm{p}}$, and for hole traps $\sigma_{\mathrm{p}} \gg \sigma_{\mathrm{n}}$, and if both the electron and hole traps act as nonradiative recombination centers (i.e., $\sigma_{\mathrm{n}} = \sigma_{\mathrm{p}} = \sigma$), the equation for the recombination rate can be written as

$$R_{\mathrm{SRH}} = \frac{N_t \sigma v_{th}(np - n_i^2)}{n + p + 2n_i \cosh[(E_t - E_i)/k_B T]} \tag{4.8.5}$$

From the analysis of the dependence of R_{SRH} as a function of $(E_t - E_i)$, it follows that the nonradiative recombination rate increases as E_t approaches E_i (i.e., the mid-energy gap), and the maximum recombination rate occurs at centers with energy levels located at or near the mid-energy gap. For a defect-related energy level located near the band edges (i.e., shallow levels), the thermal emission to the corresponding band results in reduction of the recombination process. Thus, to summarize briefly, the defect levels can be either a recombination center or a trap depending on E_i. In other words, the defect level is a trap if $|E_t - E_i|$ is large and E_t is located near either valence or conduction band. On the other hand, if E_t is located near mid-energy gap (i.e., E_i), it is an efficient recombination center. The definitions of the electron lifetime in a *p*-type semiconductor and of the hole lifetime in an *n*-type semiconductor, respectively, are as follows:

$$\tau_n^{\mathrm{SRH}} = \frac{1}{N_t \sigma_n v_{th}} \tag{4.8.6}$$

$$\tau_p^{\mathrm{SRH}} = \frac{1}{N_t \sigma_p v_{th}} \tag{4.8.7}$$

Another important mechanism of nonradiative recombination is that of *surface* (and *interface*) *recombination,* which can have a major effect on semiconductor devices. In general, the free surface of a semiconductor contains dangling bonds and impurities (from the ambient) that lead to (i) the formation of surface states (and corresponding energy levels in the energy gap of a semiconductor) and (ii) the band bending at the surface. These surface states act as nonradiative recombination centers. The semiconductor interfaces are also of great importance, since they have a major effect on minority carrier transport. An important procedure of the removal of dangling bonds (and associated interface states) is the *passivation.* One of the most technologically important interfaces is that formed between Si and SiO_2. Such Si/SiO_2 interface reduces the interface state density by orders of magnitude. The surface recombination process can be characterized by the *surface recombination velocity* $S = N_{st}\sigma_s v_{th}$. The surface recombination is one of the major sources of the degradation of the device performance of junction devices, and especially of the laser diodes. For some of the most important semiconductors, the surface recombination velocity is of the order of 10^3 cm s^{-1} (or less) for Si, and it is about 10^6 cm s^{-1} for GaAs.

Additional nonradiative recombination mechanism is due to the *Auger recombination* process, which involves three particles (see Section 7.6.1). In this case, the energy due to the recombination of an electron and a hole in a band-to-band transition is given to another free carrier, which is likely to be a majority (rather than a minority) carrier, since the probability of the energy transfer is proportional to the density of these free carriers. Thus, in this process, no

generation of photons is involved, and the energy transferred to another free carrier eventually dissipates through the emission of phonons. This recombination mechanism is often observed in the cases of high-injection levels in semiconductor lasers and light-emitting diodes.

4.8.3. Recombination Rate

As mentioned earlier, various excitations may lead to the generation of charge carriers in excess of the thermal equilibrium densities, and the recombination of electron–hole pairs restores that equilibrium. The rate R at which recombination occurs is proportional to the product of the concentration of occupied states (i.e., electrons) in the conduction band and the concentration of unoccupied states (i.e., holes) in the valence band. Thus, in general, we may write

$$R = B n_0 p_0 \tag{4.8.8}$$

where B is a constant and n_0 and p_0 are the equilibrium concentrations of electrons and holes. If Δn and Δp are the excess carrier concentrations, we can write

$$-\frac{d\Delta n(t)}{dt} = B[n_0 + \Delta n(t)][p_0 + \Delta p(t)] - B n_i^2 \tag{4.8.9}$$

Assuming that $n_0 p_0 = n_i^2$ and considering the case of low-level injection, i.e., the excess concentrations are small, for an extrinsic material (e.g., p-type), this equation can be simplified to

$$-\frac{d\Delta n(t)}{dt} = B p_0 \Delta n(t) \tag{4.8.10}$$

Thus, if the generation process ceases, the initial excess carrier concentration $\Delta n(0)$ will decay exponentially

$$\Delta n(t) = \Delta n(0)\exp(-B p_0 t) = \Delta n(0)\exp(-t/\tau_n) \tag{4.8.11}$$

where $\tau_n = (B p_0)^{-1}$ is called the *recombination lifetime* or *minority carrier lifetime.* Similarly, the decay of excess holes in n-type semiconductor occurs with decay constant, $\tau_p = (B n_0)^{-1}$.

In general, the description of such optical processes as the formation of the luminescence involves the analysis of the generation, diffusion, and recombination of minority carriers. Under equilibrium conditions, the generation of electron–hole pairs is balanced by recombination processes. These recombination processes may include radiative band-to-band and band-to-impurity recombination, and non-radiative recombination processes such as trap-assisted recombination, Auger recombination, and surface recombination.

As mentioned earlier, recombination centers with energy levels in the gap of a semiconductor are radiative or nonradiative, depending on whether the recom-

bination results in the emission of a photon or not. These centers are characterized by a recombination rate $R \propto 1/\tau$, where τ is a recombination time. Note that the carrier diffusion length $L = (D\tau)^{1/2}$, where D is the diffusion coefficient (see Section 4.11).

For both competitive radiative and nonradiative centers present, the observable lifetime is

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{4.8.12}$$

or

$$\tau = \frac{\tau_r \tau_{nr}}{\tau_r + \tau_{nr}} \tag{4.8.13}$$

where τ_r and τ_{nr} are the radiative and nonradiative recombination lifetimes, respectively. In general, τ_{nr} is the resultant of several nonradiative recombination processes that may be present in a given material, i.e.,

$$\frac{1}{\tau_{nr}} = \sum_i \frac{1}{\tau_{nri}} \tag{4.8.14}$$

The *radiative recombination efficiency* (or *internal quantum efficiency*) η, which is defined as the ratio of the radiative recombination rate R_r to the total recombination rate R, can be written as

$$\eta = \frac{\tau}{\tau_r} = \frac{1}{1 + (\tau_r/\tau_{nr})} \tag{4.8.15}$$

Note that, when $\tau_r \gg \tau_{nr}$, i.e., when the nonradiative recombination process is faster, radiative recombination efficiency is relatively small. On the other hand, when $\tau_{nr} \gg \tau_r$, i.e., when the radiative recombination process is faster, radiative recombination efficiency is relatively greater.

For a semiconductor that contains only one type of radiative and one of nonradiative recombination center, the radiative recombination efficiency η can be expressed as, using a relationship $\tau = (N\sigma v_{th})^{-1}$

$$\eta = \frac{\tau}{\tau_r} = \frac{1}{1 + (N_{nr}\sigma_{nr}/N_r\sigma_r)} \tag{4.8.16}$$

where N_r and N_{nr} are the densities of the radiative and nonradiative recombination centers, respectively, σ_r and σ_{nr} are the radiative and nonradiative capture cross-sections, and v_{th} is the carrier thermal velocity. It should be emphasized that, since the rate of luminescence emission is proportional to radiative recombination efficiency, in the observed emission intensities one cannot distinguish between radiative and nonradiative processes in a quantitative manner. (In general, η

depends on temperature, the particular dopants and their concentrations, and the presence of various defects.) In principle, it is possible in some cases to ensure that $N_r > N_{nr}$, but typically $\sigma_{nr} \gg \sigma_r$.

In semiconductors, the measured minority carrier lifetime, τ, in principle depends on various recombination mechanisms. For bulk semiconductors (including thick semiconductor layers), the minority carrier lifetime τ_{bulk} is

$$\frac{1}{\tau_{bulk}} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{4.8.17}$$

In semiconductors, one has also to include additional recombination mechanism, i.e., nonradiative surface (or interface) recombination. Thus, the measured values of the lifetime are effective values determined by the bulk and surface lifetimes

$$\frac{1}{\tau_{measured}} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} + \frac{1}{\tau_{surf}} \tag{4.8.18}$$

In general, whereas τ_{bulk} depends on the density of recombination centers in the bulk of the material, τ_{surf} is determined to a large extent by the recombination centers on the surface. Surface states arise due to the abrupt change at the surface of the three-dimensional band structure associated with the bulk of the material. In addition, impurity atoms and oxide layers (on the surface of the material) may also produce discrete energy levels. It should be noted that in pure crystals, τ_{bulk} is typically long, so that surface recombination will dominate. (It should be emphasized that in such a case, the lifetime of the material may depend strongly on surface treatment procedures.)

4.8.4. Luminescence Centers

As discussed earlier, the energy levels in the gap of a semiconductor (or insulator) can be categorized as *shallow* and *deep levels* according to their depth from the nearest band edges. (Note, however, that there is some arbitrariness involved in the definition of deep levels.) In general, deep levels have large capture cross-sections for carriers, and they are efficient recombination centers. In some cases, impurities and a variety of defects may introduce deep levels in the energy gap of a semiconductor. The information about these deep centers is crucial in the analysis of luminescence processes, since deep centers usually act as efficient recombination centers or traps and control the carrier lifetime.

Luminescence properties of semiconductors may also depend strongly on isoelectronic impurities. Such centers are formed by replacing one atom of the crystal by another atom from the same group of the periodic table. For example, the substitution of N for P in GaP leads to the formation of a deep localized acceptor level that can trap an electron, and subsequently, through the Coulombic attraction, it can also attract a hole. Thus, the isoelectronic trap is an efficient radiative recombination center.

In general, shallow-level centers are relatively simple, and they can be analyzed by employing the effective mass approximation. Another relatively well understood case is that of rare earth ion impurities and of transition-metal impurities that form deep states corresponding to basically sharp levels of the element. This is due to the fact that the electrons in the partially filled *f* and *d* states of rare earth or transition-metal impurities are screened from the surrounding matrix by the outer electrons of the ion, and, thus they experience only a slight perturbation that can be described by a crystal field. Between these two extremes of (i) shallow centers and (ii) the screened atomistic levels of rare earth and transition-metal impurities, occur deep levels due to impurities and a wide variety of defects (e.g., vacancies, interstitials, antisite defects and their complexes, dislocations, stacking faults, grain boundaries, or precipitates). However, no universal theory is available to account for all these cases.

In general, the probability of direct electron–hole recombination is small, and most recombination events occur through levels in the energy gap as illustrated in the example shown in Fig. 4.17. In this case, the electron–hole pairs are generated by an incident energetic photon or electron. The carriers dribble down to the band edges within about 10^{-11}–10^{-12} s and reach thermal equilibrium with the lattice by the most likely process of emission of phonons. The carriers (both in the conduction and valence bands) can move through the crystal until they are trapped temporarily. At sufficiently high temperatures, an electron can be re-excited back to the conduction band by acquiring sufficient energy from the lattice vibrations. The probability of such an escape is given by the Boltzmann factor $P \propto \exp(-E_t/k_B T)$, where E_t is the depth of the trap. In this expression, the coefficient of proportionality is given by $N\sigma v_{th}$, where N is DOS in the band, σ is the capture cross-section (i.e., a measure of the effectiveness of the localized level in capturing excess carriers passing nearby), and v_{th} is the carrier thermal velocity. At room temperature, electrons can be trapped from seconds to days, depending on E_t. A carrier (e.g., an electron) may eventually encounter a recombination center. The distinction between traps and recombination centers is as follows. The center is a trap, if following the capture of a carrier of one type, the probability

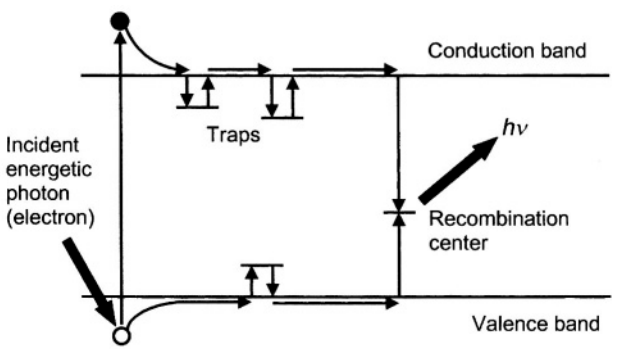

FIGURE 4.17. Schematic diagram of excitation, trapping, and recombination mechanisms in luminescence with trapping levels for both electrons and holes.

of re-excitation is higher than the capture of a carrier of the opposite type. On the other hand, if the capture of a carrier of one type is followed by a capture of a carrier of the opposite sign, it is predominantly a recombination center. To summarize briefly, for a trap, the capture cross-section of one type of a carrier is much larger than that of the opposite type, whereas a recombination center has a large capture cross-section for both types of carriers.

For a case of only one type of recombination center present, the carrier lifetime is $\tau = (N\sigma v_{th})^{-1}$, where N in this case is the density of recombination centers (generally in the range between about 10^{12} and 10^{19} cm^{-3}). Depending on the nature of the Coulomb interaction, the range for capture cross-section is between about 10^{-25} cm^2 (for Coulomb repulsive capture) and 10^{-12} cm^2 (for Coulomb attractive capture). For a carrier thermal velocity of the order of 10^7 cm s^{-1}, the lifetime τ of the carriers can thus vary from about 10^{-14} s to tens of hours.

4.9. SPONTANEOUS AND STIMULATED EMISSION

As discussed earlier, in luminescence processes the recombination of individual excited carriers occurs randomly. The excited carriers in such processes remain in their excited state for a characteristic time, i.e., the *carrier lifetime,* prior to encountering and recombining with a carrier of opposite charge. This is the *spontaneous recombination* (i.e., no external effect triggers the emission). This random process produces emitted light (i) that travels in all directions and (ii) with the waves being out of step with each other (i.e., *incoherent emission* of light).

The concept of *light amplification by stimulated emission of radiation* (LASER) involves the *stimulated emission* in a system with two states having energies E_1 and E_2 ($E_2 > E_1$) and in the presence of electromagnetic radiation. In this process, an excited electron is stimulated to de-excite, i.e., the incoming photon (of appropriate energy $h\nu = E_2 - E_1$) increases the probability that the electron returns to the ground state. This results in the emission of a second photon that is exactly in-phase with the incident photon and having the same energy $h\nu$. The additional photons in this process may cause other excited carriers to de-excite in a similar manner, leading to the build up of an intense *coherent emission* of light (i.e., all the waves, moving in the same direction, are exactly in step, and add together constructively as they are in-phase). Eventually, this process dominates the spontaneous emission rate in the laser.

LASER depends on *population inversion,* i.e., obtaining more electrons in the excited state than in the ground state. In many cases, the population inversion is realized by using the *optical pumping* (i.e., intense irradiation of the active medium with an external source). In semiconductor lasers, the population inversion is established by using an electrical current flow through the active medium of the device.

The process of stimulated emission requires (i) an active medium (e.g., gas or solid), (ii) a source of external excitation energy (e.g., photons), and (iii) forming the optical cavity. This results in the monochromatic and coherent light with very low beam divergence (i.e., collimated or focused beam of high intensity).

It should be noted that the excited state in such processes must in fact correspond to a metastable state (i.e., a state with sufficiently long lifetime for stimulated emission to occur before spontaneous emission). In addition, as mentioned earlier, the emitted photons must be confined in the optical cavity for sufficient time, so that to permit these photons to stimulate additional emission from the excited states.

The relationship between spontaneous and stimulated processes, for a system with two states with energies E_1 and E_2 ($E_2 > E_1$) and in the presence of electromagnetic radiation with energy density $\rho(\nu)$, are analyzed in terms of *Einstein coefficients,* A_{21}, B_{12}, and B_{21}, which represent three possible transitions between these states, i.e., they characterize the rates of *spontaneous emission, absorption,* and *stimulated emission,* respectively, between two energy levels E_1 and E_2. (For schematic illustration of these processes, see Fig. 4.18.) The transitions are as follows:

(a) Spontaneous emission, E_2 to E_1, with the rate proportional to the population (number of atoms) N_2 at E_2, so that the transition rate in this case is $A_{21}N_2$.
(b) Stimulated absorption, E_1 to E_2, with the rate proportional to the population (number of atoms) N_1 at E_1 and to the number of photons with energy $h\nu = E_2 - E_1$, so that absorption rate in this case is $B_{12}N_1\,\rho(\nu)$.
(c) Stimulated emission, E_2 to E_1, with the transition rate in this case being $B_{21}N_2\,\rho(\nu)$.

At thermal equilibrium, the transition rates for spontaneous and stimulated emission and the transition rate for absorption are balanced, i.e.,

$$B_{12}N_1\,\rho(\nu) = A_{21}N_2 + B_{21}N_2\,\rho(\nu) \tag{4.9.1}$$

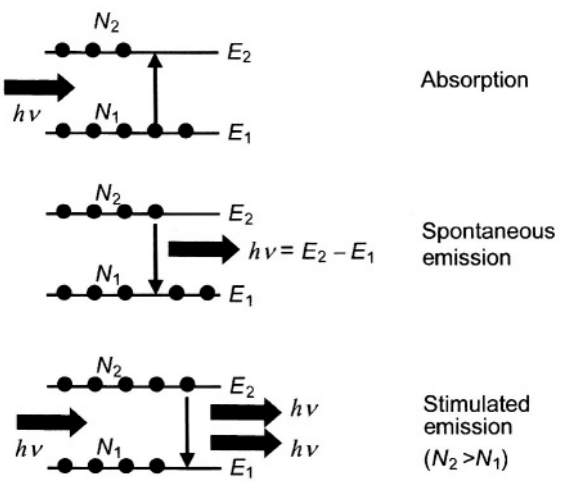

FIGURE 4.18. Schematic illustration of types of transition processes for a two-level system. For stimulated emission, the inversion condition ($N_2 > N_1$) results in the generation of amplified coherent beam.

Einstein relations show that coefficients of stimulated absorption and emission are equal:

$$B_{12} = B_{21} \tag{4.9.2}$$

and that the coefficient of spontaneous emission, A_{21}, is related to coefficients of induced (or stimulated) processes as

$$A_{21} = B_{21} \frac{8\pi n_r^3 h\nu^3}{c^3} \tag{4.9.3}$$

where h is Planck's constant, c is the speed of light in vacuum, and n_r is the refractive index.

The relationship between absorption and spontaneous emission can be derived from the Van Roosbroeck–Shockley principle of detailed balance, which allows calculating the shape of the emission band from the experimentally determined values of the absorption coefficient. The relationship between the equilibrium emission intensity $L(h\nu)$ at a photon energy $h\nu$ and the absorption coefficient $\alpha(h\nu)$ for an intrinsic semiconductor can be expressed as

$$L(h\nu) = \frac{8\pi n_r^2 h^2 \nu^2 \alpha(h\nu)}{h^3 c^2 [\exp(h\nu/k_B T) - 1]} \tag{4.9.4}$$

where n_r is the refractive index, c is the speed of light, and h is Planck's constant.

4.10. EFFECTS OF EXTERNAL PERTURBATIONS ON SEMICONDUCTOR PROPERTIES

Electrical and optical properties of semiconductors, and the band structure in general, are strongly dependent on various external perturbations, such as temperature and applied fields (e.g., electric and magnetic fields).

In general, there are two main causes of the temperature dependence of band states in any semiconductor. First, the band structure is a function of the crystal lattice spacing, so lattice dilation may be expected to contribute to a change in the energy position of band states. The second contribution is due to electron–phonon interactions. The total coefficient of the temperature dependence of band states range from about 10^{-4} to 10^{-3} eV K^{-1}. The dilation contribution, which can be calculated from the thermal expansion and pressure coefficients, amounts to about 20–50% of the total temperature dependence of band states.

For practical applications, the temperature dependence of the energy gap of many semiconductors can be fitted by the empirical expression

$$E_g(T) = E_g(0) - bT^2/(T + \theta) \tag{4.10.1}$$

where $E_g(0)$ is the energy gap at 0 K, and b and θ are constants. This expression represents satisfactorily experimental data for many semiconductors. For example, for GaAs, $E_g(0) = 1.519$ eV, $b = 5.405 \times 10^{-4}$ eV K^{-1}, and $\theta = 204$ K. The use of these values in Eq. (4.10.1) gives the dependence shown in Fig. 4.19.

The application of the electric field on a semiconductor produces changes in its electrical and optical properties, which are of great importance in various optoelectronic applications, such as optical modulators. These include the effects related to (i) changing the absorption of the incident light in the presence of the applied electric field (referred to as electroabsorption) in the bulk materials (i.e., *Franz–Keldysh effect*) and quantum well structures (i.e., *quantum-confined Stark effect*) and (ii) changing the refractive index, i.e., *electro-optic effects* (*Pockels effect* and *Kerr effect*).

The electric-field-induced change of interband optical absorption is described by the Franz–Keldysh effect. In this case, with the application of an electric field, the local band structure changes in such a way that the conduction and valence band edges become spatially tilted. This enables the electronic states to tunnel into the energy gap, resulting in the possibility of the optical absorption to occur for photon energies lower than the energy gap energy. Thus, to summarize briefly, the states in the conduction and valence bands, which are separated by the energy gap, in the presence of the electric field overlap due to the tunneling into the energy gap. In bulk semiconductors, this results in an increase of the absorption coefficient below the energy gap and the *Franz–Keldysh oscillations* above the energy gap. These oscillations are due to the alternating constructive and destructive overlap of the oscillatory part of the electron and hole wave functions. The absorption below the energy gap increases with increasing electric field (due to the increasing penetration of the envelope wave functions into the energy gap), and so does the magnitude and period of the Franz–Keldysh oscillations. As a consequence, the absorption edge, in the presence of the electric field, appears to be shifted towards lower energies; however, to be more precise, the absorption edge is broadened and not shifted.

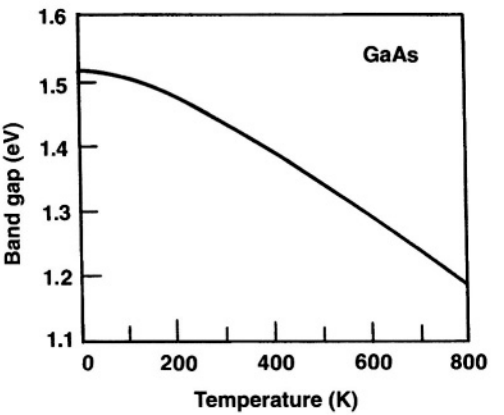

FIGURE 4.19. The energy gap as a function of temperature for GaAs. The curve was calculated from Eq. (4.10.1) using $E_g(0) = 1.519$ eV, $b = 5.405 \times 10^{-4}$ eV K^{-1}, and $\theta = 204$ K.

The absorption edge broadening, or tailing, is exponential and its extent increases with increasing electric field. Various electro-optic effects, such as the Franz–Keldysh effect in bulk materials can be employed in electro-optical waveguide switches and modulators, which are main components for the signal processing in fiber-optic communication technology. The electroabsorption modulator, based on the Franz–Keldysh effect, employs the fact that due to the steep absorption edge in direct energy-gap semiconductor (e.g., GaAs), significant variations in the near-band-edge absorption can be induced by the electric-field application.

The quantum-confined Stark effect relates to the shift of the exciton absorption peaks in quantum wells in the presence of electric fields that are applied perpendicularly to the quantum well layers. (With increasing fields, the absorption edge shifts to longer wavelengths). In such a case, the high potential energy barriers due to the walls of the quantum wells preclude the field ionization of the exciton, which can still be preserved in the presence of high electric fields. Note that, in order to ensure a relatively short distance between the electron and hole, the quantum well size should be much smaller as compared to the bulk excitonic diameter. Typically, the exciton can be confined in such a quantum well in the presence of high electric fields of up to about $10^5\ \mathrm{V\,cm^{-1}}$ at room temperature. Thus, large Stark shifts can be achieved.

The electro-optic effect relates to the change in the refractive index in the presence of an electric field. The optical modulators based on this effect typically operate by employing changes in the beam's optical path length and the corresponding variations in its direction, reflection or transmission. There are two types of electro-optic effects, i.e., the linear (Pockels) effect and the quadratic (Kerr) effect. The linear (Pockels) electro-optic effect occurs in a material that does not have a center of symmetry (e.g., $LiNbO_3$). In such a case, the change in refractive index Δn_r caused by the applied electric field $\mathcal{E}$ can be expressed as $\Delta n_r = (n_r^3 r_{eff}\,\mathcal{E})/2$, where r_{eff} is an effective electro-optic coefficient that depends on the sample orientation and the beam polarization. The quadratic (Kerr) electro-optic effect is dominant in materials that have a center of symmetry (such materials do not exhibit linear effect).

4.11. BASIC EQUATIONS ON SEMICONDUCTORS

This section summarizes some of the basic equations that are useful for the analysis and modeling of semiconductors and semiconductor devices.

4.11.1. Poisson's Equation

This equation describes the dependence of the electric field on the space charge density (denoted as ρ)

$$\nabla \cdot \boldsymbol{\mathcal{E}} = \rho/\varepsilon \tag{4.11.1}$$

where $\rho = e(N_d^+ - N_a^- - n + p)$.

4.11.2. Continuity Equations

In semiconductors at nonequilibrium conditions, for processes such as generation and recombination and carrier transport, the carrier densities within a given unit of volume of the material varies as a function of time.

The *continuity equations* for the carriers describe changes in carrier densities with time in terms of (i) the incoming and outgoing flux of carriers, (ii) the carrier generation (G), and (iii) the carrier recombination (R). The current continuity equations for the carriers, i.e., electrons and holes, respectively, are

$$\frac{\partial n(x,t)}{\partial t} = \frac{1}{e}\nabla \mathbf{J}_{\mathrm{n}}(x,t) + G(x,t) - R(x,t) \tag{4.11.2}$$

$$\frac{\partial p(x,t)}{\partial t} = -\frac{1}{e}\nabla \mathbf{J}_{\mathrm{p}}(x,t) + G(x,t) - R(x,t) \tag{4.11.3}$$

where $\mathbf{J}_{\mathrm{n}}$ and $\mathbf{J}_{\mathrm{p}}$ are the electron and hole current densities, respectively, under steady state conditions.

4.11.3. Carrier Transport Equations

In semiconductors, the transport of carriers is diffusive for nonuniform distribution of carriers. The diffusion current density (for electrons) in this case is

$$\mathbf{J}_{\mathrm{diffusion}} = eD_{\mathrm{e}}\nabla n \tag{4.11.4}$$

where D is the diffusion coefficient.

In the presence of an electric field $\mathcal{E}$, the total current density will be

$$\mathbf{J} = \mathbf{J}_{\mathrm{drift}} + \mathbf{J}_{\mathrm{diffusion}} \tag{4.11.5}$$

Thus, for electron and hole current densities, respectively, the carrier transport equations consist of two components, i.e., the drift component due to the electric field and the diffusion component due to the carrier concentration gradient

$$\mathbf{J}_{\mathrm{n}} = ne\mu_{\mathrm{e}}\mathcal{E} + eD_{\mathrm{e}}\nabla n \tag{4.11.6}$$

$$\mathbf{J}_{p} = pe\mu_{\mathrm{h}}\mathcal{E} - eD_{\mathrm{h}}\nabla p \tag{4.11.7}$$

An equation that links carrier mobility μ with the diffusion process (through the diffusion constant D) is the *Einstein relation*

$$D = \mu\frac{k_{\mathrm{B}}T}{e} \tag{4.11.8}$$

4.12. SUMMARY

Intrinsic semiconductors are materials whose properties are native to the material; it often implies an undoped semiconductor. The concept of a *hole* is related to an unoccupied state in the valence band, and it can be regarded as a positive charge carrier that can contribute to the conduction process. *Extrinsic semiconductors* are those whose properties are influenced or controlled by intentionally added impurity atoms, or by the presence of impurities and/or defects; it often refers to a doped material. *Dopants* are specific impurity atoms that are intentionally added to a semiconductor in controlled amounts in order to increase either the electron or the hole concentration. Among dopants, *donors* are impurity atoms which increase the electron concentration, i.e., *n*-type dopant, whereas *acceptors* are impurity atoms which increase the hole concentration, i.e., *p*-type dopant. In general, *n-type semiconductor* is a donor-doped material, or a semiconductor containing more electrons than holes, and *p-type semiconductor* is an acceptor-doped material, or a semiconductor containing more holes than electrons. The *majority carriers* are the most abundant carriers in a given semiconductor sample, i.e., electrons in*n*-type, and holes in *p*-type semiconductor; and *minority carriers* are the least abundant carriers in a given semiconductor sample, i.e., holes in *n*-type, and electrons in *p*-type semiconductor.

The determination of both the carrier concentrations and energy distributions in semiconductors is of great importance in elucidating their electrical and optical properties. This involves such concepts as the *probability of carrier occupancy* of a state at energy E (i.e., *Fermi–Dirac occupation statistics*) and the *density of states,* or DOS that represents the number of available electronic states per unit energy and per volume.

In nondegenerate semiconductors (at equilibrium), the product of the electron and hole density, i.e., the *np* product, depends on the effective conduction band and valence band densities of states, the energy gap of a semiconductor, and the temperature, and it is independent of the Fermi level position and of the individual electron and hole concentrations.

The basic semiconductor equations include those related to (i) the dependence of the electric field on the space charge density (i.e., *Poisson's Equation*), (ii) changes in carrier densities with time (i.e., the *continuity equations*), and (iii) the transport equations consisting of the drift component and the diffusion component (i.e., *carrier transport equations*).

PROBLEMS

4.1. Find the position of the Fermi level in a semiconductor, wherein the probability that a state is occupied in the conduction band is equal to the probability that a state is empty in the valence band.

4.2. Find the position of the Fermi level (relative to E_i) at 300 K in Si doped with 10^{16} cm^{-3} donors and 10^{15} cm^{-3} acceptors that are fully ionized.

4.3. Explain why the absorption coefficient in indirect-gap semiconductors is generally lower as compared to the direct-gap materials.

4.4. Describe the procedure for distinguishing in luminescence spectra the features due to excitons and those due to the donor–acceptor recombination.

This page intentionally left blank

5

Applications of Semiconductors

5.1. INTRODUCTION

One of the principal utilities of semiconductors in a wide range of *electronic devices* (i.e., devices that employ the transport properties of carriers in the material) and *optoelectronic devices* (i.e., devices for the generation and detection of light) is related to their capability to form various electrical junctions and resulting electrostatic inhomogeneities and built-in electric fields. These junctions include (i) a *p–n homojunction* (between two semiconductor regions of opposite doping types), (ii) a *metal–semiconductor junction* (or *Schottky barrier*), and (iii) a *heterojunction* (formed between two dissimilar semiconductors that are joined together by deposition or epitaxial growth). Electronic devices based on such junctions are typically employed as rectifying elements (*p–n diodes*), or as parts in various transistors, or in a wide range of light emitting and light detecting devices. Among these, a *p–n* junction is the most important one for both microelectronic and optoelectronic device applications. This chapter will outline the basic characteristics of a *p–n* junction, as well as of other types of junction and semiconductor devices (for extensive discussion on various semiconductor junctions and their applications in various electronic devices, see books in Bibliography Section B2).

5.2. DIODES

5.2.1. The p–n Junction

A *p–n* junction is formed between two semiconductor regions of opposite doping types. Such junctions formed within a single semiconductor are referred to as homojunctions. For the purpose of illustration, this can be depicted by bringing two oppositely doped regions together and aligning their conduction and valence band energies (see Fig. 5.1). (Note that typical methods of the actual formation of a *p–n* junction, which include diffusion, ion implantation, or epitaxial growth, are

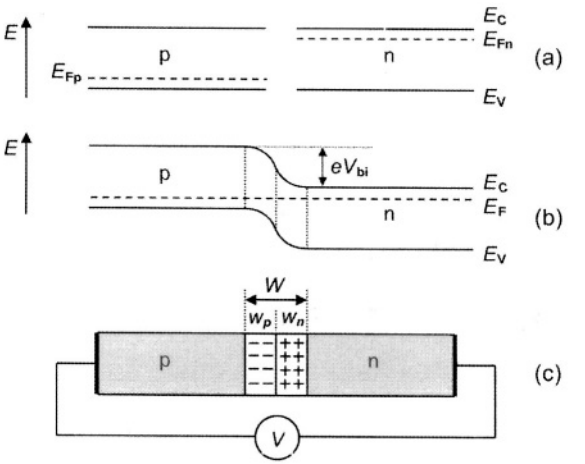

FIGURE 5.1. Schematic illustration of the *p–n* junction: (a) the energy diagrams of a *p*- and *n*-type semiconductors prior to junction formation, (b) the energy diagram after the junction is formed (in thermal equilibrium), and (c) a *p–n* junction showing the depletion region (or space–charge region).

described briefly further.) It is assumed that (i) the *n*- and *p*-doped regions are uniformly doped and (ii) the transition between the two regions is abrupt. (Such a structure is referred to as an *abrupt p–n junction.*) The electrons (in *n*-type region) and holes (in *p*-type region) close near the junction diffuse across it. The electrons diffuse into the *p*-type region, whereas the holes diffuse into the *n*-type region, and encountering each other, electrons and holes recombine, thus leading to the formation of a region (around the junction) that is depleted of mobile carriers. This region is called the *depletion region* (*W*). As a result of this diffusion of electrons and holes across the junction, the immobile ionized donors and acceptors in *n*- and *p*-type regions are no longer compensated, resulting in the formation of space–charge regions near the junction (see Figs. 5.1 and 5.2). This is associated with the concentrations of acceptor (N_a) and donor (ons, which produce a net negative charge on the *p*-type side of the junction and a net positive charge on the *n*-type side, respectively. Such a build-up of oppositely charged regions results in the formation of the junction potential, which effectively prevents further migration of free carriers. Any free carrier, entering the depletion region, is experiencing a force that pushes it back away from the depletion region (i.e., the depletion region is kept free of charge carriers). In other words, the presence of uncompensated charge due to the ionized donors and acceptors produces an electric field that results in a drift of carriers in the opposite direction. The *internal* (*built-in*) *potential* is essentially formed as the result of the Fermi energy difference between the *n*- and *p*-type regions. In the thermodynamic equilibrium (no voltage bias is applied), the Fermi level (which is near the valence band edge in a *p*-type region and near the conduction band in an *n*-type region) of a *p–n* junction must be constant across the junction, which necessitates the band bending through

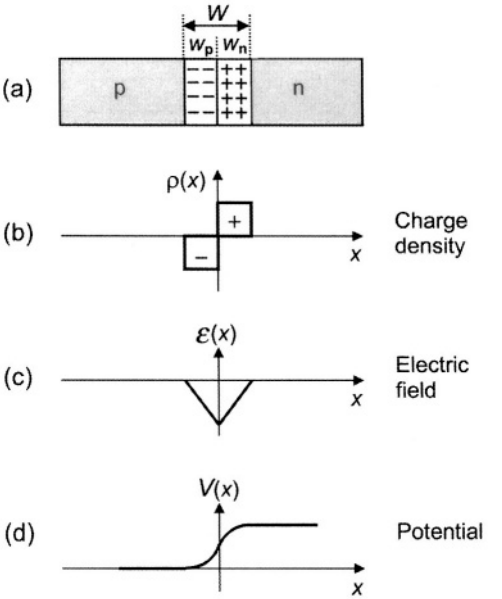

FIGURE 5.2. Schematic illustration of the p–n junction (for $N_a = N_d$) in equilibrium: (a) a p–n junction showing the depletion region (or space–charge region), (b) space–charge distribution, (c) electric field distribution, and (d) potential across junction.

the junction [see Fig. 5.1(b)]. The equilibrium built-in voltage V_{bi} is related to the difference in the Fermi levels (of the two semiconductors) prior to their equalization, i.e.,

$$eV_{bi} = E_{Fn} - E_{Fp} \tag{5.2.1}$$

In order to move across the depletion region, free charge carriers require extra energy to overcome the forces of the space–charge region. In other words, the junction behaves like a barrier for charge flow. Such a barrier is depicted in Fig. 5.1(b) as band bending of the conduction and valence bands in the depletion region. Such a depiction represents the condition of the electrons that now have to "move uphill" in order to traverse across the depletion region from the n-type side to the p-type side. The opposite is true for holes. Thus, free charge carriers require energy to traverse across the depletion region; this can be accomplished by the application of a voltage between the two ends of the p–n junction diode [see Fig. 5.1(c)]. However, depending on polarity, the application of such a voltage may either assist in overcoming the barrier, or vice versa. This effectively results in a rectifying characteristic of a diode, which allows the flow of electrical current in one direction but not in the another. This depends on the *forward-bias* or *reverse-bias* conditions of such a diode. If a voltage is applied to the junction [see Fig. 5.1(c)], it is referred to as *forward biased* when a positive voltage is applied to the p-doped region, and it is *reversed biased* when a negative voltage is applied to

the *p*-doped region. Thus, to summarize briefly, the general performance of a *p–n* junction, related to the application of the external bias, can be understood in terms of the *forward-bias* and *reverse-bias* conditions. For forward-bias conditions, the free electrons and holes are pushed towards the junction, thus providing them with additional energy to traverse the junction. Whereas, for reverse-bias conditions, the electrons and holes are pulled away from the junction, making it more difficult for them to traverse the depletion region. This essentially implies that in the forward bias, the potential barrier is lowered, whereas in the reverse bias, the barrier is raised (see Fig. 5.3). Such a performance of the *p–n* junction is employed in many electronic device applications.

Typical methods of the formation of a *p–n* junction include diffusion, ion implantation, or epitaxial growth. In the case of diffusion, the suitable dopant (in sufficient concentration) is diffused (using heat) in the appropriate region, and this results in the formation of a junction. One can also employ ion implantation of, e.g., *n*-type semiconductor with acceptor ions, which results in sharp junctions. (Note that the bombardment with high-energy ions during the implantation process also induces damage to the crystal structure, i.e., it results in the generation of various defects.) It is also possible to employ epitaxial deposition techniques, which allow the formation of various layers of semiconductors together with required dopants included during the growth. Using such techniques allows fabrication of very abrupt junctions (i.e., with no counter-doping in corresponding regions).

As shown in Fig. 5.2, the space–charge density [see Fig. 5.2(b)] changes sharply near the edges of the depletion region. Such a depiction is referred to as the *depletion approximation*, which essentially defines two separate regions, i.e., (i) a depletion region with negligibly low carrier densities and (ii) quasi-neutral (homogeneous)

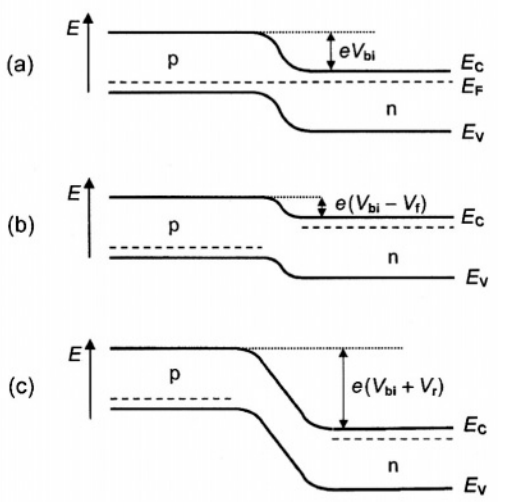

FIGURE 5.3. Schematic illustration of the energy band diagrams of a *p–n* junction (a) in equilibrium, (b) under forward bias, and (c) under reverse bias.

region within which the charge density is assumed to be zero [see Fig. 5.2(b)]. (Note, however, that in practical cases, the carrier densities do not abruptly fall to zero at the edges of the depletion region on both sides, resulting in both the carriers and the space charge having a distribution at the depletion region edge.)

From the analysis of the flow of charge carriers across the *p–n* junction and from the charge neutrality condition for the abrupt junction, one can derive analytical relationships for the diode in equilibrium. In this case, the total current is zero (i.e., the net current across the junction due to both electrons and holes is zero), as the diffusion current and drift current are equal and opposite and, thus, they cancel each other out. Note that (for both types of carrier) the diffusion current is from the *p*-side to the *n*-side, and the drift current is from the *n*-side to the *p*-side. The conditions of the diffusion and drift components cancelling out at equilibrium can be expressed for the *x*-direction as (see Section 4.11)

$$e\mu_e n(x)\mathcal{E}(x) + eD_e \frac{dn(x)}{dx} = 0 \tag{5.2.2}$$

$$e\mu_h p(x)\mathcal{E}(x) - eD_h \frac{dp(x)}{dx} = 0 \tag{5.2.3}$$

The electric field $\mathcal{E}(x) = -dV(x)/dx$, where $V(x)$ is the potential and $n(x)$ and $p(x)$ are the carrier densities at a distance x from the junction. Thus, these equations can be written as

$$-e\mu_e n(x)\frac{dV(x)}{dx} + eD_e \frac{dn(x)}{dx} = 0 \tag{5.2.4}$$

$$e\mu_h p(x)\frac{dV(x)}{dx} + eD_h \frac{dp(x)}{dx} = 0 \tag{5.2.5}$$

For electrons (as an example), Eq. (5.2.4.) can be expressed as

$$\mu_e \frac{dV(x)}{dx} = D_e \frac{1}{n(x)} \frac{dn(x)}{dx} \tag{5.2.6}$$

By integrating this equation over the proper limits (related to the depletion region widths from both sides of an abrupt junction) and by using the Einstein relation ($D = \mu k_B T/e$), one can obtain the following equation:

$$V_n - V_p = \frac{k_B T}{e} \ln \frac{n_n}{n_p} \tag{5.2.7}$$

where V_n and V_p correspond to the potential on each side of the junction, and n_n and n_p correspond to the electron concentration at the each edge of the depletion region. Noting that the potential difference $V_n - V_p = V_{bi}$, and to a good

approximation $n_n = N_d$, and [using Eq. (4.4.9): $np = n_i^2$] $n_p = n_i^2/N_a$, the built-in voltage (or barrier voltage) can be related to doping concentrations as

$$V_{bi} = \frac{k_B T}{e} \ln \frac{N_a N_d}{n_i^2} \tag{5.2.8}$$

Thus, Eq. (5.2.8) relates V_{bi} to the given semiconductor parameters, i.e., it depends on the doping of the p- and n-regions, on temperature, and the energy gap E_g [see Eq. (4.4.11): $n_i = (N_c N_v)^{1/2} \exp(-E_g/2k_B T)$ that relates n_i and E_g].

Using Poisson's equation (see Section 4.11), which describes the dependence of the electric field on the space–charge density ρ, i.e., $\nabla \cdot \mathcal{E} = \rho/\varepsilon$, where $\rho = e(N_d^+ - N_a^- - n + p)$, one can derive the electric field distribution within the depletion region. For any point along the x-direction, the electric field gradient can be related to the local space charge as

$$\frac{d\mathcal{E}(x)}{dx} = \frac{e}{\varepsilon}(N_d^+ - N_a^- - n + p) \tag{5.2.9}$$

By neglecting the contributions from n and p within the depletion region, this equation is simplified to (assuming complete ionization)

$$\frac{d\mathcal{E}(x)}{dx} = \frac{e}{\varepsilon} N_d \tag{5.2.10}$$

$$\frac{d\mathcal{E}(x)}{dx} = -\frac{e}{\varepsilon} N_a \tag{5.2.11}$$

for two depletion regions w_n and w_p, respectively (see Fig. 5.1). These equations essentially indicate that within the depletion region the electric field distribution $\mathcal{E}(x)$, which is directed from the n-side to the p-side, has a positive slope on the n-side and a negative slope on the p-side (as shown in Fig. 5.2) with a maximum value $\mathcal{E}_{max}$ of the field at the junction (i.e., $x = 0$) and reaching zero at the boundaries of depletion region (i.e., x_n and $-x_p$ for n- and p-sides, respectively). By integrating Eq. (5.2.10) or (5.2.11) over the proper limits (related to the depletion region boundaries of an abrupt junction), $\mathcal{E}_{max}$ can be related to the built-in voltage V_{bi}. Thus, for the n-side (i.e., for $0 < x < x_n$), one can write

$$\int_{\mathcal{E}_{max}}^{0} d\mathcal{E} = \frac{eN_d}{\varepsilon} \int_0^{x_n} dx \tag{5.2.12}$$

and for the p-side (i.e., for $-x_p < x < 0$), one can write

$$\int_0^{\mathcal{E}_{max}} d\mathcal{E} = -\frac{eN_a}{\varepsilon} \int_{-x_p}^{0} dx \tag{5.2.13}$$

From these equations, one can derive $\mathcal{E}_{max}$:

$$\mathcal{E}_{max} = -\frac{eN_d w_n}{\varepsilon} = -\frac{eN_a w_p}{\varepsilon} \tag{5.2.14}$$

Since the electric field $\mathcal{E} = -dV(x)/dx$, where $V(x)$ is the potential, one can relate the electric field with the built-in potential V_{bi} from

$$-V_{bi} = \int_{-x_p}^{x_n} \mathcal{E}(x)\, dx \tag{5.2.15}$$

and,thus

$$-V_{bi} = \frac{\mathcal{E}_{max}(x_n + x_p)}{2} \tag{5.2.16}$$

or using Eq. (5.2.14) and the fact that $x_n + x_p = W$ (note that essentially x_n and x_p correspond to w_n and w_p), one can write

$$V_{bi} = \frac{eN_d w_n W}{2\varepsilon} = \frac{eN_a w_p W}{2\varepsilon} \tag{5.2.17}$$

From the condition of charge neutrality (i.e., the total negative charge in the p-side depletion region exactly balances the total positive charge in the n-side depletion region, see Fig. 5.1), one can write

$$w_p N_a = w_n N_d \tag{5.2.18}$$

where w_p and w_n are the widths of the p-side and n-side charged regions, respectively. In addition, one can express the total depletion width W as

$$W = w_p + w_n \tag{5.2.19}$$

Thus, one can write

$$w_p = W\frac{N_d}{N_a + N_d} \tag{5.2.20}$$

$$w_n = W\frac{N_a}{N_a + N_d} \tag{5.2.21}$$

The depletion width W (at equilibrium) can be related to doping concentrations and the built-in voltage as

$$W = \left(\frac{2\varepsilon V_{bi}}{e}\ \frac{N_a + N_d}{N_a N_d}\right)^{1/2} \tag{5.2.22}$$

The expressions for the widths w_p and w_n are

$$w_p = \left(\frac{2\varepsilon V_{bi}}{e} \frac{N_d}{N_a(N_a + N_d)}\right)^{1/2} \tag{5.2.23}$$

$$w_n = \left(\frac{2\varepsilon V_{bi}}{e} \frac{N_a}{N_d(N_a + N_d)}\right)^{1/2} \tag{5.2.24}$$

From the condition of charge neutrality ($w_p N_a = w_n N_d$), one can also conclude that for a specific case, e.g., for $N_a \gg N_d$, one obtains that $w_n \gg w_p$. This indicates that in such a case, the much greater fraction of the depletion region occurs on the *n*-side of the junction, i.e., in the side of the junction with lower doping. In other words, the actual magnitude of W is largely determined by the doping concentration of the lower-doped side of the junction, which basically determines the *p–n* junction characteristics. Such a structure (i.e., when one side of the junction has significantly greater doping concentration than the another side) is referred to as a *one-sided abrupt p–n junction.*

In general, the knowledge about the depletion width W is vital, since (i) it determines the limit on the dimensions of the diode, and (ii) in practical applications of a reverse-biased diode, the depletion width may also determine its breakdown voltage.

For the case of a biased diode, the expression for the depletion width has to include the applied voltage V (which can be either positive or negative for forward or reverse bias, respectively):

$$W = \left(\frac{2\varepsilon(V_{bi} - V)}{e} \frac{N_a + N_d}{N_a N_d}\right)^{1/2} \tag{5.2.25}$$

From this expression, it follows that W is increased (decreased) under reverse (forward) bias conditions (see Fig. 5.3).

Another fundamental consequence of the analysis of the net flow of charge carriers across the *p–n* junction (i.e., analysis of diffusion and recombination of charge carriers) is its current–voltage characteristic, which can be expressed as

$$I = I_0[\exp(eV/k_B T) - 1] \tag{5.2.26}$$

This equation (referred to as the *diode equation*) describes the characteristic of the *p–n* junction which is referred to as *rectification.* (This characteristic is applied to a wide range of applications.) The current across the diode increases exponentially with the application of a forward-bias voltage, whereas the current is limited to a relatively small saturation value with the application of reverse-bias voltage (see Fig. 5.4). Note that this equation and $I(V)$ characteristic in Fig. 5.4 represent an ideal case (in practical devices, however, deviations from this characteristic may occur). The saturation current I_0 depends on various parameters related to the minority carriers as

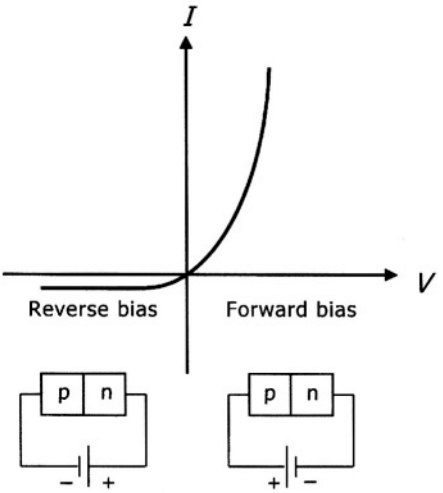

FIGURE 5.4. Schematic representations of the current–voltage curve of a p–n junction diode and of the corresponding circuits.

$$I_0 = eA\left(\frac{D_e n_p}{L_e} + \frac{D_h p_n}{L_h}\right) \tag{5.2.27}$$

where A is the cross-sectional area of the junction, D_e and D_h are the diffusion constants of electrons and holes, L_e and L_h are the diffusion lengths of electrons and holes, and n_p and p_n are the equilibrium minority carrier concentrations of electrons in p-region and holes in n-region, respectively. (Note that the diffusion constant $D = \mu k_B T/e$.) As mentioned earlier, the relationship between the minority carrier diffusion length and the diffusion constant is $L = (D\tau)^{1/2}$, where τ is the minority carrier lifetime. Using Eq. (4.4.9), i.e., $np = n_i^2$, and also recalling that for *shallow impurities* at room temperature almost the entire donor or acceptor sites are ionized and the free carrier density corresponds to the impurity concentration, the above expression for I_0 can be expressed in terms of dopant concentrations. Note that in the case of donors, the electron density n equals the concentration of donors (i.e., $n \cong N_d$), and in the case of acceptors, the density of holes p equals the concentration of acceptors (i.e., $p \cong N_a$). Thus, $n_p = n_i^2/p_p = n_i^2/N_a$, and $p_n = n_i^2/n_n = n_i^2/N_d$. Accordingly, I_0 can be expressed as

$$I_0 = en_i^2 A\left(\frac{D_e}{L_e N_a} + \frac{D_h}{L_h N_d}\right) \tag{5.2.28}$$

or alternatively, by using the expression $L = (D\tau)^{1/2}$,

$$I_0 = en_i^2 A\left[\frac{1}{N_a}\left(\frac{D_e}{\tau_e}\right)^{1/2} + \frac{1}{N_d}\left(\frac{D_h}{\tau_h}\right)^{1/2}\right] \tag{5.2.29}$$

These equations indicate the dependence of the reverse saturation current on the properties of the minority carriers. Thus, I_0 decreases as the carrier lifetimes and/or the doping concentrations increase.

Due to the charge separation in the depletion region (i.e., the presence of two layers of space charge in the depletion region), it behaves like a capacitor, and the (junction) capacitance can be expressed as

$$C = A\left[\frac{e\varepsilon N_a N_d}{2(N_a + N_d)}\right]^{1/2} \frac{1}{(V_{bi} - V)^{1/2}} \tag{5.2.30}$$

Thus, a *p–n* junction capacitance can be varied by the applied voltage; this property can be employed in electronic circuits. Such junction devices that are employed for their voltage-controlled variable capacitance are referred to as *varactor diodes.*

As mentioned earlier, Eq. (5.2.26) for $I(V)$ characteristic (see Fig. 5.4) represents an ideal case. However, in practical devices, various breakdown phenomena and deviations from the ideal characteristic may occur. The basic breakdown mechanisms include *Zener* (or *tunnel*) *breakdown* and *avalanche breakdown.* In the case of Zener breakdown, which typically occurs in heavily doped diodes at low reverse voltages, the heavy doping results in a very narrow barrier width, and the valence band electrons that are at sufficiently short distances from empty states in the conduction band, tunnel through the barrier. In the case of avalanche breakdown, which occurs at high reverse voltages, the carriers, gaining sufficient energy in the junction electric field, cause ionizing collisions that produce additional electron–hole pairs which may result in an uncontrolled current flow and rapid increase in the reverse current.

5.2.2. Schottky Barrier

The semiconductor diode structures can also be obtained by joining two dissimilar materials of high purity. For example, bringing a metal into a contact with a semiconductor (e.g., by depositing a metal onto a semiconductor) can form a metal–semiconductor diode structure (a *Schottky barrier*). Depending on the relative values of the work functions of the metal and a semiconductor and on the type of a semiconductor (i.e., *n*-type or *p*-type), one can obtain either rectifying junctions or ohmic contacts. From the equilibrium criterion of the Fermi level equalization across the junction, it follows that if the work function of the metal $e\Phi_m$ is greater than that of *n*-type semiconductor $e\Phi_s$, the flow of electrons from the semiconductor into a metal occurs. This results in the formation of a depletion layer (with a positive space charge) in a semiconductor near the junction and the accumulation of a negative surface charge on a metal (see Fig. 5.5). This is accompanied by the upward bending (within a depletion region in a semiconductor) of the energy band edges and the formation of a potential barrier having a height of $eV_{bi} = e(\Phi_m - \Phi_s)$, which prevents the electron diffusion from the conduction band of a semiconductor into the metal (see Fig. 5.5). Note that

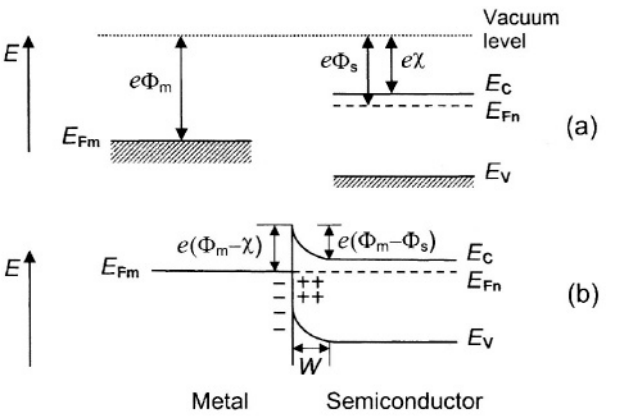

FIGURE 5.5. Schematic diagram of the band structure of metal–(*n*-type) semiconductor junction formation (for $\Phi_m > \Phi_s$) (a) before and (b) after contact; $e\Phi_m$ and $e\Phi_s$ are the work functions of the metal and of a semiconductor, respectively, and $e\chi$ is the electron affinity of a semiconductor.

the potential energy barrier height for electron movement from the metal into the conduction band of a semiconductor can be expressed as $e\Phi_B = e(\Phi_m - \chi)$. The band bending is analogous to the case of a *p–n* junction described earlier. For the *n*-type semiconductor, with the positive voltage applied to the semiconductor, the barrier height is increased, impeding the flow of electrons from semiconductor to metal (this corresponds to reverse-bias conditions). For forward-bias conditions (i.e., the negative voltage applied to the semiconductor), the barrier height is reduced, and there is a substantial electron flow from semiconductor to metal. Thus, such a metal–semiconductor junction acts as a rectifier, which can be described by the diode equation, i.e., similar to the case of a *p–n* junction (see Eq. 5.2.26). On the other hand, if the work function of the metal $e\Phi_m$ is smaller than that of *n*-type semiconductor $e\Phi_s$, the flow of electrons from the metal into a semiconductor and the Fermi level equalization across the junction result in the downward bending (in a semiconductor) of the energy band edges. In this case, in the absence of a potential barrier at the junction, the electrons can move unimpeded across the junction for both polarities of the bias voltage. In other words, in this case an ohmic contact is formed. Using analogous considerations for the case of *p*-type semiconductors, the junction is rectifying if the work function of the metal $e\Phi_m$ is smaller than that of *p*-type semiconductor $e\Phi_s$, whereas the junction is ohmic if the work function of the metal $e\Phi_m$ is greater than that of *p*-type semiconductor $e\Phi_s$. In general, the formation of metal–semiconductor junctions is somewhat unpredictable, since these types of junctions are greatly affected by the surface states that can cause band bending.

5.2.3. Heterojunctions

The junctions formed between two semiconductors with different energy gaps are referred to as heterojunctions. In general, heterostructures offer a wide range of

design choices for novel semiconductor devices (e.g., diodes, transistors, and optoelectronic devices). The main advantages are related to the control of the charge carrier transport by controlling the energy barriers and potential variations (on a quantum level) and to the ability of heterostructures to confine the optical radiation (which is especially essential in optoelectronic devices). In many applications, devices incorporate more than one heterojunction, and in such cases these are referred to as a *heterostructure.* In an ideal heterojunction, the interface is atomically abrupt. (In practical cases, such an ideal structure is nearly realized in $Al_xGa_{1-x}As/GaAs$ heterojunctions.)

The basic types of such junctions are the following:

(a) *Isotype* heterojunctions, i.e., both semiconductors having the same conductivity type, resulting in *n–n* or *p–p* junctions,

(b) *Anisotype* heterojunctions, i.e., semiconductors having different conductivity type, resulting in *p–n* heterojunctions,

(c) Heterojunctions, in which one of the semiconductors is doped and the another one undoped; in such a case, a doped material provides carriers, whereas the undoped material has higher carrier mobility due to the absence of ionized impurity scattering. In fact, the *field effect transistor* (FET), discussed below in Section 5.3.2, which incorporates such junctions (between a wider energy-gap doped semiconductor and a narrower energy-gap undoped material, e.g., doped AlGaAs and undoped GaAs) can be produced, and it is referred to as *modulation doped field effect transistor* (MODFET), or alternatively as a *high electron mobility transistor* (HEMT).

Note that, as outlined in Chapter 6, in order to obtain high-quality heterostructures (which are typically produced by employing epitaxial growth techniques), it is important to ensure that the crystal structures and the lattice constants of dissimilar materials are matched as close as possible in order to minimize the *mismatch strain* and avoid the formation of defects (i.e., *misfit dislocations*), which are detrimental in device applications. However, high-quality heterojunctions can also be realized in pseudomorphic (strained layer) structures, in which one of the semiconductors is sufficiently thin, so that no misfit dislocations at the interface are generated and the lattice mismatch between the dissimilar semiconductors is accommodated by strain.

As in the other cases of junctions described above, the basic property for elucidating the performance of a heterojunction is the energy band profile of the conduction and valence band edges as a function of position. Due to the difference in the energy gap of the constituent semiconductors, the most important consideration in the formation of a heterojunction is the energy band alignment, which results in discontinuities (in both the conduction and valence bands) that affect the properties of heterojunctions.

It should be noted that there are several theories of the band alignment. A major issue, related to the band gap discontinuities, is whether they are determined by the bulk properties of the constituent semiconductors, or are affected by the interface properties. The earlier electron-affinity model (which was

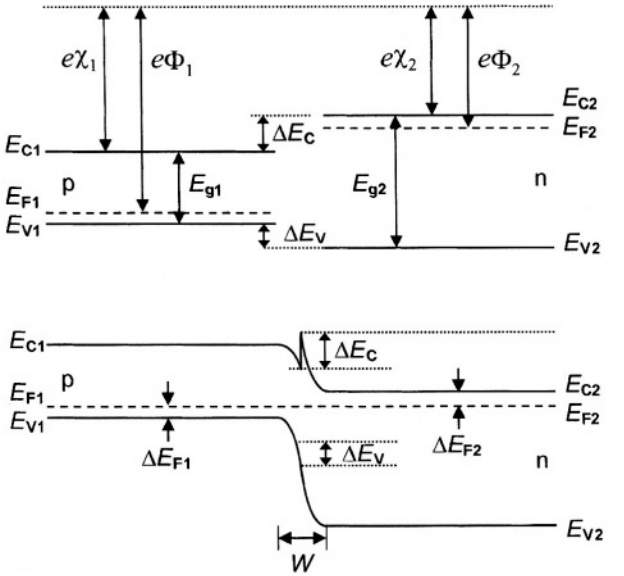

FIGURE 5.6. Schematic illustration of the p–n heterojunction: (a) the energy band diagrams of a p- and n-type semiconductors prior to junction formation and (b) the energy band diagram after the junction is formed (in thermal equilibrium).

subsequently revised) proposes the determination of the conduction band discontinuity at the interface between two dissimilar semiconductors from the difference between their electron affinities, i.e.,

$$\Delta E_c = e(\chi_1 - \chi_2) \tag{5.2.31}$$

This procedure is demonstrated for an anisotype heterojunction, consisting of a narrower-energy-gap p-type semiconductor and a wider-energy-gap n-type material, in Fig. 5.6. (However, note again that for the heterointerface, the electron affinity may not have a straightforward relation to the bulk magnitudes.)

The contact between the semiconductors results in the diffusion of electrons into the p-type region and holes into the n-type region, until the Fermi levels equalize at equilibrium. This results in the formation of a *depletion region* and in upward (downward) band bending in the n-side (p-side) region. Thus, the contact between such dissimilar materials in relation to their different electron affinities and equalizing their Fermi levels results in the formation (at the heterojunction) of discontinuities in the conduction band (ΔE_c) and the valence band (ΔE_v) with a discontinuous spike arising in the conduction band. The valence band discontinuity can be expressed as

$$\Delta E_v = (e\chi_2 + E_{g2}) - (e\chi_1 + E_{g1}) \tag{5.2.32}$$

and the energy-gap discontinuity ΔE_g can be written as

$$\Delta E_g = E_{g2} - E_{g1} = \Delta E_c + \Delta E_v \tag{5.2.33}$$

The built-in potential in this case can be written as

$$eV_{bi} = E_{g1} + \Delta E_c - \Delta E_{F1} - \Delta E_{F2} \tag{5.2.34}$$

The heterojunction, outlined above, is also referred to as a *single heterostructure.* Another important structure, which found a wide range of applications in various semiconductor devices, is based on two heterojunctions, and it is referred to as a *double heterostructure.* Typically, such a double heterostructure includes one *p–n* heterojunction and an isotype heterojunction. Such structures are typically composed of a (wider-energy-gap *p*-type)/(narrower-energy gap *p*-type)/(wider-energy-gap *n*-type) structure. In this case, under forward bias, the injected carriers from the wider-energy-gap *n*-type semiconductor into narrower-energy-gap *p*-type material are confined within that layer by the potential barrier of a heterojunction formed between narrower-energy-gap *p*-type material and a wider-energy-gap *p*-type semiconductor. For the case of the injected electron density being greater than the hole density in the narrower-energy-gap *p*-type semiconductor, charge neutrality requirement necessitates the injection of holes from the wider-energy-gap *p*-type semiconductor into narrower-energy-gap *p*-type material. Such a corresponding injection of carriers is referred to as *double injection.* Such a process can be effectively utilized in light emitting devices that require high radiative recombination rates.

5.3. TRANSISTORS

5.3.1. Bipolar Junction Transistors

The *p–n* junctions are also used as basic parts in various transistors that are extensively employed in (i) computer technology (which is based on the ability of transistors to operate as fast on–off switches), or (ii) for amplification of a current or voltage. In general, the two main types of transistors are *bipolar junction transistors* (BJTs) and *field effect transistors* (FETs). (For detailed discussion on different types of these transistors, see e.g., Streetman 1995; Sze 1981).

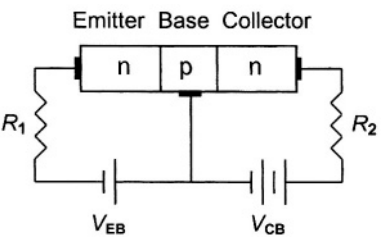

FIGURE 5.7. Schematic illustration of a common base *n–p–n* BJT.

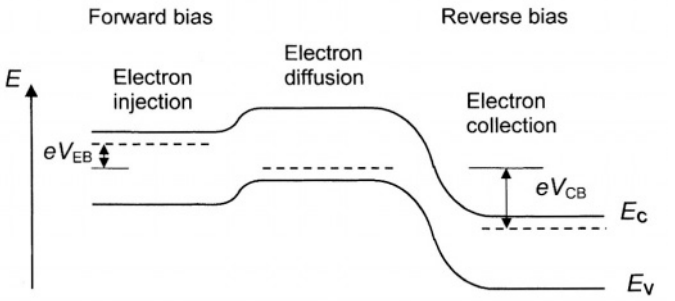

FIGURE 5.8. Schematic energy band diagram of a common base *n–p–n* BJT under normal bias conditions.

BJTs consist of two *p–n* junctions in a back-to-back configuration (i.e., either *p–n–p* or *n–p–n* junction configuration). These are three-terminal minority-carrier devices, in which a central region separating two junctions is referred to as the *base* (which is typically very narrow and lightly doped region) that is common to the input (referred to as the *emitter*) and output (i.e., *collector*) circuits. In BJTs, the term *bipolar* implies that two types of carriers are involved in its operation. In such devices (e.g., *n–p–n* transistor, see Fig. 5.7), one of the junctions (between the emitter and the base) is normally forward biased, so that the electrons can be injected from the emitter to the base (see Fig. 5.8). Whereas, the junction between the base and the collector is normally reverse biased, and thus the carriers reaching that junction are swept across into the collector region. For circuit connections, BJTs require two input and two output terminals; but having three transistor terminals, one of them has to be common to the input and output loops. Thus, for the BJT, there are three possible configurations for circuit connection: the common base, common emitter, and common collector. In the common–base configuration (see Fig. 5.7), the voltage gain can be obtained, and it is essentially determined by the ratio of R_2/R_1, where R_1 and R_2 are the input and output

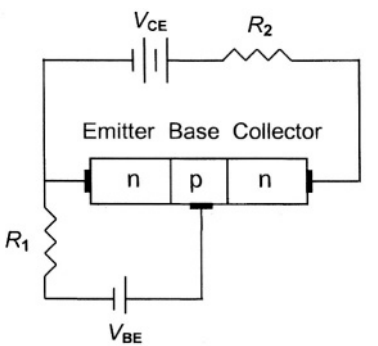

FIGURE 5.9. Schematic illustration of a common emitter *n–p–n* BJT.

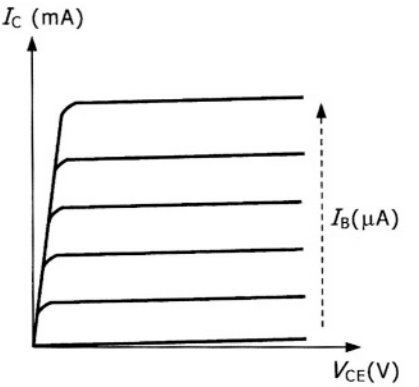

FIGURE 5.10. Current–voltage characteristic of a common emitter *n–p–n* BJT, showing typical output characteristics, i.e., I_C as a function of V_{CE} for different values of I_B.

(or load) resistors, respectively. For $R_2 \gg R_1$, the voltage gain can be on the order of several hundreds. In the common-emitter connection configuration (which is the most typically employed configuration), shown in Fig. 5.9, a current gain can also be obtained. The typical current–voltage characteristics of such a transistor is presented in Fig. 5.10. In general, BJTs can be employed as an amplifier or a switch (depending on the external circuit configuration). In the case of a switch, the emitter–collector currents can be turned fast "on" or "off" depending on changes in base voltage.

As mentioned above, for the operation of the BJTs, the junction between the emitter and the base is normally forward biased, so that the electrons can be injected from the emitter to the base. The doping in the emitter region in such cases (i.e., *n–p–n* BJTs) is much greater than the doping in the base, resulting in negligible injection of holes from the base into the emitter. However, the relatively low doping of the base region results in high series resistance of the base, which consequently limits high frequency applications. In heterojunctions (see Section 5.2.3), the flow of carriers (both electrons and holes) can be individually controlled due to the formation of different conduction band and valence band discontinuities at the heterointerface. Thus, in a *heterojunction bipolar transistor* (HBT), the doping in the base region can be increased in a structure incorporating a wider-energy-gap emitter and a narrower-energy-gap base, such as an $Al_xGa_{1-x}As/GaAs$ heterojunction. The structure of a transistor in such a case is *n* ($Al_xGa_{1-x}As$)–*p* (GaAs)–*n* (GaAs). Another heterojunction system that can be employed in an HBT is Si/$Si_{1-x}Ge_x$ (note that the energy gap of Si is greater than that of $Si_{1-x}Ge_x$).

5.3.2. Field Effect Transistors

Another type of a transistor is based on FETs, such as a *metal-oxide-semiconductor field effect transistor* (MOSFET), which is relatively easier to fabricate (as compared to BJTs) and which allows one to realize higher packing

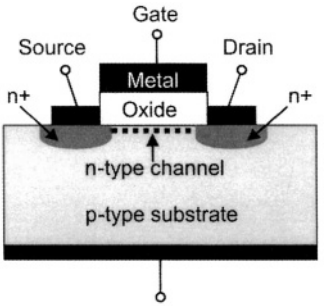

FIGURE 5.11. Schematic illustration of an enhancement-mode MOSFET. Application of a positive gate voltage induces a thin *n*-channel (inversion layer) in the *p*-type semiconductor along the surface (between the source and drain).

density in *integrated circuits* (ICs) (see further). Since the operation of MOSFET involves one type of carrier only, it is also referred to as *unipolar* transistor. In addition, since the carriers involved are the majority carriers, the transistor operation is less influenced by external factors. The operation of MOSFET (see Fig. 5.11) is based on controlling the flow of current by a voltage applied to a *gate* electrode, which provides the function analogous to that of the base in a BJT. The control of the current through the transistor is maintained by employing the electric field induced by the gate, which is isolated electrically from the rest of the transistor (i.e., from source and drain) by a gate insulator (e.g., a native SiO_2 in the case of Si). By applying a positive gate voltage, electrons are attracted to the region below the gate and form a thin *n*-type channel (i.e., an inversion layer) in the *p*-type semiconductor along the surface (adjacent to the oxide) between the source and drain. This allows the current flow through the channel. Thus, the gate voltage controls the resistance of the channel, i.e., the MOSFET can function as a voltage-controlled variable resistor. To summarize briefly, in the absence of an applied gate voltage, electrons cannot pass from the source to the drain, and in this condition the MOSFET is in its "off" state. By applying a positive voltage to the gate, negative charges build up in the *p*-type semiconductor at the interface between the semiconductor and oxide. This results in the formation of a conductive pathway between the n^+-type regions (i.e., source and drain) and the flow of electrons from the source to the drain. In this case, the MOSFET is in its "on" state. Such a device is also referred to as an *enhancement-mode* MOSFET, i.e., it is normally in an "off" state, and with the application of a positive voltage to the gate it is in an "on" state. Another type of such transistors is a *depletion-mode* MOSFET, which is normally in an "on" state, and with the application of a negative voltage to the gate (resulting in the constriction of the *n*-channel), the device is in an "off" state. It should be emphasized that one of the great advantages of the MOSFET is its input resistance, which is the largest as compared to other types of transistor. MOSFETs, which are extensively employed in ICs, can be produced by forming on opposite sides of the gate, e.g., two heavily doped n^+-regions (acting as the

source and *drain*) in a *p*-type substrate (typically made of silicon). Such a structure is referred to as an *n*-type channel MOSFET (or NMOS). In a *p*-type channel MOSFET (or PMOS), p^+-regions (acting as the *source* and *drain*) are formed in an *n*-type substrate.

As mentioned earlier (in Section 5.2.3), the FET, based on heterojunctions (between wider energy-gap doped semiconductor and narrower energy-gap undoped material) can be effectively employed in a MODFET, or as it is also referred to as a HEMT.

5.4. INTEGRATED CIRCUITS

The IC typically consists of large numbers of such circuit elements as resistors, capacitors, diodes, and transistors, which are combined in many different configurations that are designed to carry out various functions, e.g., digital data storage and amplification. In general, there are two types of ICs: a *monolithic IC* and a *hybrid IC.*

In a monolithic IC, all the circuit components are fabricated next to each other (and interconnected) into (or on top) a single semiconductor chip by using various device fabrication steps (see Section 6.1). In the wafer fabrication process, each wafer may contain several hundred identical ICs (i.e., chips), with each of them containing millions of submicron circuit elements (in *very large-scale integration,* or VLSI). The finished wafer is diced into individual chips for the final steps of *assembly* and *packaging* (including attachment, making electrical connections, and protection against moisture and contaminants). The final product, which can be readily handled, is connected to a circuit board.

A hybrid IC typically consists of separate components (including monolithic ICs), which are attached on an insulating substrate with other circuit elements and are electrically interconnected.

The great utility of such an integration (e.g., in monolithic ICs) is in the ability of joining large numbers of various circuit elements on a single segment of a semiconductor. This also facilitates a mass production of such ICs.

It should be noted at this juncture that one of the major efforts in microelectronic applications of semiconductors is directed toward continuous miniaturization, i.e., the reduction of the size of circuit elements, such as transistors. This tendency is motivated by the facts that increasingly smaller transistors will provide faster switching capability and, hence, shorter processing time, and they would also facilitate incorporation of more complex processors and an increasing number of their components per unit area. However, there are certain limits to such a miniaturization, since in principle, there is a fundamental limit for the size of a conventional transistor. Conventional digital computing employs a transistor to switch electrical current flow “on” or “off”. However, when it is miniaturized beyond the scale, when the smallest features of a transistor are reduced to less than about 50 nm, it will no longer function as a switching device, since quantum-mechanical effects, such as tunneling of electrons through potential barriers, will begin to dominate the operation of a conventional device. Specifically, continuous

miniaturization of the MOSFET devices to sizes down to about 100 nm can be realized with conventional design features. However, below that size several factors begin to hinder further miniaturization. Some of the effects that limit the scalability include (i) *tunneling* that limits packing density of devices, (ii) increased *electric field* due to voltages applied over shorter distances, which may result in avalanche breakdown in the transistor due to high-energy electrons, (iii) increasing *heat dissipation* in ICs, (iv) *shrinking depletion regions* (ability to block current from the source to the drain, for a transistor off-state, continuously deteriorates at the scale of less than about 50 nm), and (v) *statistical dopant concentration fluctuations* in nanoscale structures. The novel transistor designs such as *multi-gate transistors* (e.g., double-gate MOSFET) may overcome some of these problems. Such devices provide better control of the leakage current, and therefore allow further transistor miniaturization. A specific device structure, referred to as a *vertical replacement-gate transistor,* having sub-50-nm gate-length, can be manufactured with conventional fabrication methods. In this structure, the channel is enclosed by a double-gate, facilitating current flow along both vertical surfaces of a rectangular semiconductor section. This offers increased current flow and faster operation. Other examples of the potential nanoelectronic devices are *single electron devices* and *resonant tunneling devices* (see Section 6.11).

5.5. LIGHT EMITTING AND DETECTING DEVICES

5.5.1. Light Emitting Devices

Some of the vital components in photonics communications technology are the semiconductor light emitting and detecting devices, such as *light-emitting diodes* (LEDs), *diode lasers,* and *photodetectors,* which can be made sufficiently small and very efficient for various applications. (In the literature, the light emitting and detecting devices are generally described by a term *optoelectronics,* which refers to the technology related to the generation of light, amplification and control of light, and detection of light.) Such devices can be produced in a variety of types and configurations. Basically, an LED is a semiconductor *p–n* junction diode, which employs radiative recombination of injected minority carriers with the majority carriers and subsequent emission of light from the forward-biased junction. This effect is also referred to as an *injection electroluminescence,* i.e., the conversion of electrical energy into light. In this process, electrons injected in the conduction band fall into the valence band and release the extra energy which is emitted as a photon. In such a process, photons are emitted in random directions, hence it is referred to as incoherent light. As noted in Chapter 1, it is the *energy gap* of a semiconductor that determines the energy (or wavelength) of the emitted photon (see also Section 4.8), and the availability of a wide variety of semiconductors with appropriate energy gaps makes devices suitable for the emission of light in the desired wavelength ranges (see Table 6.4).

The same process is used in semiconductor *diode lasers.* (Laser is a light source generating coherent and near-monochromatic light by employing stimulated

emission of radiation.) The diode laser differs from an LED in (i) the operating current (i.e., much greater current for achieving *optical gain*) and (ii) having opposite ends cleaved parallel to each other. (This results in the formation of aligned mirrors that reflect the generated light back and forth for achieving amplification of light and generation of *stimulated emission.*) In this process, conduction-band electrons, which are stimulated by photons, fall into the valence band, and photons emitted as a result of this stimulated emission constitute coherent light, i.e., (i) the emitted photons have the same energy as the incident photons and (ii) they are in step with each other. To summarize briefly, a diode laser is an LED having mirrors on two opposite surfaces that form the *laser cavity.* In such a forward-biased diode, electrons that are injected into the active region of the junction recombine with holes, generating a spontaneous emission of photons, which in turn cause the recombination of additional electron–hole pairs that results in stimulated emission. As compared to the LEDs, the semiconductor diode lasers typically provide light output with narrower wavelength range, higher powers, and more directionality.

Some examples of semiconductors that are employed in optoelectronic applications are Si, Ge, GaAs, InP, InAs, GaN, and ternary (e.g., $Al_xGa_{1-x}As$, $GaAs_{1-x}P_x$) and quaternary (e.g., $Ga_xIn_{1-x}As_yP_{1-y}$) alloys with adjustable characteristics such as the *energy gap*, which is the principal parameter that determines the wavelength of the emitted or absorbed electromagnetic radiation (and thus determines the range of applications of a given semiconductor). The III–V compound semiconductors, such as GaAs, InP, GaP, InAs, GaN, and their corresponding ternary and quaternary alloys play especially important role in optoelectronic applications. This is mainly due to the wavelength range at which these materials emit and absorb light efficiently. For example, $Ga_xIn_{1-x}As_yP_{1-y}$ is suitable for light emission in the range between about 1300 and 1700 nm (i.e., within the most important range in fiber-optic communication).

In general, besides the emission wavelength range, the LEDs are characterized by the direction of emission as well. Thus, depending on device geometry (as well as internal structure), the LED types are distinguished between the *surface-emitting* and the *edge-emitting* types that are shown in Fig. 5.12, which illustrates homojunction diode structures. In practice, such devices are often based on heterostructures, which have several advantages that are outlined below.

It should be emphasized that light-emitting devices, such as LEDs and diode lasers, are superior compared to other sources of optical radiation, since they

(a) provide radiation with higher efficiency;
(b) have longer operational life times;
(c) generally require lower voltages;
(d) emit radiation in a very narrow wavelength range, which can be tuned in a wide range of wavelengths.

The LED efficiency (or the external conversion efficiency) η_{ext} basically depends on three factors: (i) the injection efficiency, (ii) the recombination efficiency, and (iii) the extraction efficiency. In an asymmetric p–n^+ junction diode

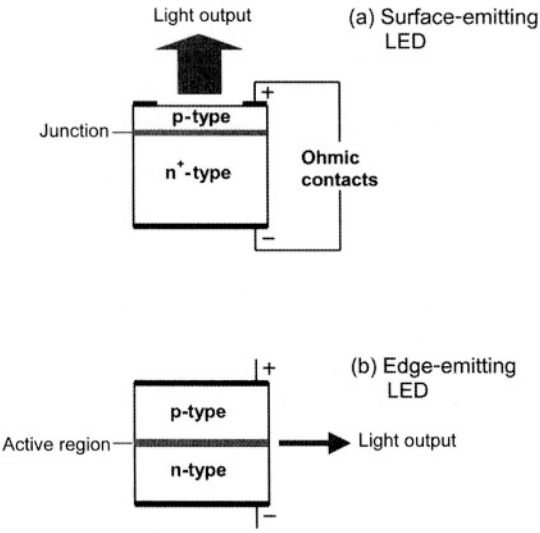

FIGURE 5.12. Schematic diagram of (a) surface-emitting LED, and (b) edge-emitting LED. Note that light is actually emitted in different directions, but typically the manner of a device packaging allows light output in one direction only.

(i.e., the electron injection is dominant as compared to hole injection), the injection efficiency η_{inj} can be derived from junction theory by evaluating the fraction of the total diode current due to diffusion of electrons that are injected into the *p*-side of the junction, and it can be related to semiconductor parameters as

$$\eta_{inj} = \frac{1}{(\mu_h L_e N_a)/(\mu_e L_h N_d) + 1} \tag{5.5.1}$$

which demonstrates that high η_{inj} can be obtained for $N_d \gg N_a$. (This can be realized by employing a p–n^+ junction diode.)

The recombination efficiency η_{rec} is related mainly to the type of a semiconductor, i.e., direct- or indirect-gap material (see Sections 4.7 and 4.8), type and concentration of carriers, presence of defects (e.g., dislocations), and the proximity to surface. In this context, it is essential that light be generated at sufficient distances from defect sites. The radiative efficiency (i.e., the number of photons generated per number of injected electrons) can be generally enhanced by reducing the nonradiative recombination processes. (For a discussion on non-radiative recombination mechanisms, see Section 4.8.)

The extraction efficiency depends primarily on self-absorption of the generated light in the material and reflection of light at the interface. (Note that the self-absorption depends on the absorption coefficient of the material at the emission wavelength.) One method of minimizing the absorption of light is by employing a heterojunction device with the top layer having a wider energy gap

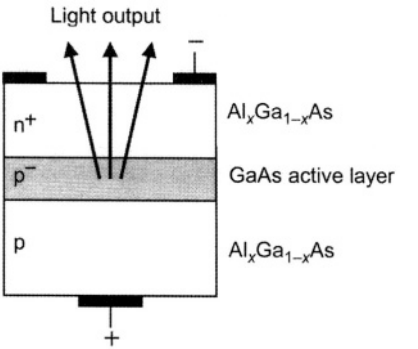

FIGURE 5.13. Schematic diagram of a DH-LED.

(i.e., a transparent layer for the generated light). In this case, however, a judicious choice of the material is required to ensure a close lattice matching for preventing the formation of interface defects that act as nonradiative recombination centers. One example of such a structure is based on $Al_xGa_{1-x}As$ system that facilitates the formation of a double heterostructure (DH) LED (see Fig. 5.13), which is typically produced by epitaxial growth (see Section 6.1). Such a structure greatly reduces some of the problems related to the homojunction devices shown in Fig. 5.12a. (In such homojunction devices, the relatively large spacing between the junction and the surface results in substantial self-absorption of the generated radiation, whereas the close proximity of the junction to the surface leads to an enhanced nonradiative surface recombination.) The DH-LED (see Fig. 5.13) alleviates these problems. In this case, the top $Al_xGa_{1-x}As$ layer, which has a wider energy gap than GaAs, performs as a window that allows the generated photons to be emitted with no self-absorption; in addition, the injected carriers from the n^+-type $Al_xGa_{1-x}As$ into thin GaAs layer are confined within that layer by the potential barrier of a heterojunction formed between p^--type GaAs and *p-type* $Al_xGa_{1-x}As$.

Depending on the angle of incidence, the emitted radiation may be internally trapped due to the total internal reflection at angles above the critical angle $\theta_c = \sin^{-1}(n_2/n_1)$, where n_2 and n_1 are the refractive indices of air (or another medium) and of the semiconductor, respectively. (Note that the relatively high refractive indices of the semiconductors used in LEDs result in small θ_c.) The reflection losses can be reduced by geometrical shaping of the semiconductor–air interface into a hemisphere or by employing encapsulation (typically, a transparent glass or plastic having a refractive index of about 1.5) shaped as a dome, which is a more practical method (such devices are referred to as dome LEDs).

The edge-emitting LEDs are typically heterostructure devices that can be fabricated by employing various materials systems (see Chapter 6) with advantageous features of both carrier and optical confinement, and higher power densities.

The major applications of LEDs are in (i) displays, (ii) light sources, and (iii) fiber-optic communication.

5.5.2. *Light Detecting Devices*

The *photodiode* is a semiconductor device (an optoelectronic transducer), which produces a photocurrent in response to absorbed incident optical power. Unlike light-emitting devices that are forward biased, in order to make these devices more sensitive and faster, semiconductor junctions in photodiodes are reverse biased. [In this case, in the absence of illumination, the dark current, which is limited to I_0, can be kept very small, see Eq. (5.2.26), and Fig. 5.4.] The basic principle of a photodiode is related to the generation of electron–hole pairs (by the absorption of incident photons with energies $h\nu \geq E_g$) inside or near the depletion region (i.e., within a minority carrier diffusion length of the depletion region edge), which results in subsequent separation of these charges in opposite directions through the junction and the flow of current in the device. (It is, however, important to emphasize that the presence of high densities of various defects, e.g., dislocations and grain boundaries, which typically act as recombination centers, present in the material may substantially reduce the carrier diffusion length.) Due to the presence of the junction field, the excess electrons and holes in the depletion region are swept into the *n*- and *p*-regions, respectively. The total current is

$$I = I_0[\exp(eV/k_BT) - 1] - I_{phg} \tag{5.5.2}$$

where I_{phg} is the photogenerated current, which can be expressed as $I_{phg} = eGA(W + L_e + L_h)$, where G is the generation rate of electron–hole pairs (i.e., the number of photogenerated pairs per second and per unit volume). Equation (5.5.2) indicates that the diode characteristic (depicted in Fig. 5.14) is shifted down by a constant amount, which depends primarily on illumination intensity. From Eq. (5.5.2), it also follows that I_{phg} corresponds to the magnitude of the short-circuit current (i.e., for V=0). By setting the current to zero, the open-circuit voltage can be obtained from $I_0[\exp(eV_{oc}/k_BT) - 1] = I_{phg}$:

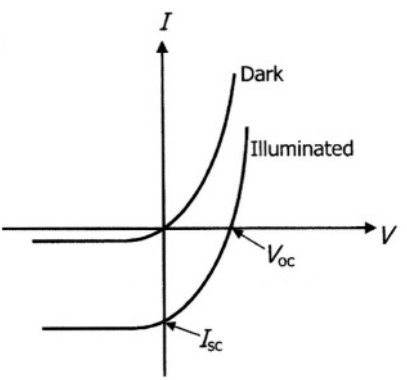

FIGURE 5.14. Schematic representations of the current–voltage characteristic of a *p–n* junction diode in the dark and under illumination.

$$V_{oc} = \frac{k_B T}{e} \ln\left(\frac{I_{phg}}{I_0} + 1\right) \tag{5.5.3}$$

Note that since the absorption process of photons with given energies by a semiconductor (and thus, the spectral response of a photodiode) depends on its energy gap E_g, a broad spectral range can be covered by employing various semiconductors with suitable values of E_g (see Table 6.4). Some of the important semiconductors for photodiodes include Ge, Si, GaAs, and $Ga_xIn_{1-x}As$. It should be emphasized that each of these photodiodes has a specific dependence of the photoresponse on wavelength. Thus, e.g., Si detector has a maximum photoresponse at about 900 nm, whereas $Ga_xIn_{1-x}As$ has a nearly flat responsivity between about 1200 and 1700 nm and is typically optimized for such important wavelengths (for fiber-optic communication) as 1300 and 1550 nm.

Related devices also include *photovoltaic devices* (*solar cells*), which are essentially photodiodes that (i) have a large light-sensitive area designed for both the efficient absorption of light and collection of photogenerated carriers (in this case, the junction is located near the illuminated surface, so that the photogenerated carriers can diffuse to the junction before they recombine), and (ii) operate under conditions for delivery of power into an external load (see Fig. 5.15). The open-circuit voltage, which is also referred to as the photovoltage, essentially provides the source of power in photovoltaic cells. The magnitude of the open circuit voltage is limited to the V_{bi}. Comparing the photodiode and the solar cell operation, one should consider the third and fourth quadrants of the $I(V)$ characteristic (see Fig. 5.14). In the third quadrant, the $I(V)$ characteristic corresponds to the reverse bias with the current increasing as a function of illumination intensity; in this case, the diode can be employed as a photodetector. In the fourth quadrant, the positive photovoltage and the negative current result in the negative power dissipation in the device, indicating the flow of power from the device into the external circuit. Thus, this case corresponds to

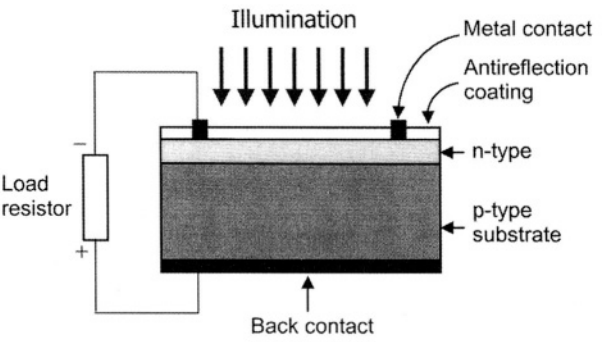

FIGURE 5.15. Schematic illustration of main features of the solar cell. (Dimensions are not to scale.) The thickness of the n-type layer is typically less than 1 μm. The top surface is coated with an antireflection coating, and the top contact is a metal grid that is designed for maximum efficiency.

a photovoltaic cell, which acts as an energy source. To summarize briefly, both the photodiode and solar cell can be characterized by the short circuit current (I_{sc}) and the open circuit voltage (V_{oc}). In the short circuit case, the device operates as a current source to a relatively low impedance load, facilitating an application as a photodetector. In the open circuit case, a positive (forward bias) voltage (V_{oc}) is generated across the terminals with the maximum value corresponding to that of the built-in potential V_{bi} (which is typically smaller than E_g/e), and in this case, the device functions as a power source. For the case of a solar cell (referring to the fourth quadrant), it should be noted that the power (i.e., I times V) provided by the solar cell corresponds to the maximum (i.e., $I_m V_m$) somewhere between I_{sc} and V_{oc} on the curve. In fact, one of the most important characteristics of the solar cell is the *fill factor* (FF), which is the ratio of $I_m V_m / I_{sc} V_{oc}$. The selection of a semiconductor for optimal efficiency depends on three factors, i.e., I_{sc}, V_{oc}, and FF. In general, the solar irradiance spectrum determines the suitability of a given semiconductor for solar cell applications. The semiconductors with a narrower energy gap are more suitable for the utilization of a wider fraction of the solar spectrum, and in principle, I_{sc} increases with decreasing E_g. In contrast, wider energy-gap semiconductors provide greater values of V_{oc}. Thus, the optimization for the spectral response determines the choice of semiconductors that can be efficiently employed in solar cells. The optimal materials are those with the energy gap in the range between about 1.4 and 1.6 eV, although some narrower energy-gap materials (such as Si) facilitate some of the most practical applications. The solar cell energy conversion efficiency η can be expressed as

$$\eta = \frac{V_{oc} I_{sc} \mathrm{FF}}{P_{in}} \tag{5.5.4}$$

where P_{in} is the power of the incident solar radiation. The efficiencies of typical solar cells are in the range between about 10% and 20%. This largely depends on the type of materials employed in a solar cell, i.e., whether it is a monocrystalline, polycrystalline, or amorphous material (see Chapter 6). Note that at standard conditions, there is a theoretical maximum level of solar cell efficiency, which is about 28% for Si and 30% for GaAs. There are, however, several ways to obtain higher efficiencies. These include, e.g., the following:

(a) Surface structuring to reduce reflection loss (e.g., formation of a pyramid structure on the cell surface, so that incoming light strikes the surface several times),
(b) Concentrator cells utilizing higher light intensities by employing mirrors and lenses for focusing light,
(c) Using *multijunction solar cells* (sometimes also referred to as *tandem cells*, or *stacked cells*); these are based on integrating several cells (with different materials having various values of the energy gap) into a single structure. Such solar cells contain two or more cell junctions in such a configuration that each of them is optimized for a specific portion of the solar irradiance spectrum. (Note that, since the typical solar cells employ a single junction

and only photons with energies $h\nu \geq E_g$ of the solar cell material can be used for generating carriers, the response of such single-junction cells is limited to the specific fraction of the solar irradiance spectrum.) Although, in principle, the total solar irradiance spectrum covers a range from infrared to ultraviolet, great portion of that energy cannot be utilized by most solar cells, since it is either below E_g of the solar cell material or it is excessively high for the material to absorb within the bulk. The multijunction structure is a stack of single-junction cells in descending order of E_g values. The top cell, having highest E_g and capturing the high-energy photons, allows transmission of the rest of the photons that are subsequently absorbed by cells with lower E_g.

In functional solar cell systems, large numbers of single cells are connected in series for generating higher system voltage output, and such modules are connected in parallel to increase the current output.

A wide variety of both the photodiodes and solar cells have been developed for different applications, and various types of semiconductor materials are employed in such applications (see Chapter 6 and Table 6.4).

In general, the spectral response of practical devices depends on the energy gap and absorption coefficient of the semiconductor, i.e., the long-wavelength cut-off is determined by its E_g, and the short-wavelength cut-off is determined by its absorption coefficient α (which is typically sufficiently large, so that the incident optical energy is mostly absorbed near the material's surface).

Other types of photodetectors include (i) *photoconductive detectors* (for the description of photoconductivity, see Section 7.2.3), (ii) the *p–i–n* (or *PIN*) *photodiodes* (an undoped *i*-region is inserted between *n*- and *p*-regions; due to the higher resistivity of the fully depleted *i*-region, an applied bias falls practically completely across it; the photoresponse can be optimized by adjusting the thickness of the *i*-region), (iii) the *Schottky photodiodes*, (iv) the *avalanche photodiodes* (these operate under very high reverse bias and provide valuable internal amplification of the photogenerated current), (v) the *phototransistors* (provide an internal gain), and (vi) *vidicons* (these employ the photoconductivity for converting optical images into electrical signals).

Unlike junction photodetectors, which require the formation of a semiconductor diode, in photoconductive detectors an external bias is applied across the electrodes (i.e., a metal such as gold) that are deposited on a semiconductor through a mask to provide the comb-like configuration for improved sensitivity of the device. In photoconductive detectors, irradiation with photons, whose energies are greater than E_g, produces photogenerated electron–hole pairs that can contribute to the conductivity (this is the case of *intrinsic photoconductivity*). Thus, as mentioned above, the spectral response of a specific photodetector depends on the energy gap of a semiconductor employed in the device. For example, CdS is a typical material for the detection of visible light, whereas several narrow energy-gap semiconductors (see Section 6.4), which have the energy gap below about 0.5 eV, are used as infrared detectors. In addition to an *intrinsic type* (i.e., the operation depends on the detection of photons with

energies near E_g), there are also photodetectors of *extrinsic type*, which detect photons having energies lower than E_g. In this case, photoconductivity can still be produced by excitation of carriers from impurity levels in the energy gap. (Note that only one kind of carriers is generated, and this is the case of *extrinsic photoconductivity.*) Thus, the far-infrared radiation detection can be realized, provided the detector is cooled to cryogenic temperatures (in order to reduce the background noise), so that the energy for ionization is provided only by photons.

5.6. SUMMARY

The main applications of semiconductors in various devices are related to the effects produced by the controlled addition (to a semiconductor material) of dopants that facilitate the formation of built-in electric fields and corresponding *junction devices.* A wide range of semiconductor devices based on such junctions are typically employed as *rectifying elements* (e.g., *p–n diodes*), or as parts in various *transistors* that can be employed as current (or voltage) controlled switches and amplifiers in microelectronics technology, or as optoelectronic devices that are employed in photonics technology.

In the *p–n junction,* the *depletion region* and the *internal (built-in) potential* are formed, resulting in a rectifying diode characteristic, which allows the flow of electrical current in one direction but not in the another depending on the *forward-bias* or *reverse-bias* conditions of such a device. This behavior of the *p–n* junction is employed in various electronic device applications. Other junctions include *metal–semiconductor junctions* (or *Schottky barriers*), and *heterojunctions* that are formed between two dissimilar semiconductors. The important properties of hetero-structures are related to their ability to control the carrier transport by controlling the energy barriers and potential variations and their ability to confine the optical radiation, which is especially important in optoelectronic devices.

The two main types of transistors are BJTs and FETs. These are extensively employed in computer technology as fast on–off switches, or in devices for amplification of a current or voltage.

The *light-emitting devices* employ radiative recombination of injected minority carriers with the majority carriers and subsequent emission of light from the forward-biased junction. In such devices, the *energy gap* of a semiconductor determines the energy (or wavelength) of the emitted photon, and the availability of a wide variety of semiconductors with appropriate energy gaps makes such devices suitable for the emission of light in the desired wavelength ranges.

The *photodiode* is a semiconductor device that produces a photocurrent in response to absorbed incident optical power. Unlike light-emitting devices that are forward biased, in order to make these devices more sensitive and faster, semiconductor junctions in photodiodes are reverse biased.

Various semiconductors are commonly employed in a wide range of structures and devices, which cannot be all outlined in the limited format of this book. For more details and extensive discussions on various semiconductor junctions and

devices, and on a wide range of semiconductor device applications, see books in the Bibliography Section B2.

PROBLEMS

5.1. Sketch *p–n* junction characteristics (depicted in Fig. 5.2) for the case of $N_a \gg N_d$.

5.2. An abrupt Si *p–n* junction is doped with $N_a = 10^{17}\ \text{cm}^{-3}$ and $N_d = 5 \times 10^{15}\ \text{cm}^{-3}$. Calculate (a) the built-in voltage V_{bi} and (b) the depletion region widths W, w_p and w_n in equilibrium at 300 K. (For the required data, see Table 4.2 and Appendices A and B.)

5.3. Using the data given in Problem 5.2, calculate the width of the depletion region for a forward bias of 0.3 V and a reverse bias of 3 V.

5.4. The $I(V)$ curve of the solar cell under illumination is depicted in Fig. 5.14. Draw that curve and indicate the area (of a rectangle) corresponding to the power output at the maximum power point.

6

Types of Semiconductors

6.1. INTRODUCTION (SEMICONDUCTOR GROWTH AND PROCESSING)

This chapter provides an outline of different types of semiconductors. In principle, a classification scheme can be proposed based on such criteria as (i) the grouping of elements in the periodic table (e.g., group IV or group III–V), or (ii) the magnitude of the energy gap that defines many applications of a particular semiconductor, or (iii) the structure (e.g., crystalline, polycrystalline, or amorphous). For convenience and completeness, we will use various classification systems, so that the description or information on a specific semiconductor may appear in different sections of the chapter. It should be noted that, although the emphasis in the book is on common inorganic semiconductors, for completeness we also outline briefly, in Section 6.10, some of the issues and possible applications of organic semiconductors.

In this chapter, some commonly employed semiconductors are reviewed. For a complete overview of a wide range of semiconductors, see (Bibliography Section B2) *Semiconductors—Basic Data* (Madelung, 1996) and *Semiconductor Materials* (Berger, 1997).

In general, majority of semiconductors in various applications are prepared as bulk crystals or thin films. Bulk crystals are typically produced as single crystal (cylindrical) *boules* by employing, e.g., the *Czochralski pulling technique* or the *Bridgman directional solidification technique.* In the Czochralski growth method (see Fig. 6.1), a seed crystal (having the required crystal orientation), which is attached to a holder, is inserted into a crucible of the molten material and then is slowly pulled from the melt. This results in crystal growth by solidification of the molten material on the seed crystal surface with the crystal structure and orientation of the growing material being identical to those of the seed crystal. Thus, a cylindrical crystal bar (referred to as a boule) is produced. The crucible is placed inside the graphite, which is heated by employing radio-frequency (RF) induction coils. In order to ensure the crystal growth uniformity, the growing crystal and/or the crucible are rotated at several revolutions per minute. The typical pulling speeds are on the order of several centimeters per

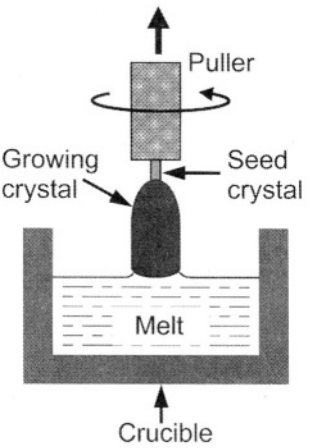

FIGURE 6.1. Schematic illustration of the Czochralski crystal growth method. (The growth chamber and the RF inductive heating coils are not shown.)

hour. The boule diameter can be controlled by changing the temperature, pulling speed and rotation rate. The boules are sliced into wafers (with thickness in the range between about 0.1 and 1 mm), which are subsequently etch-polished on one surface prior to further processing for various device applications or for use as substrates for the deposition of thin films. The properties of such substrates have a major effect on the subsequent growth and properties of the *epitaxial layers,* i.e., layers with the same crystal structure and crystallographic orientation as the substrate. In the epitaxial growth, *homoepitaxy* refers to growth of a given layer on the substrate made of the same material, whereas in *heteroepitaxy* the layer material and the substrate material are different.

The major effort in growth of semiconductors is devoted to low-dimensional structures, which became possible with the advances in such thin-film deposition techniques as *molecular beam epitaxy* (MBE) and *chemical vapor deposition* (CVD) methods. Between these two methods, MBE is an ultrahigh vacuum (UHV) method (i.e., base pressure at about 10^{-11} Torr), whereas CVD method is a low-vacuum technique (i.e., about 10^{-2} Torr or greater). A CVD method for growing epitaxial layers by employing metal-organic gases is referred to as *metal-organic chemical vapor deposition* (MOCVD). These epitaxial techniques are commonly employed for the formation of various structures with desired properties (i.e., *energy-band-gap engineering*), such as *superlattices* and *quantum wells* (QWs) (see Section 6.11).

A schematic illustration of an MBE system, which offers a great control of the film deposition process, is shown in Fig. 6.2. In this UHV technique, several source cells (referred to as effusion cells) supply fluxes of molecular beams of various species that impinge upon a heated substrate (e.g., about 580°C in the case of GaAs). Each separate effusion cell, which is used for individual constituents of required material, is equipped with a shutter; thus, the fluxes can be strictly

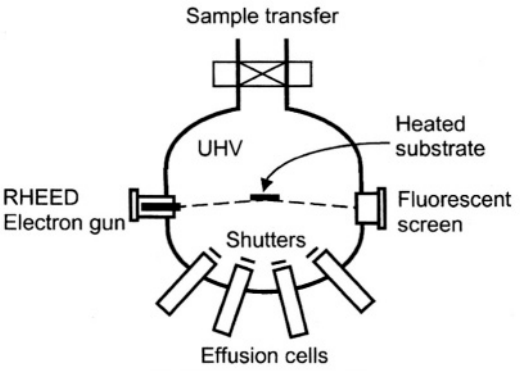

FIGURE 6.2. Schematic illustration of the UHV MBE system, including effusion cells that direct molecular beams of various elements of interest (e.g., Al, Ga, and As) on to a heated substrate.

controlled (i.e., initiated or terminated). The slow growth rates of less than 1 nm s^{-1} allow growing high-quality crystalline materials, and opening or closing shutters allows controlling the film stoichiometry and doping levels within a monolayer. As discussed in Section 6.11, by using MBE technique for growing epitaxial layers, it is possible to construct artificial structures in which layers of different materials and thickness alternate; in such structures, the period of the layers may be as small as a few monolayers. The main advantages of this technique (with its superior clean environment) include its suitability for monolayer growth with atomically abrupt interfaces and great uniformity, stoichiometry and impurity control. As discussed in Section 7.5, reflection high-energy electron diffraction (RHEED) can be employed as an *in situ* monitoring tool in MBE (see Fig. 6.2). In this method, high-energy electrons (in the range between about 5 and 50 keV), striking the sample at a grazing angle between about 1 and 5°, are scattered by the first few atomic layers of sample surface. The pattern produced on a phosphor screen positioned opposite the electron gun can be monitored, and from the features of such RHEED patterns (i.e., from the spacing and symmetry of the features, and from intensity oscillations), information on, e.g., the surface structure and coverage, as well as on the degree of surface roughness and film growth mechanism can be derived. Thus, this method (i.e., the time evolution of the RHEED pattern during epitaxy) allows monitoring of structural changes in the film during its growth or during its post-growth annealing.

The thickness of epitaxial layers is typically less than 1 μm. In order to obtain a high-quality epitaxial layer (and to minimize strain), it is important to ensure that (i) the crystal structures of the epitaxial layer and the substrate are similar and (ii) their lattice constants are matched as close as possible in order to minimize the *mismatch strain* (i.e., a strain at the interface between the layer and substrate) and avoid the formation of defects (i.e., *misfit dislocations*), which typically act as nonradiative recombination centers or deep traps and thus are detrimental in device applications. In this context, it is useful to refer to a diagram presenting an energy gap and lattice constant for several important direct- and indirect-gap

semiconductors, as shown in Fig. 6.3. As can be seen from this figure, e.g., GaAs is closely lattice-matched to AlAs and to any ternary alloy $Al_xGa_{1-x}As$ with the energy gap varying in the range between about 1.4 and 2.2 eV. (In Fig. 6.3, the lines joining the binary compounds correspond to the energy gap and lattice constant values for the ternary compounds. In other words, e.g., the points on the line connecting GaAs and AlAs represent various ternary alloys of $Al_xGa_{1-x}As$.) Thus, the energy gap of these $Al_xGa_{1-x}As$ alloys can be adjusted within the technologically useful range from visible to near infrared spectral regions. This figure also demonstrates that the lattice matching for $Ga_xIn_{1-x}As$ growth on InP substrate can be realized for only a given value of E_g (and hence of x). In the context of the lattice matching in the epitaxial growth, an additional concern is due to the fact that materials, which may be lattice matched at a given temperature (e.g., the growth temperature), may be mismatched due to the difference in thermal expansion coefficient at a different temperature (e.g., on cooling to lower temperature).

Semiconductor device fabrication processes involve a variety of steps (see, e.g., Jaeger, 1988; Mayer and Lau, 1990 in Bibliography Section B2). These, in general, may involve (i) *lithography* (i.e., the process of transferring patterns to a semiconductor wafer), (ii) *wet* and *dry etching* (i.e., immersing samples in various solutions in the case of wet etching, or using ion beams for local etching), (iii) *diffusion* or *ion implantation* (for the introduction of dopant atoms into the semiconductor wafer), (iv) thin film deposition, and (v) contact formation. Using *photolithography* methods, these processes allow manufacturing semiconductor devices on the sub-micron scales. It should be noted that the repeated lithographic process steps are some of the most crucial steps in the device manufacturing process. An important step in these processes is the deposition of a *resist,* which is

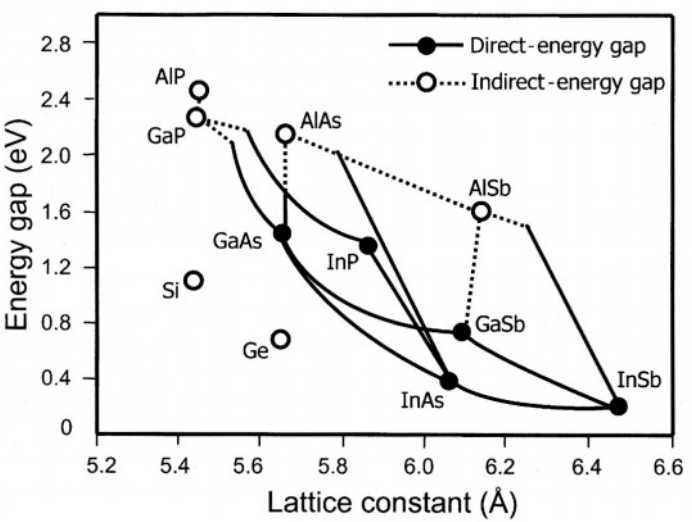

FIGURE 6.3. Plot of the energy gap vs. lattice constant for both direct- and indirect-gap semiconductors.

exposed to photons (in photolithography), or electron beam (in *electron-beam lithography*), or X-rays (in *X-ray lithography*), or ion-beam (in *ion-beam lithography*). With an exposure of selected areas only, the resist may be patterned and subsequently used to mask various device-processing steps. Two classes of resist employed in lithographic processes are *negative* and *positive resists.* The resist, which is polymerized on exposure (thus the resist becomes insoluble in a developer solution and it is harder to remove relative to the unexposed area), is called a negative resist. Whereas a positive resist is softened on exposure and thus the exposed resist becomes more soluble and it is easier to remove relative to the unexposed area. A schematic diagram of the photolithographic patterning process steps is shown in Fig. 6.4 with the associated description of these steps listed below: (note that the goal is to extract part of the layer in order to provide, e.g., interconnection paths).

(a) The wafers are completely covered with a thin film (about 1 μm thick), which may be either a metal, a dielectric, or a semiconductor.
(b) Resist, which is dispensed on the wafer, spreads out into a thin uniform coating as a result of spinning the wafer at a few thousand revolutions per minute; subsequently the resist is baked dry.

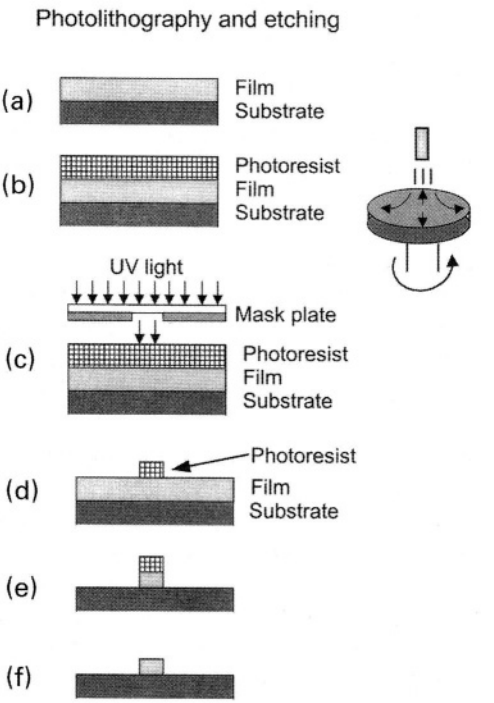

FIGURE 6.4. Schematic diagram of the photolithography and etching process.

(c) The wafer, which is coated with the resist, is exposed to UV light through a mask plate containing the required pattern. The resist may be either negative or positive. In the case of a negative resist, exposure to light causes cross-linking of the polymer bonds, making the exposed resist harder to remove as compared to the unexposed areas, and thus a negative resist remains on the surface area in exposed regions. In the case of a positive resist, exposure to light breaks the polymer bonds, and thus the exposed resist is easier to remove relative to the unexposed area.
(d) After the exposure, the wafers are dipped or sprayed with a developer solution that removes the softer parts of the resist.
(e) The wafer with the resist is baked in order to toughen the resist film; this is followed by the etching of the wafer to remove the unprotected parts of the top layer.
(f) The residual resist is removed, and the above steps can be repeated for the next layer added to the wafer.

For integrated circuit manufacturing, a given set of the above process steps may be repeated numerous times depending on the levels of integration. To summarize briefly, semiconductor wafer processing includes the following important steps:

(a) thin-film deposition;
(b) lithography (transfer of patterns from masks to the resist on the wafer surface);
(c) chemical and physical etching.

Thin-film deposition methods include epitaxy (typically, for single crystal semiconductors), CVD (for dielectrics or polycrystalline semiconductors), evaporation (for metals), sputtering (for metals or insulators), and spin-on (for organic compounds and polymers).

Two basic methods of etching are wet (chemical) and dry (physical) etching (see Fig. 6.5). Wet chemical etching employs various dilute acid and alkaline solutions (the rate of etching and shape of etched pattern depend on materials, etchants, and temperature). Some of these produce isotropic etching (i.e., etching is equal in all directions), whereas others are sensitive to crystal plane. Wet etching usually produces no physical damage; however, in this case (i) it is difficult to form fine patterns, (ii) the specific shape of etched patterns depends on crystal orientation, and (iii) the method is likely to etch under the edge of the mask (see Fig. 6.5).

Dry (physical) etching, which employs plasma etching, offers improved etching control of features with increasing aspect ratio. In this case, the etching is performed in a vacuum chamber containing gases with chemically active ions (e.g., fluorine or chlorine ions, generated by RF excitation) that attack the resist and the exposed areas of the semiconductor differentially. In this process, the etch rates can be controlled by varying such conditions as RF power density, temperature, gas composition and flow rates. Dry etching is an anisotropic etching process with no undercutting problem that is typical of wet etching. This method allows forming directional etching profiles (see Figs. 6.5 and 6.6). Note, however, that damaged

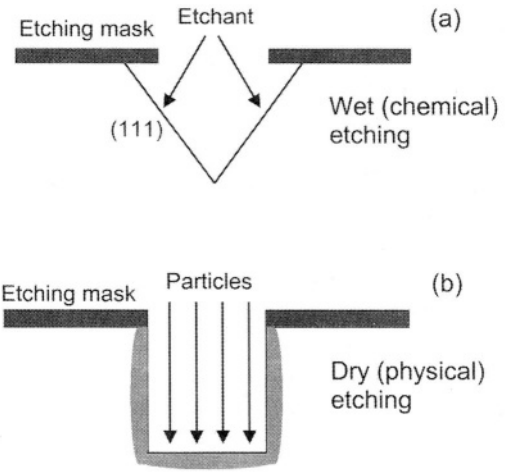

FIGURE 6.5. Schematic illustration of wet (chemical) and dry (physical) etching of InP crystal.

layer may also form, and this requires some chemical etching to remove it. Other types of dry etching include, e.g., *sputter etching* (i.e., sputtering of surface atoms using energetic noble gas ions such as Ar^+, with no chemical reactions involved), and *reactive ion etching* (RIE) that combines the plasma and ion beam removal of the layers.

In addition to these steps, the fabrication of integrated circuits may also include steps such as *oxidation* (to form a silicon dioxide layer on the silicon surface), *doping* (by diffusion or ion implantation), *metallization* and *interconnection* between electronic components, and *packaging* (to produce the final product that is protected from moisture and contaminants).

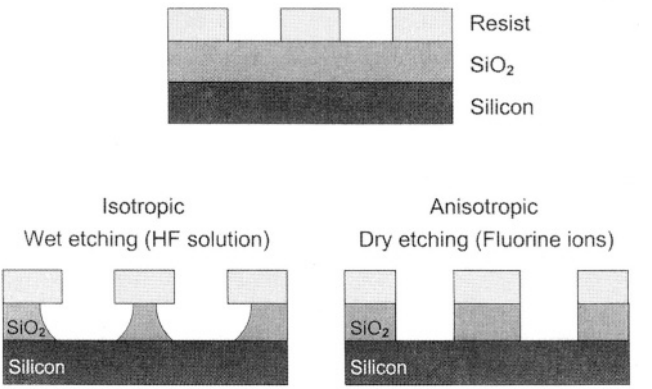

FIGURE 6.6. Schematic illustration of the formation of grooves in SiO_2 using resist patterns and isotropic wet chemical etching or anisotropic dry (fluorine ion plasma) etching.

Note that each wafer may contain several hundred identical integrated circuits (referred to as *chips*), with each of them containing millions of submicron circuit elements (in very large-scale integration, or VLSI). The finished wafer is subsequently diced into individual chips (the size of the chip may vary from about 1 mm × 1 mm to 10 mm × 10 mm). As mentioned in Chapter 5, the integrated circuit typically consists of large numbers of circuit elements such as resistors, capacitors, diodes and transistors, which are combined in many different configurations (for carrying out such functions as, e.g., digital data storage and amplification). In a monolithic integrated circuit, all the circuit components are fabricated next to each other into or on top of a single semiconductor chip by using various fabrication steps outlined above. This ability of joining large numbers of various circuit elements on a single piece of a semiconductor facilitates a mass production of such integrated circuits.

6.2. ELEMENTAL SEMICONDUCTORS

The important elemental semiconductors are group IV materials, such as silicon (Si), germanium (Ge), and diamond (C). These group IV elemental materials all have diamond crystal structure, i.e., each atom is in a tetrahedral configuration with four nearest atoms, and thus they are also referred to as tetrahedrally-bonded semiconductors. Another group IV elemental semiconductor having such a structure is α-Sn ($E_g = 0.08$ eV), which is also referred to as gray Sn. Other elemental semiconductors, having various structures differing from diamond structure, include group III element boron, group V material phosphorus, and group VI materials such as sulphur (S), selenium (Se), and tellurium (Te).

It should be emphasized that currently Si is the most important material used in electronic devices (e.g., integrated circuits). Some of the important advantages of Si over other semiconductors are (i) a relative ease of passivating the surface by oxidizing it in a controlled manner and forming a layer of stable native oxide that substantially reduces the surface recombination velocity, (ii) its hardness that allows large wafers to be handled safely, (iii) its thermal stability, up to 1100°C, that allows high-temperature processing related to diffusion, oxidation and annealing, and (iv) its relatively low cost. The basic limitations of Si are due to (i) the magnitude and type of its energy gap (i.e., the value of $E_g = 1.12$ eV and the fact that it is an indirect energy-gap material, which limit the optoelectronic applications of this material) and (ii) the relatively lower carrier mobility (as compared to, e.g., GaAs).

As mentioned above, the basic properties of Si related to the magnitude and type of its energy gap (i.e., the fact that it is an indirect energy-gap material) limit the optoelectronic applications of this material. However, emerging materials based on Si nanostructures (e.g., Si nanocrystals, quantum wires and dots, and porous Si) and Si_xGe_{1-x} layers grown on Si substrate, appear to be promising materials in various applications. In nanostructures, quantum confinement of carriers leads to (i) increased electron–hole wave function overlap (and hence, increased photon

emission efficiency) and (ii) high-energy shift (i.e., blue shift) of the emission peak. Porous Si, which can be obtained from the anodic etching of crystalline Si in aqueous HF solution, contains a network of pores and crystallites with sizes of the order of several nanometers. This material exhibits relatively efficient luminescence, which is several orders of magnitude higher than that in crystalline Si, and it is thought to be related to the quantum confinement effects in nanocrystalline Si.

As mentioned earlier, heteroepitaxy involves a single crystal layer growth on a substrate of a different material. One of the main objectives of heteroepitaxy is to "engineer" materials and structures with unique properties (i.e., artificial structures). One of the examples of practical applications of energy-band-gap engineering is that of SiGe/Si heterostructures and superlattices, which are attractive since such systems combine highly developed Si-based technology with the benefits derived by the introduction of Ge in Si-based devices (e.g., transistors), leading to substantial improvements in the performance of conventional Si devices related to such properties as operating frequency and power capabilities. (Note that SiGe has significantly higher carrier mobility than Si.) The smaller energy gap and larger refractive index of SiGe imply that these heterostructures are also suitable in optoelectronic applications, such as waveguide detectors (in the 1.3–1.6 μm wavelength range) that can be employed in fiber-optic communication. An important approach in such structures is that of *strained layer heteroepitaxy*; in this case, if the layers are sufficiently thin, the lattice mismatch between the dissimilar semiconductors is accommodated by strain, and no misfit dislocations at the interface are generated. Note that this accommodation by strain results in a tetragonal distortion of the diamond unit cell. Consequently, the lattice distortion leads to the changes in the band structure and heterojunction band offsets (i.e., the differences, between dissimilar semiconductors, in energies of the valence band maxima and the conduction band minima). Thus, this allows flexibility in band-structure engineering. The typical values of the *critical thickness* in such a material system are between about 1 nm for pure Ge (note about 4% lattice mismatch between Si and Ge) and about 100 nm for $Si_{0.8}Ge_{0.2}$ (i.e., 1 % mismatch).

Many semiconductors, in principle, can be grown on Si substrates. For example, growth of III–V compounds on Si substrate is attractive, since such heterostructures would enable to integrate optical devices in the III–V compounds with Si circuitry on a monolithic chip. III–V compounds offer a wide range of applications in optoelectronic devices, whereas Si offers both a convenient electronic device technology and a large area substrate that is mechanically stronger than, e.g., GaAs and also has a larger thermal conductivity. The important issues of concern in obtaining high-quality epitaxial heterostructures (e.g., GaAs/Si) are (i) the presence of high dislocation densities due to the lattice constant mismatch between the epitaxial layer and substrate, (ii) residual stresses in the epitaxial layers due to the difference in thermal expansion coefficients of the epitaxial layer and substrate, and (iii) the formation of structural defects (i.e., antiphase boundaries) due to the epitaxial growth of a polar crystal (e.g., GaAs) on a nonpolar substrate (e.g., Si). Various approaches are being used to overcome these problems. These include, e.g., (i) thermal cycle annealing, (ii) growth interrupts, (iii) selective area growth, and (iv) insertion of strained-layer superlattices.

In general, growth of various semiconductors on Si substrate may allow complete optoelectronic integration between different devices and functions in a monolithic design. Such a monolithic integration of GaAs and Si, e.g., would allow to combine GaAs optoelectronic or high-speed components with Si circuitry. Thus, various intrachip and interchip combinations of optical communication devices with Si signal processing devices can be realized. Besides GaAs/Si system, other useful heterostructures are, e.g., $Ga_xIn_{1-x}As_yP_{1-y}$ on Si for fiber-optic communication, and InSb or $Hg_{1-x}Cd_xTe$ on Si for infrared detection.

The ability to deposit diamond films at relatively low temperatures has catalyzed its wide range of potential applications, including both passive and active electronic device applications. Having both a very high thermal conductivity and a very high electrical resistivity, diamond can be used as electronic substrates and heat sinks (i.e., heat-spreading components), especially in high-power microelectronic device applications. Owing to its excellent optical transparency in a wide wavelength range, diamond protective coatings can be used as optical windows in harsh environments. Another possible application of diamond films is in flat-panel displays, where an array of diamond tips can be employed as cold cathode field emission sources (owing to the negative electron affinity of diamond). Some challenges (at this juncture) in active electronic device applications, such as diodes and transistors, include difficulties with (i) growing high-quality films on foreign substrates, (ii) *n*-type doping and (iii) fabricating actual devices. Thus, although the basic electronic properties (e.g., carrier mobility) of diamond, in principle, are superior as compared to other semiconductors, the difficulties involved with (i) growing high-quality material, (ii) realizing shallow *n*- and *p*-type doping, and (iii) developing suitable device processing techniques, hinder (at this juncture) its electronic device applications.

6.3. COMPOUND SEMICONDUCTORS

In this section, we will only review the most commonly used compounds. For a more complete list of a wide range of compound semiconductors, see, in Bibliography Section B2, *Semiconductors—Basic Data* (Madelung, 1996) and *Semiconductor Materials* (Berger, 1997).

At this juncture, a general observation on compound semiconductors should be emphasized. The effects of impurities and defects in compound semiconductors are significantly different from those in elemental materials. For example, the donor and acceptor impurities in compound semiconductors may be introduced on either sublattice, and nonstoichiometry effects may become more crucial.

6.3.1. III–V Compounds

The III–V compounds (e.g., GaAs, GaP, GaN, A1As, InSb, InAs, and InP) are important semiconductors for various device applications. In general, these materials crystallize with a relatively high degree of stoichiometry and most of them can easily be obtained *n*- and *p*-type. Many of these compounds (e.g., GaAs, InAs, InP, and InSb) have direct energy gaps and high carrier mobilities;

thus, the common applications of these semiconductors are in a variety of optoelectronic devices for both the detection and generation of electromagnetic radiation, and also in high-speed electronic devices. The energy gap of these compounds, which are useful for optoelectronic applications, ranges (at room temperature) from 0.17 eV for InSb to 3.44 eV for GaN, covering the wavelength range from about 7.29 to 0.36 μm, i.e., from infrared through visible and to ultraviolet spectral ranges. Materials such as GaAs and InP are also extensively used as substrates for a wide variety of electronic and optoelectronic devices. In this context, the two bulk crystal growth techniques employed for providing such substrates are the *liquid encapsulated Czochralski* (LEC) and *horizontal Bridgman* (HB) methods. Among these two methods, the HB growth method can produce the crystal wafer substrates with lower dislocation densities, and thus they are more applicable for optoelectronic devices such as light-emitting devices; whereas the LEC technique is employed in producing crystal wafers with larger diameter, which are more advantageous in electronic device applications such as transistors and integrated circuits.

The main thrust of research and development in the III–V compounds, including one of the most promising semiconductors, GaAs, is directed at growth of epitaxial layers, which are superior to bulk crystals. (Note that the epitaxial layers of these semiconductors are grown at much lower temperatures than their melting point.) Several major challenges in the development of compound semiconductors such as GaAs, in comparison with elemental materials such as Si, are related to the more complex defect structure in GaAs that has two sublattices and the issue of the deviation from stoichiometry in compound materials, in general. The zincblende III–V compounds have no center of symmetry, and their ⟨111⟩ direction is crystallographically polar. In such structures, the parallel {111} surfaces of a crystal segment may have different properties depending on whether they terminate with atoms from group III or V. The deviation from stoichiometry, which may result in either the Ga-rich or As-rich material, affects, e.g., the formation of edge dislocations, which may have different properties depending on whether a dislocation line consists of atoms from one group or another. Many native defects and their complexes are responsible for both shallow and deep levels in these compounds. In GaAs, e.g., As vacancies act as shallow donors, whereas Ga vacancies act as deep acceptors, and a dominant intrinsic defect (i.e., As_{Ga} antisite defect), acting as EL2 center and giving rise to a deep level, is thought to be accountable for the semi-insulating property of undoped material.

Among these compounds are narrow energy-gap semiconductors (with the energy gap below about 0.5 eV, e.g., InSb and InAs), which are extensively employed in infrared optoelectronic device applications. Some III–V compounds, such as GaN and A1N, are the wide energy-gap semiconductors that are useful materials in optoelectronic devices operating in the visible and ultraviolet spectral regions, as well as in high–temperature and high-power devices. In this context, an important feature of III–V nitrides (i.e., InN, GaN, and A1N) is that they form a continuous alloy system with direct energy gaps ranging from about 2 eV (InN), 3.4 eV (GaN) and 6.3 eV (A1N); in other words, such a system would have a continuous range of energy-gap values throughout the visible and ultraviolet ranges.

In addition to the binary III–V compounds, materials such as ternary (e.g., $Al_xGa_{1-x}As$ and $GaAs_{1-x}P_x$) and quaternary (e.g., $Ga_xIn_{1-x}As_yP_{1-y}$) alloys with "tunable" properties are also used in specific device applications. In such cases, these compounds are typically grown as epitaxial layers in heterojunction systems on substrates such as GaAs or InP, e.g., $Al_xGa_{1-x}As$ on GaAs, or $Ga_xIn_{1-x}As_yP_{1-y}$ on InP. In such systems, it is highly desirable to match the lattice structure and lattice constant of the epitaxial layer and the substrate. (For the values of lattice constants of various semiconductors, see Table 2.3.) An example of such lattice matching is that of AlAs and GaAs, and thus $Al_xGa_{1-x}As$/GaAs system with lattice matching can be obtained. Similarly, lattice–matched $Ga_xIn_{1-x}As_yP_{1-y}$ /InP system can be obtained. By choosing the composition (i.e., x and y) in such cases, it is possible to select a particular semiconductor property (e.g., the energy gap) to fit the specific device requirement. GaAs can be also grown on other lattice-matched substrates, e.g., GaAs on Ge. As described in the preceding section, growth of GaAs on Si substrate is especially attractive for the integration of optoelectronic devices based on GaAs with Si circuitry on a monolithic chip. However, one of the important issues of concern in such epitaxial heterostructures is the lattice constant mismatch between the epitaxial layer and substrate, which results in generation of high dislocation densities in the epitaxial layer. There are several ways to reduce the presence of these defects in such mismatched heterostructures. These include, e.g., (i) selective area growth, (ii) incorporation of strained layer superlattices, and (iii) thermal cycle annealing. As mentioned in the preceding section, additional issues of concern in GaAs/Si also include the presence of residual stresses in the epitaxial layers due to the difference in thermal expansion coefficients of these materials, and the presence of antiphase domains due to the polar (GaAs)–on–nonpolar (Si) growth.

In addition to a wide range of applications of III–V compounds in optoelectronic devices, the possibility of forming *multiple quantum* well (MQW) structures (i.e., alternating thin layers of different materials), based on these compounds, offers new possibilities in "engineering" the electronic band structure and in designing novel semiconductor devices (see Section 6.11).

6.3.2. II–VI Compounds

The Zn and Cd-chalcogenides (i.e., compounds with O, S, Se, and Te) cover a wide range of electronic and optical properties due to the wide variations in their energy gap. These compounds are also relatively easily miscible, which allows a continuous "tuning" of various properties. However, the preparation of high-quality materials and the processing technologies are not sufficiently developed in comparison with those related to Si and some III–V compounds. The II–VI compounds are typically *n*-type as grown, except ZnTe, which is *p*-type. Among these compounds, in CdTe the conductivity type can be changed by doping, and thus *n*- and *p*-type materials can be obtained. Others, such as ZnSe, ZnS and CdS, can be doped to produce a small majority of holes. For device applications, it is possible (i) to form heterojunctions in which the *n*- and *p*-sides of the junction are of different II–VI compound semiconductors, and (ii) to use metal–semiconductor

and metal–insulator–semiconductor structures for carrier-injection device applications. Since all the II–VI compound semiconductors have direct energy gaps, efficient emission or absorption of electromagnetic radiation can be expected in these materials. Thus, these semiconductors are important mainly for their optical properties. In addition to the binary II–VI compounds, materials such as ternary (e.g., $Zn_{1-x}Cd_xS$ and ZnS_xSe_{1-x}) and quaternary (e.g., $Zn_{1-x}Cd_xS_ySe_{1-y}$) alloys with "tunable" properties are also of interest.

An important issue in device applications of these compounds, except in the case of CdTe, is a phenomenon of apparent self-compensation, i.e., the generation of native point defects of a type that compensate dopant atoms and thus hinder the formation of *p–n* junctions in most of these semiconductors. Self-compensation is plausible in these compounds, since they are partially ionic, and, thus, e.g., anion and cation vacancies can act as acceptors and donors, respectively.

Some important commercial applications of these compounds include, e.g., phosphors in lighting and various display applications (CRT and thin-film electroluminescent flat displays, e.g., ZnS : Ag and ZnS : Mn), infrared photodetectors for imaging systems ($Hg_{1-x}Cd_xTe$), protective windows and optical elements (ZnSe and ZnS), nuclear radiation detectors (CdTe), and solar cells (CdS and CdTe). Other potential applications include, e.g., short-wavelength light-emitting devices (e.g., blue laser), and integrated optoelectronic systems (with Si and GaAs) for optical processing and computation.

6.3.3. IV–VI Compounds

The lead chalcogenides (i.e., PbS, PbSe, and PbTe) are characterized by narrow energy gaps, high carrier mobilities, and high dielectric constants (see Table B1 in Appendix B). The unique feature of the direct energy gap in these compounds is that it increases with increasing temperature (i.e., the energy gap has a positive temperature coefficient, PTC), in contrast to the temperature behavior of the energy gap in other elemental and compound semiconductors that have a negative temperature coefficient. Main applications of these compounds are in light-emitting devices and detectors in the infrared spectral region.

6.3.4. I–III–VI_2 (Chalcopyrite) Compounds

The I–III–VI_2 compound semiconductors such as, e.g., $CuAlS_2$, $CuGaS_2$, and $CuInSe_2$ are of interest in various device applications. These semiconductors have the direct energy gaps in the range between about 1 and 3.5 eV, and they crystallize in the tetragonal structure, which is close to the structure of a mineral chalcopyrite (i.e., $CuFeS_2$). In addition, materials such as $CuAlS_2$ and $CuGaS_2$ can be obtained as *p*-type materials, making them interesting candidates for heterojunctions with wide energy-gap *n*-type II–VI compound semiconductors. Some possible applications of the I–III–VI_2 compound semiconductors include, e.g., (i) light-emitting devices based on a wide energy-gap material such as $CuAlS_2$ with the energy gap of 3.5 eV that would accommodate visible (including blue) luminescence centers and (ii) photovoltaic solar cells based on $CuInSe_2$ and $CuInS_2$ with the energy gaps of 1.04 and 1.53 eV, respectively. In solar cells, $CuInSe_2$,

having a high absorption coefficient, can be used as an efficient light absorbing layer in a heterojunction structure such as $CdS/CuInSe_2$.

6.3.5. Layered Compounds

Layered compounds are examples of a specific covalent–van der Waals structure. Among this type of compounds that have attracted interest in various applications are (i) layered transition metal dichalcogenides, such as MoS_2 and ZrS_2 (with energy gaps of about 2 eV), (ii) compounds such as HgI_2 and PbI_2 (with energy gaps of about 2.1 and 2.3 eV, respectively), and (iii) GaSe and InSe (the energy gaps of about 2 and 1.2 eV, respectively). In these materials, the bonding is covalent within the layers, which are held together by the weak van der Waals bonding. The fact that layered compounds are strongly bound in two directions (by the covalent bonding) and weakly bound in the third direction (i.e., *c*-axis) leads to anisotropy of the structural and physical properties of these semiconductors. These compounds are essentially quasi-two-dimensional systems, and, thus, some physical properties in such structures may exhibit a two-dimensional behavior.

Since only weak van der Waals forces act between the layers of these compounds, it is possible to introduce a variety of "foreign" atoms and some organic molecules between the layers forming the intercalation compounds; such an intercalation process has been extensively demonstrated in layered transition metal dichalcogenides. These intercalation compounds exhibit significant changes in a wide variety of properties as compared to the nonintercalated materials. The process of intercalation is reversible, and thus such layered compounds may be used, e.g., as cathodes in rechargeable high-energy-density batteries, in which the cell reaction is the reversible intercalation of the layered crystals.

The presence of weak van der Waals bonds between the layers also accounts for easy cleavage, implying somewhat delicate handling of bulk materials in various device applications. For example, one of the problems related to applications of such promising semiconductors as HgI_2 and PbI_2 in nuclear detectors is the fact that their layered crystal structure makes these materials soft and fragile.

It should be noted that these layered compounds must be differentiated from superlattices, which can be considered as artificially produced compositional layered structures, i.e., periodic arrays of thin layers of different compositions. The term superlattice is also used in crystallography to indicate an additional periodicity in the crystal structure. In semiconductor science and technology, superlattice structures (i.e., two dissimilar semiconductors in a periodic layered array) imply those that exhibit quantum effects.

6.4. NARROW ENERGY-GAP SEMICONDUCTORS

Narrow energy-gap semiconductors are those that have the energy gap below about 0.5 eV. Such semiconductors are extensively employed in such infrared optoelectronic device applications as detectors and diode lasers. The properties of some of the narrow energy-gap semiconductors are given in Table 6.1.

TABLE 6.1. Properties of some common narrow energy-gap semiconductors at room temperature (300 K)

	E_g (eV)	Transition
InSb	0.17	d
InAs	0.36	d
PbSe	0.28	d
PbTe	0.31	d
PbS	0.41	d

The energy gap E_g; Transition d indicates that it is a direct energy-gap material.

Photoconductive lead sulphide (PbS) and lead selenide (PbSe) detectors can be employed in the spectral range between about 1 and 6 μm.

Another important material used as a detector in the infrared range is $Hg_{1-x}Cd_xTe$. The $Hg_{1-x}Cd_xTe$ epitaxial layers can be grown, e.g., on CdTe substrates, lattice-matched $Cd_{1-x}Zn_xTe$ substrates, or CdTeSe substrates. The energy gap in $Hg_{1-x}Cd_xTe$ can be varied in the range between 0 and 1.56 eV depending on x.

In addition to these semiconductors, superlattice structures and QWs can also be employed in infrared detector applications. For example, HgTe–CdTe superlattices may offer substantially less tunneling noise and better control over cut-off wavelength as compared to $Hg_{1-x}Cd_xTe$ material having the same energy gap.

6.5. WIDE ENERGY-GAP SEMICONDUCTORS

The wide energy-gap semiconductors are also referred to as refractory semiconductors, since they can be employed in high-temperature applications. (Some of these semiconductors are listed in Table 6.2.) In addition to having wide energy gaps, some typical common properties of diamond, SiC and III–V nitrides include, e.g., high thermal conductivity, high saturation electron drift velocity, high breakdown electric field, and superior physical and chemical stability. To expand further, it should be noted that (i) the wide energy gap enables these materials to emit and detect light in the short-wavelength region, including blue and ultraviolet, and the wide energy gap also ensures that various electronic devices can operate at relatively very high temperatures (greater than about 600°C) without experiencing intrinsic conduction effects, (ii) high thermal conductivity indicates that, since the excess heat generated during the operation can dissipate readily, the wide energy-gap semiconductor devices can operate at very high power levels, (iii) high saturation electron drift velocity implies that the wide energy-gap semiconductor devices can operate at high frequencies (i.e., RF and microwave), and (iv) high breakdown electric field enables the realization of high-power electronic devices and it also allows high device packing density for integrated circuits. It should be noted in this context that the thermal conductivities of the II–VI compound

TABLE 6.2. Properties of common wide energy-gap semiconductors at room temperature (300 K)

	E_g (eV)	Transition	μ_e ($cm^2\ V^{-1}\ s^{-1}$)	μ_h ($cm^2\ V^{-1}\ s^{-1}$)	Thermal conductivity ($W\ cm^{-1}\ K^{-1}$)	a (Å)
C (diamond)	5.47	i	2200	1600	10	3.567
SiC (W, 6H)	3.0	i	600	40	3.6	3.081
SiC (Z, 3C)	2.3	i	1000	40	3.2	4.36
AlN (W)	6.28	d	135	14	2.0	3.11
GaN (W)	3.44	d	1000	30	1.5	3.189
ZnSe (Z)	2.7	d	600	80	0.19	5.668
ZnS (Z)	3.68	d	165	40	0.27	5.41

The energy gap, E_g; transition (*d* or *i*) indicates whether it is a direct- or indirect-energy gap; mobility of electrons, μ_e, and holes, μ_h (note that these values depend strongly on the purity of the material measured); thermal conductivity; lattice constant, *a*. Lattice: D, diamond; W, wurtzite (hexagonal); Z, zincblende (cubic). Note that some values in the table are approximate, and some compounds (e.g., SiC and ZnS) can be grown in either W (hexagonal, 2H) or Z (cubic, 3C) structures, and they can also exhibit various polytypic structures (e.g., 4H and 6H); in the case of the wurtzite structure, for complete description, a lattice constant *c* is also required (see Table 2.3).

semiconductors (ZnSe and ZnS) are much lower than those of other wide energy–gap materials quoted in this Section (see Table 6.2).

Thus, the potential devices based on the wide energy-gap semiconductors (based on such properties as thermal conductivity, breakdown electric field, saturation velocity), in principle, should be superior to present devices, especially in high power, high temperature, high frequency, and short-wavelength devices. In practice, however, in many cases of the wide energy-gap semiconductors, the difficulties involved with growing high-quality materials (and with appropriate doping type) and fabricating actual devices, hinder their applications at this juncture.

Some major present applications of some of the wide energy-gap semiconductors are in optoelectronic devices (e.g., short-wavelength light emitters and detectors) operating in the blue, violet and ultraviolet spectral regions, as well as in high-temperature and high-power devices. A short-wavelength (e.g., blue) laser, e.g., would substantially increase the storage density for optical recording, as compared to the devices operating at longer wavelengths (i.e., between the red and the near infrared). With the use of wide energy-gap semiconductors it also becomes feasible to produce, e.g., full-color flat-panel displays.

The high-temperature electronic devices usually imply those that operate at temperatures greater than about 125°C (i.e., the typical upper limit for commercial electronic devices); high-temperature electronic devices also include those power transistors and circuits in which high junction temperatures (greater than about 200°C) are generated during the operation at room temperature.

As mentioned above, diamond has a wide range of potential applications. These, in principle, include both passive and active electronic device applications. Diamond has both a very high thermal conductivity and a very high electrical resistivity, and thus it can be used as heat sinks in electronic devices; it also has

excellent optical transparency in a wide wavelength range, and thus diamond protective coatings can be used as windows in various applications.

As mentioned in Chapter 2, some semiconductors can crystallize in several different structures, leading to existence of different polytypes. SiC can be grown in either hexagonal (2H) or cubic (3C) structures, and it can also exhibit various polytypic structures (e.g., 4H and 6H). Some of these polytypes, e.g., 3C and 6H, are of particular interest for electronic devices due to their relatively high electron mobility and large energy-gap values. The main limitation of SiC in laser applications is its indirect gap, but the material can be doped either *n*- or *p*-type, and various devices can be fabricated. These include high-temperature, high-power, and high-frequency devices, as well as light-emitting devices. For example, the realization of blue light-emitting diodes (LEDs) is accomplished by forming an abrupt *p–n* junction (aluminum- and nitrogen-doped, respectively) in SiC (6H polytype). It should be noted that the relatively high thermal conductivity of SiC also allows the high-density integration of SiC devices. Additional advantage of SiC, as compared to other wide energy-gap semiconductors, is that it can be oxidized to form a stable, electrically insulating oxide. (In this case, the thermal oxidation of SiC results in SiO_2–SiC interface, or SiO_2 forming on SiC surface.)

III–V nitride semiconductors, such as InN, GaN, and A1N are useful materials in various optoelectronic device applications, since in wurtzite structure they form a continuous alloy system with direct energy gaps ranging from about 2.0 eV (for InN), 3.4 eV (for GaN) and 6.3 eV (for A1N), i.e., such an alloy system would have a continuous extent of energy-gap values throughout the visible and ultraviolet ranges. One of the major problems related to the thin-film growth (using, e.g., MOCVD method) of these compounds is the choice of a suitable substrate. For example, the sapphire, which is used widely as the substrate, has a very large lattice mismatch with III–V nitride semiconductors. However, the crystal quality of GaN layer can be substantially improved by incorporating a thin buffer layer such as A1N (or GaN buffer layer grown at lower temperatures than actual GaN layer), prior to the deposition on sapphire. For the growth of the ternary (e.g., GaInN) and quaternary (AlGaInN) compounds, it is also of great importance to select appropriate (lattice-matched) substrates (for the energy gap vs. lattice constant plot, see Fig. 6.7). The control of conductivity in these compounds is also difficult to realize because of the very high group V equilibrium vapor pressure. The *p*-type doping in GaN is not readily obtainable; however, with an appropriate post-growth treatment of Mg-doped GaN layer, sufficiently high hole concentrations can be obtained. This allows the fabrication of *p–n* junctions for LEDs with the relatively high output powers. Cubic (zincblende) GaN has some potential advantages over hexagonal (wurtzite) GaN. First of all, the value of the energy gap of cubic GaN is about 3.23 eV at room temperature, which is about 0.2 eV lower than that of hexagonal GaN. This fact is important for optoelectronic applications, since it implies that one can obtain a material for the visible spectral region (i.e., blue and green regions) with lower In content (alloyed to GaN) as compared to hexagonal case. (Note that with lower In content, reduced phase separation within the InGaN system would be expected.) Also, cubic GaN has higher carrier mobility as compared with the hexagonal case; this is due to the lower phonon scattering in

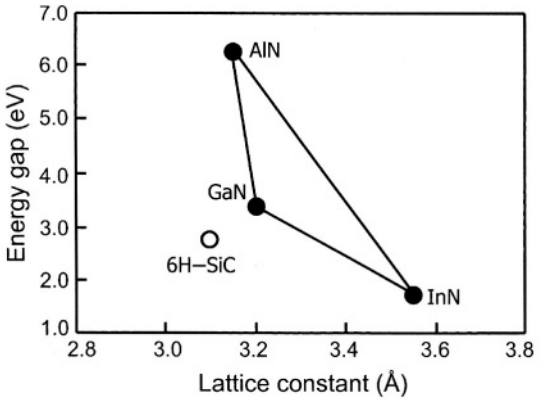

FIGURE 6.7. Plot of the energy gap vs. lattice constant for III–V nitride semiconductors and 6H–SiC.

higher crystallographic symmetry. In addition, since the epitaxially grown cubic GaN exhibits easy cleavage with respect to the substrate, it may offer better alternative material for the realization of laser diodes with cleaved cavities.

One of the main difficult problems related to the fabrication of light-emitting devices based on wide energy-gap II–VI compound semiconductors, such as ZnSe and ZnS, is the growth (e.g., MOCVD method) of *p*-type material with adequately high carrier concentrations. As mentioned above, an important limitation in device applications of such compounds is a phenomenon of self-compensation, i.e., the generation of native point defects of a type that compensate dopant atoms and thus hinder the formation of *p–n* junctions. Self-compensation in these compounds is due to the fact that anion and cation vacancies can act as acceptors and donors, respectively. For example, Zn vacancies in ZnSe and ZnS act as acceptors, whereas Se and S vacancies in ZnSe and ZnS act as donors. The effect of self-compensation can be reduced by growing material at lowest temperature possible. Some success with *p*-type doping in ZnSe was realized in Li-doped material (Li as shallow acceptor substituting for Zn), and thus *p–n* junctions for LEDs could be fabricated.

In addition to the binary wide energy-gap semiconductors discussed above, there are also some ternary semiconductors with wide energy gaps that are of interest in possible device applications. These are, e.g., I–III–VI_2 (chalcopyrite) compound semiconductors such as $CuAlSe_2$ and $CuAlS_2$ with the energy gap of 2.7 and 3.5 eV, respectively. It is interesting to note that the melting points (of about 1500 K) of these specific compounds are lower as compared to the binary compounds with the similar energy gap.

6.6. OXIDE SEMICONDUCTORS

Oxide semiconductors are also referred to as semiconductor ceramics. These are often polycrystalline and polyphase materials with grain sizes in the range between about 1 and 10 μm. The properties of the grains and grain boundaries

play a crucial role in both the understanding and applications of these materials. In this context, the control of the composition and microstructure of the grains and, especially, grain boundaries are most important issues in developing these materials. It has been established that (i) the grain boundaries generally have an associated space–charge region controlled by the defect structure of the material, (ii) the grain boundaries are paths for the rapid diffusion for various impurities, and (iii) grain boundary segregation, precipitation, and oxidation typically affect various properties of these materials.

Some examples of the oxide semiconductors (with corresponding energy gaps) are Cu_2O (2.1 eV), Bi_2O_3 (2.8 eV), ZnO (3.4 eV), SnO_2 (3.7 eV), $BaTiO_3$ (3 eV), $SrTiO_3$ (3.3 eV), and $LiNbO_3$ (4 eV). These materials are employed in a variety of electronic devices and sensors, such as (i) PTC thermistors, (ii) varistors (i.e., resistors with nonlinear, but symmetric, current–voltage characteristics) that are used for the protection of electronic devices and circuits, (iii) capacitors of high-dielectric constant that can be employed in MOS structures (in dynamic random access memory, DRAM), (iv) gas sensors, and (v) electro-optic modulators.

6.7. MAGNETIC SEMICONDUCTORS

Semiconductor compounds that contain magnetic ions (e.g., Cr, Mn, Fe, Co, Ni, and Eu) may also exhibit magnetic properties. For example, some oxide semiconductors, such as FeO, NiO, and EuO, exhibit various magnetic properties (e.g., FeO and NiO are antiferromagnetic materials, and EuO and EuS are semiconducting ferromagnetic materials). Other magnetic semiconductors include diluted magnetic semiconductors, such as $Cd_{1-x}Mn_xTe$, $Pb_{1-x}Mn_xTe$, and $In_{1-x}Mn_xAs$, in which a fraction of nonmagnetic cations is substituted by magnetic ions. Such materials exhibit both semiconducting properties and magnetic behavior. Depending on the amount of the magnetic ions substituting for nonmagnetic cations and also on temperature, diluted magnetic semiconductors can exhibit various magnetic properties.

The diluted magnetic semiconductors, such as $Cd_{1-x}Mn_xTe$, have attracted increasing interest because of some interesting properties associated with such compounds. For example, the energy gap and lattice parameters, as well as the development of various magnetic properties in such a compound can be adjusted to a desired value depending on x, and these effects can be investigated in a systematic manner. These compounds also exhibit large magneto-optical effect, which may facilitate their application in optical modulators. One of the interesting possibilities in the studies of diluted magnetic semiconductors is the growth of low-dimensional structures (using MBE) such as magnetic QWs confined by nonmagnetic barriers. The growth of QW structures based on diluted magnetic semiconductors would allow to investigate the low-dimensional magnetic phenomena. In addition, in these materials there may be a possibility of "tuning" the energy gap of a semiconductor and band offsets in heterostructures by using external magnetic fields.

For a general overview on diluted magnetic semiconductors, see Furdyna and Kossut (1988) in Bibliography Section B2.

6.8. POLYCRYSTALLINE SEMICONDUCTORS

Grain boundaries play a crucial role in determining the properties of polycrystalline semiconductors. These semiconductors can be further classified as (i) microcrystalline and nanocrystalline materials that are usually prepared as thin films and (ii) large grain materials in the form of sliced ingots and sheets. The grain size in polycrystalline materials depends on the substrate temperature during thin film growth, the thickness of the film, and also on post-growth annealing treatment of the film. As mentioned in Section 6.6, the grain boundaries generally have an associated space–charge region controlled by the defect structure of the material, and the grain boundaries are paths for the rapid diffusion of impurities affecting various properties of polycrystalline materials. An important consequence of the presence of potential barriers on grain boundaries in a polycrystalline semiconductor is the increase of its electrical resistivity. One of the important processes is the decoration of grain boundaries, i.e., the process in which precipitates of impurity elements segregate to the boundaries.

In general, the grain boundaries introduce allowed levels in the energy gap of a semiconductor and act as efficient recombination centers for the minority carriers. This effect is important in minority-carrier devices, such as photovoltaic solar cells and it is expected that some of the photogenerated carriers to be lost through recombination on the grain boundaries. Typically, the efficiency of the device will improve with increasing grain size. In this context, the *columnar grain structure* (i.e., grains in a polycrystalline material extend across the wafer thickness) is more desirable as compared to the material containing fine grains that do not extend from back to front of a device structure. In order to prevent significant grain-boundary recombination of the minority carriers, it is also desirable that the lateral grain sizes in the material be larger than the minority carrier diffusion length. It should also be mentioned that the possible preferential diffusion of dopants along the grain boundaries (and/or precipitates of impurity elements segregated at the boundaries) may provide shunting (i.e., conducting) paths for current flow across the device junction.

It should be noted that the hydrogen passivation of grain boundaries in polycrystalline silicon devices, such as photovoltaic cells, is an effective method of improving their photovoltaic performance efficiency. This improvement is associated with the mechanism similar to that of the passivation of dangling bonds in amorphous silicon (see Section 6.9). It should be added that the hydrogen passivation of other defects, such as dangling bonds at vacancies and dislocations, is also beneficial in improving the performance of photovoltaic cells.

6.9. AMORPHOUS SEMICONDUCTORS

Amorphous semiconductors have found a wide range of applications in various devices. These materials can be (relatively) inexpensively produced as thin films deposited on large-area substrates. Some common examples include

the use of amorphous selenium as a photoreceptor material in electrophotographic copiers, and of hydrogenated amorphous silicon in solar cells and flat-panel displays.

Some of the important amorphous semiconductors include *amorphous chalcogenides* (e.g., *a*-Se and $a\text{-As}_2\text{Se}_3$) and *tetrahedrally-bonded amorphous semiconductors* (e.g., *a*–Si:H). [For an introduction, see Street (1991) in Bibliography Section B2.]

As mentioned in Chapter 2, amorphous semiconductors have only short-range order with no periodic structure (see Fig. 6.8). In such cases, some information about the structure (i.e., about the atomic array, or atomic distribution) can be obtained by plotting the so-called *radial distribution function,* which is the probability, *P(r),* of finding an atom at a distance *r* from a given atom. In crystalline solids, such a radial distribution function exhibits series of sharp peaks indicative of the long-range order. The curve representing an amorphous material indicates the presence of the short-range order only. This also implies that the number of nearest neighbors to any given atom, on average, is not much different from the corresponding number in the crystalline material. In amorphous materials, there are certain bond length and bond angle variations, but on average the density is similar to that corresponding to the crystalline material. In amorphous semiconductors, defects are of different kind as compared to crystalline materials. In the case of amorphous materials, the main defects are those related to the deviations from the average coordination number, bond length, and bond angle; other defects include, e.g., dangling bonds, deviations from an optimal bonding arrangement, and microvoids.

The electronic band structure of amorphous semiconductors is substantially different from that in the crystalline semiconductors. In crystalline materials, the periodicity of the atomic structure and the presence of long-range order result in a band structure with allowed and forbidden electronic levels, with sharp band edges and a fundamental energy gap separating valence band from the conduction band. In amorphous semiconductors, there is still a fundamental energy gap based on the short-range bonding between the atoms; however, the sharp

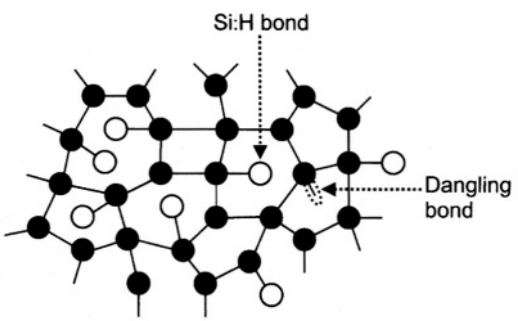

FIGURE 6.8. Schematic illustration of a two-dimensional continuous random network of atoms having various bonding coordination.

band edges of the crystalline semiconductor are replaced in the amorphous material by *exponential band tails* due to localized states related to the structural disorder (i.e., bond length and bond angle deviations that broaden the distribution of electronic states); in addition, defects (i.e., dangling bonds) introduce electronic levels in the energy gap. It should be noted that the transition from the localized states to the extended states results in a sharp change in the carrier mobility, leading to the presence of the mobility edges (for corresponding conduction and valence bands), or the mobility gap. The carrier mobility in the extended states is higher, and the transport process is analogous to that in crystalline materials; whereas in the localized states, the mobility is due to thermally-activated tunneling between the localized states (i.e., *hopping conduction*) and it is lower as compared to the extended states.

As mentioned above, in amorphous semiconductors, the allowed energy bands have band tails in the energy gap. The typically observed exponential energy dependence of the absorption edge, or exponential absorption edge (i.e., the *Urbach edge*) provides an important parameter for characterizing the material's quality and it usually depends on the deposition method and deposition conditions. Several models have been proposed to explain this widely observed behavior. In amorphous semiconductors, the shape of the absorption edge can be explained in terms, of the joint density of states (DOS) of the valence and conduction band tails. [For more details, see Street (1991) in Bibliography Section B2.] As noted above, in the amorphous material the exponential band tails are related to the structural disorder, i.e., bond length and bond angle deviations that broaden the distribution of electronic states; thus, the slope of the Urbach edge can be related to the material's quality.

An important consequence of the long-range disorder in amorphous semiconductors is that one can no longer use the periodic potential $V(\mathbf{r})$ and derive $E(\mathbf{k})$ relationship. The energy bands in this case are described by a DOS distribution $N(E)$. And since momentum conservation rules or direct and indirect optical transitions no longer apply, in the case of amorphous silicon, e.g., this results in very high absorption coefficient, allowing the use of only micrometer-scale thin films for absorption of solar energy (unlike its indirect-gap crystalline counterpart having low optical absorption).

To summarize briefly, one of the important features of amorphous semiconductors is the presence, in the energy gap of the material, of high density of defect-induced localized states that have a crucial effect on the electronic properties of these materials. These localized states make doping difficult or impossible.

Among the various amorphous semiconductors, *hydrogenated amorphous silicon* (*a*-Si:H), which contains substantial amount of hydrogen (up to about 15%), has become one of the most attractive materials for a variety of applications. The feasibility of using *a*-Si:H in such device applications as, e.g., photovoltaic solar cells and thin-film field-effect transistors (employed in liquid crystal displays and large-area detectors including medical X-ray imaging) is based on the effect of hydrogen passivation of defects (e.g., by attaching hydrogen to the dangling bonds) in amorphous silicon matrix (see Fig. 6.8). Such an effect leads to the removal of various states from the energy gap of the material; thus, doping

TABLE 6.3. The optical energy gap (E_{04}, i.e., photon energy corresponding to the absorption coefficient of 10^4 cm^{-1}) of selected amorphous semiconductors[a]

	E_{04} (eV)
a-C:H	0.5–3
a-Si:H	1.3–2.0
a-GaAs	1.0–1.5
a-Se	2.0
a-As_2Se_3	2.0

[a]In the case of hydrogenated amorphous materials, such as *a*-C:H and *a*-Si:H, the energy gap, as well as other properties, depend on the amount of hydrogen present in the material.

and the formation of semiconductor junctions and the fabrication of various electronic devices become possible. Note that *a*-Si:H network may contain a wide variety of bonds with a range of bond lengths and bond angles, and internal surfaces. In such a structure, the irradiation of the material with photons or electrons, or application of a bias, or thermal cycling may cause the generation of metastable defects induced by nonequilibrium carriers and the migration of hydrogen, leading to the generation of dangling bond defects. It should be noted that the annealing of the irradiated samples in the temperature range between about 150 and 200°C typically leads to a recovery of the physical properties.

Some of the amorphous semiconductors are listed in Table 6.3.

6.10. ORGANIC SEMICONDUCTORS

The general advantages of organic semiconductors include their diversity and relative ease of changing (accommodating) their properties to specific applications. Some examples of organic semiconductors include materials, such as anthracene, $C_{14}H_{10}$, and polyacetylene, $(CH)_n$. The electrical conductivity of polyacetylene can be varied by many orders of magnitude by doping with donors (e.g., alkali metals) or acceptors (e.g., iodine or AsF_5). Early experiments, several decades ago, have demonstrated that anthracene, as well as others, are photoconductors. In fact, it should be noted that the first practical application of anthracene (as the photoconductor) was as a photoreceptor material in imaging systems (i.e., electrophotography). Currently, various organic photoreceptors are widely employed in these applications, offering low cost and relative ease of preparation in flexible configurations.

Typically, these materials exhibit carrier trapping and low mobility, which limit their applications in electronic devices. However, some important applications of organic semiconductors with conjugated bonds are emerging in various electronic and photonic applications, such as transistors, LEDs, solar cells, and nonlinear optical materials. (Along the chain, a conjugated polymer has alternating single and double bonds between the carbons, i.e., –C = C–C = C–.)

One of the promising applications of organic semiconductors is in relatively inexpensive light-emitting diode (LED), covering whole spectrum of colors, including blue. The main advantages of organic materials in such applications include low operating voltages, color tunability, and relative simplicity of device fabrication. The actual device incorporates an organic semiconductor (i.e., light-emitting layer), which is sandwiched between two electrodes having dissimilar work functions. In such a case, light emission results from the recombination of electrons injected from a lower-work-function-electrode with holes injected from a higher-work-function-electrode into the organic layer (i.e., double injection into the light-emitting organic semiconductor). Important issues of concern in such applications are the device stability and the operational lifetime of organic LED. In this context, the control of metal/polymer interface is of great importance.

Some other applications of organic semiconductors may be realized in combination with inorganic semiconductors in hybrid inorganic/organic devices such as nanocrystal-based-quantum-dot/organic systems. For example, poly (*p*-phenylene vinylene) (PPV), which has the energy gap of about 3 eV, has been used in combination with CdSe nanocrystals in an electroluminescent structure. In this case, the luminescence is due to the recombination of holes, injected into a PPV layer, with electrons, injected into a multilayer film of CdSe nanocrystals. The luminescence spectrum (in the visible spectral region) in such systems can be adjusted by changing the nanocrystal size.

Although the inadequate stability of organic semiconductors may hinder some of their applications, their high reactivity accompanied by variations in conductivity also implies a possible application of organic semiconductors in various sensor instruments (e.g., gas sensors and biosensors).

It should be noted that there are also efforts directed at fabricating inexpensive all-polymer integrated circuits (or polymeric integrated circuits), based on field-effect transistors incorporating the conducting, insulating, and semiconducting components that are made of polymers. Such inexpensive (and relatively simple to fabricate) devices deposited on flexible substrates, such as polyimide, offer mechanical flexibility that can be employed in such high-volume applications as, e.g., product identification (i.e., electronic labels).

For a general overview on organic semiconductors and their applications, see, e.g., Farges (1994) in Bibliography Section B2.

6.11. LOW-DIMENSIONAL SEMICONDUCTORS

Recent developments related to nanoengineered materials have demonstrated that the nanostructured semiconductors offer increasingly greater flexibility and control in designing various nanoscale structures and devices. In this context, the main motivation is related to continuous trends towards (i) increasing miniaturization of various structures and devices, (ii) improving dimensional precision, and (iii) controlling and designing various materials properties.

One of the important features of nanostructures (with typical sizes in the range between about 1 and 50 nm) is the flexibility of controlling (and designing)

the properties of such materials by controlling the sizes of nanostructures. Such nanostructures exhibit structural, optical, and electronic properties that are unique to them and that are different from both macroscopic materials and isolated molecules.

Nanostructures have dimensions in the range between about 1 and 50 nm. In this range, the properties of semiconductors are modified and correspond to those that are characteristic of the quantum-mechanical electronic confinement, and they exhibit fundamentally different properties as compared to the bulk structures. The characterization of such nanoscale structures can be accomplished by using various scanning probe microscopies (SPM), as well as electron microscopy techniques and optical spectroscopy methods (see Chapter 7).

As noted earlier, in nanoscale structures with the dimensions commensurate with the de Broglie wavelength of the charge carriers, the electronic energy levels exhibit quantum confinement effects. Low-dimensional structures (see Fig. 6.9) include (i) QWs, where the charge carrier motion is allowed in two dimensions only (referred to as one-dimensional confinement), (ii) *quantum wires* where the charge carrier motion is allowed in one dimension only (referred to as two-dimensional confinement), and (iii) *quantum dots* (QDs), where the charge carrier motion is allowed in zero dimensions (referred to as three-dimensional confinement). For these low-dimensional structures, Fig. 6.9 shows their corresponding DOS as a function of energy. It should be noted that electrons propagating in the QW are also referred to as a *two-dimensional electron gas,* and those propagating in the quantum wire are called a *one-dimensional electron gas.*

The DOS (for a conduction band) in a three-dimensional system, i.e., a bulk material is a parabolic function of energy (see Chapter 4), and can be written as

$$g_{\mathrm{n}}(E) = 4\pi(2m_e^*/h^2)^{3/2}(E - E_{\mathrm{c}})^{1/2} \tag{6.11.1}$$

For the two-dimensional system (i.e., QW) and one-dimensional system (i.e., quantum wire), quantum confinement effect results in the presence of discrete sub-bands (see Fig. 6.9). The DOS in these cases, i.e., quantum wires (1D), QWs (2D), and bulk (3D), can be related to energy, in a general form, as $g(E) \propto E^{D/2-1}$,where D is the dimensionality of the system. In QWs ($D = 2$) and quantum wires ($D = 1$), the DOS in each of the sub-bands can be described by $E^{D/2-1}$dependence (see Fig. 6.9).

In QWs, the carriers (i.e., electrons and holes) are confined in one dimension only within the well thickness, but in the plane of the confining layer the behavior of carriers corresponds to that of a bulk material. The electron energy levels in a QW can be found from the solution of the Schrödinger equation (see Section 3.3), which, for an infinitely deep potential well, have a form

$$E_n = \frac{\hbar^2\pi^2}{2m_e^* L^2}\, n^2 \tag{6.11.2}$$

where $n = 1,2,3,\cdots$; the dependence of the energy levels on $(1/L)^2$ is the *quantum size effect.* The dispersion relation for the conduction band in this case can be

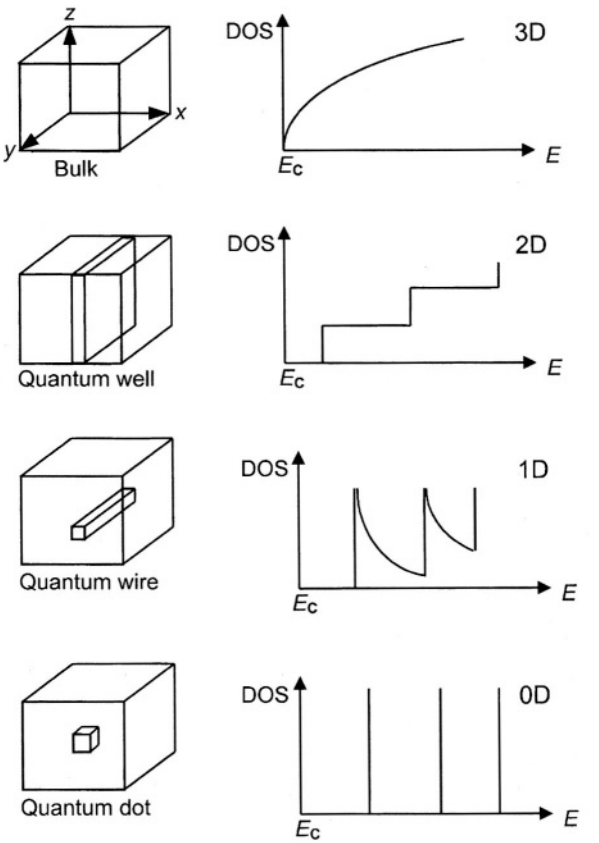

FIGURE 6.9. Schematic representation of low-dimensional structures and their corresponding DOS as a function of energy.

expressed as (e.g., for the infinite motion in y–z plane and the finite motion in the confinement direction corresponding to x-axis)

$$E(k) = E_c + E_{n_x} + \frac{\hbar^2(k_y^2 + k_z^2)}{2m_e^*} \tag{6.11.3}$$

where E_{n_x} is given by Eq. (6.11.2), with L_x being the width of the well

$$E_{n_x} = \frac{\hbar^2\pi^2}{2m_e^* L_x^2} n^2 \tag{6.11.4}$$

For the conduction band, the DOS per unit area in a two-dimensional system (i.e., QW) for one sub-band can be expressed as (for energies greater than $E_c + E_{n_x}$)

$$g(E) = \frac{m_e^*}{\pi\hbar^2} \tag{6.11.5}$$

Thus, in this case, for each quantum number n_x, the DOS is constant (i.e., it is independent of energy), and the overall DOS is the sum of these for all values of n_x, resulting in a staircase-type distribution (see Fig. 6.9), with a step height given by Eq. (6.11.5).

The quantum wire essentially behaves as a potential well that confines carriers (both electrons and holes) in two directions. For the conduction band, the dispersion relation is

$$E(k) = E_c + E_{n_x} + E_{n_z} + \frac{\hbar^2 k_y^2}{2m_e^*} \tag{6.11.6}$$

In this case, the DOS as a function of energy in each of the sub-bands can be described by $E^{-1/2}$ dependence (see Fig. 6.9).

In zero-dimensional structure (i.e., QDs), the carriers are confined in three directions, and the expression for energy is

$$E(k) = E_c + E_{n_x} + E_{n_y} + E_{n_z} \tag{6.11.7}$$

where E_{n_y} and E_{n_z} have the same form as E_{n_x} in Eq. (6.11.4). The DOS in this case is described by a set of discrete δ-functions as shown in Fig. 6.9. QDs are also referred to as artificial atoms.

It should be noted that the DOS for the low-dimensional structures, shown for the conduction band in Fig. 6.9, has a similar distribution in the valence band.

In general, various types of nanoscale structures, shown in Fig. 6.10, include (i) nanocrystallites with an aspect ratio from unity to infinity, (ii) lithographically produced nanostructures (e.g., nanowires and nanodots), (iii) multilayers (e.g., semiconductor superlattices), and (iv) nanograined (nanocrystalline) materials

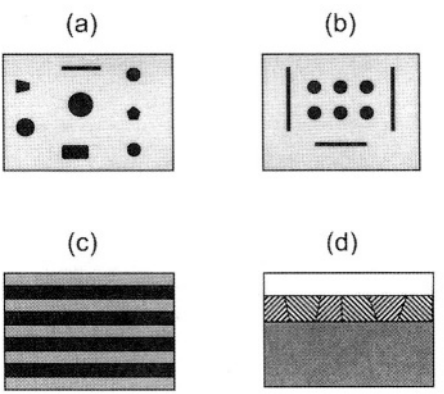

FIGURE 6.10. Schematic illustration of different types of nanostructures: (a) nanocrystallites with an aspect ratio from unity to infinity, (b) lithographically produced nanostructures (e.g., nanowires and nanodots), (c) multilayers (e.g., semiconductor superlattices), and (d) nanograined (nanocrystalline) materials (e.g., coatings or buried layers).

(e.g., coatings or buried layers). One of the important features of semiconductor nanostructures (with typical sizes in the range between about 1 and 50 nm) is the flexibility of controlling (and designing) the properties of such materials by controlling the sizes of nanostructures. For example, the energy gap of semiconductors (e.g., Si, CdSe, and InAs) increases with size r as $1/r^2$. As mentioned above, such nanostructures exhibit structural, optical, and electronic properties that are unique to them and that are different from both macroscopic materials and isolated molecules.

The continuing advances in semiconductor growth and materials processing techniques catalyzed research and possible applications of low-dimensional systems, such as QWs and superlattices. By using MBE or MOCVD techniques for growing epitaxial layers, it is possible to construct artificial structures in which layers of different materials and thickness alternate. Such structures, in which the period of the layers may be as small as a few monolayers, have the properties of two-dimensional systems. The QW is essentially a double heterojunction structure consisting of a layer (about 10 nm) of a narrower energy-gap semiconductor surrounded by wider energy-gap semiconductor (e.g., AlGaAs/GaAs/AlGaAs) and having a sharp compositional transition at the interfaces (see Fig. 6.11). This is the type-I (also referred to as straddling) alignment with such band line-up that the energy minima for both electrons and holes occur in the narrower energy gap material, i.e., both the electrons and holes are confined in the same layer. In type-II system (also referred to as staggered), the energy band line-up is such that the electrons are confined in one layer, whereas the holes are confined in a different layer. In type-III (also referred to as broken-gap) alignment, the energy gaps of different layers do not overlap.

For realistic QWs, the case of a bound electron in a finite potential well is considered, implying that the wave functions decay away exponentially into the

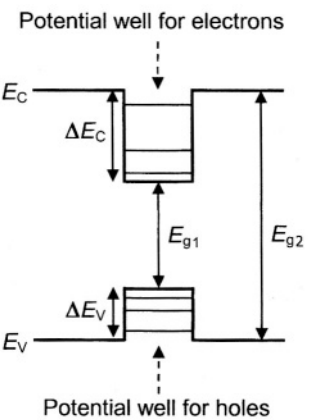

FIGURE 6.11. The energy band diagram of a single-quantum-well structure; e.g., thin (about 10 nm) GaAs layer (with narrower energy gap E_{g1}) between $Al_xGa_{1-x}As$ confining layers (with wider energy gap E_{g2}). The conduction and valence band offsets, ΔE_c and ΔE_v are also shown.

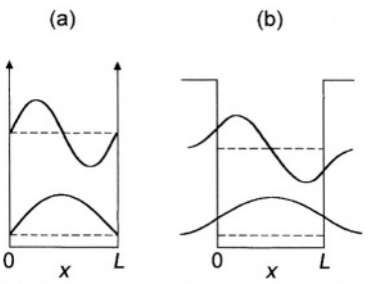

FIGURE 6.12. Schematic diagram of a ground state and first excited state energy levels and associated wave functions for (a) an infinitely deep square potential well and (b) a square potential well of finite depth.

potential barriers (see Fig. 6.12). This penetration through barriers by tunneling constitutes the basis for coupled QW systems, such as the coupled QW (i.e., two QWs separated by a thin barrier). In such quantum structures, the properties of the well, such as the values of the energy levels, depend on the potential well depths (the conduction and valence band offsets, ΔE_c and ΔE_v), QW width and on the effective masses of the confined carriers. The distinct cases of QW structures are *single-quantum well,* SQW (i.e., a structure with one QW) and *multiple-quantum well,* MQW (i.e., periodic arrays of QWs). If the wells in MQW structure are separated by relatively thin barriers (between 1 and 10 nm), the confined states in each well interact through the barrier by tunneling, resulting in the broadening of the discrete energy levels and the formation of the *minibands* extending through the structure (see Fig. 6.13). Such a structure is called a *superlattice,* which has essentially "engineered" properties corresponding to a one-dimensional crystalline material with a new electronic band structure, which can be modified by changing the composition and thickness of the layers. The term superlattice signifies the fact that extra periodicity is present (in addition to crystalline one) with a longer period. Also, in this case the periodic potential is relatively weaker, resulting in a band structure on a relatively lower energy scale; thus, such bands are termed minibands (in contrast with the bands due to the regular crystalline structure of atoms). This structure is also referred to as a

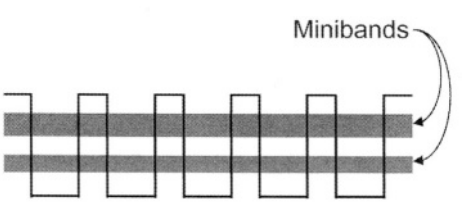

FIGURE 6.13. Schematic diagram of a superlattice structure (e.g., GaAs/AlGaAs); the gray bands correspond to miniband energy ranges.

compositional superlattice. It should be noted that it is also possible to form a periodic QW structure by two identical but differently doped semiconductors; such a structure is called a *doping superlattice,* or *nipi structure,* in which intrinsic layers, *i,* are present between the *n*- and *p*-type layers. Some advantages of doping superlattices (compared to heterojunction compositional superlattices) include (i) absence of possible problems related to the lattice-constant mismatch and (ii) the fact that, in principle, there is no restriction on types of materials, provided those can be grown with *n*-type and *p*-type doping.

It is important to emphasize that a more detailed explanation of experimental observations (e.g., luminescence) in such structures requires considerations regarding various levels associated with both light-hole and heavy-hole states in the valence band (see Fig. 6.14). To summarize briefly, the importance of such QW structures for optoelectronic applications is based mainly on the facts that (i) the emitted wavelength of a device can be tailored for a specific application by an appropriate choice of QW parameters (note that this is accomplished without changing the materials composition) and (ii) the confinement of both the electrons and holes in the narrow layer results in the formation of strongly bound excitons, which tend to recombine radiatively, thus leading to high radiative recombination efficiency. Some useful applications of semiconductor QWs are based on their optical properties in devices such as semiconductor lasers and photodetectors incorporating QWs in their active layers.

QD is a fragment of matter that is smaller than the electronic Bohr radius in all three coordinates. As mentioned earlier, in QDs, the electronic transitions are discrete, and, most importantly, they are tunable as a function of size. Typically, QDs have dimensions in the range between about 1 and 50 nm. An important issue

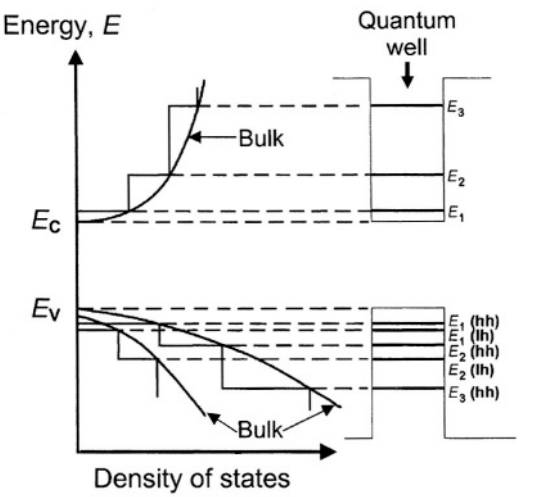

FIGURE 6.14. Schematic diagram of the DOS with corresponding energy levels for both conduction and valence bands (including both light holes lh and heavy holes hh) in a QW. The DOS corresponding to the bulk material is also shown (i.e., the solid lines corresponding to the parabolic dependence).

in such materials is to elucidate the effect of surfaces on their properties, which are very sensitive to surface properties due to the high surface to volume ratio. Surface states typically have the dominant role in these systems, acting as efficient traps for electrons and holes, and thus, surface passivation is essential for fabricating practical semiconductor devices based on such low-dimensional structures. Possible technological applications of semiconductor QDs are based on their optical properties (e.g., the emission wavelength controlled by the crystallite size) and on nonlinear optical properties. Preparation of QDs with a narrow size distribution and designed surface properties are also essential for their applications.

Several types of nanoscale synthesis methods include (i) lithography and patterning using electron beams, X-rays, or ions, (ii) methods employing self-assembly, and (iii) nanometer-scale fabrication of semiconductor using SPM techniques. In the former case (i.e., conventional lithography), the process starts with the deposition of thin films, e.g., using MBE or MOCVD that allow nanoscale control in the vertical (growth) direction, followed by lithographic processes for nanoscale patterning in lateral dimensions. In the self-assembly process, atoms and molecules are arranged into organized structures due to thermodynamic considerations. Self-assembly process may occur due to the interactions between adsorbed molecules on solid surfaces, or self-assembled nanocrystals can be prepared by precipitation in solution involving rapid introduction of reagents into a hot solvent resulting in growth of the nanocrystallites (i.e., colloidal QD formation). Nanometer-scale fabrication of semiconductor using SPM techniques (see Section 7.4.3) is based on employing various types of tip–substrate interactions, such as attractive and repulsive forces, electron beams, and electric fields. For example, in the tunneling mode, at sufficiently close separation between the tip and the sample surface and when a positive bias is applied, tunneling of electrons from surface-bound atoms to the tip allows selective removal of surface atoms, followed by the transfer of tip-adsorbed atoms to the surface when a negative bias is applied (i.e., in the field-emission mode). In addition, the STM tip can also produce a focused (on nanometer scale) beam of electrons (with energies up to about 20 eV), which can induce various processes such as chemical reactions, bond breaking, and surface migration. Employing parallel processing with STM tip arrays can facilitate the practicality of STM-based fabrication processes. Note, however, that an array of individually addressable 10^4–10^6 tips or more could be required to provide high-throughput nanolithography tool. Some of the main issues of concern in this case include uniformity, placement accuracy, and reliability of the technique.

Possible applications of semiconductor nanostructures are based on the optical properties (e.g., the emission wavelength is controlled by the crystallite size), and electronic phenomena related to, e.g., *single electron tunneling* and *resonant tunneling.*

Single electron devices are based on localization of single electrons on nanoscale device elements, which are coupled by tunneling barriers. These devices, which are based on the *Coulomb blockade,* operate by employing controlled electron tunneling across the potential barriers. This phenomenon is typically observable in the QD (or as often referred to as an island), in which the addition of an uncompensated electron charge produces an electric field that prevents the addition of the successive

electrons due to the Coulomb repulsion. Thus, the addition of an extra electron to the island necessitates overcoming an energy barrier. The application of a sufficient potential makes it possible for a single electron to overcome the respective energy barriers, resulting in a current flow. Further increases of the potential lead to additional current flow, which is depicted by the *Coulomb staircase* (representing the dependence of current on voltage) indicating a stepwise transfer of electrons (in increments of a single electron) from source to drain. This is due to the fact that tunneling is a discrete process, and thus, the charge across the potential barriers flows in multiples of the charge of an electron. The important issues related to the practical applications of such nanoelectronic devices are realizing room-temperature devices and economically viable fabrication methodologies. The realization of the room-temperature devices necessitates the value of thermal energy (i.e., $k_B T$) being much less compared to the Coulomb blockade energy (i.e., $E_C = e^2/2C$, where C, the effective capacitance of the island, is dependent on the size of the island and its distance to electrodes), which can be increased by reducing the island size. Thus, in this case, the energy, E_C, required to add an extra electron to the island, must be much greater than $k_B T$. (Typically, room temperature operation requires an island size being in the range between about 1 and 3 nm.)

The *resonant tunneling diode* (RTD) consists of an emitter and collector regions, and a double-tunnel-barrier structure that contains a QW. Such a structure is shown in Fig. 6.15, which presents the energy band diagram of a QW structure consisting of the GaAs layer (narrower energy gap) sandwiched between the layers of AlGaAs (wider energy gap). In such a structure, each layer is about 5–10 nm thick, thus the tunneling through such a thin AlGaAs layer can readily occur. Note that the outer n^+-GaAs layers are heavily doped (to provide electrons). As discussed in Section 3.3, because of the boundary conditions, only certain (discrete) energy levels are allowed in such a QW. In such a structure, if initially the electron ground state in the GaAs QW is above the Fermi level of the n^+-GaAs layer [see Fig. 6.15(a)], no significant tunneling is expected through AlGaAs layers. However, if a sufficient voltage is applied in order to elevate the Fermi level of the n^+-GaAs layer (at the left of the diagram) into alignment with the electron ground state (in the GaAs well), the latter now becomes accessible for tunneling [see Fig. 6.15(b)]. This is referred to as *resonant tunneling,* and as electrons subsequently tunnel out of the GaAs well through the AlGaAs barrier (at the right of the diagram), this results in the current flow in the diode. With further increase in the applied voltage, since the Fermi level (at the left of the diagram) is no longer in alignment with the electron ground state in the GaAs well (the Fermi level is now at higher energy) and since no other states are accessible for tunneling, the current through the device decreases [see Fig. 6.15(c)], resulting in the so-called *negative resistance* region. With further increase in the applied voltage, the condition for the resonant tunneling will again occur provided there are other allowed states in the well. Thus, to summarize briefly, the basic principle of the operation of RTD is based on the fact that electrons can only traverse from the emitter to the collector if they are lined up with a resonant energy level. The main application of the RTD is in *multiple-state memory devices* having multiple peaks in their $I(V)$ characteristic.

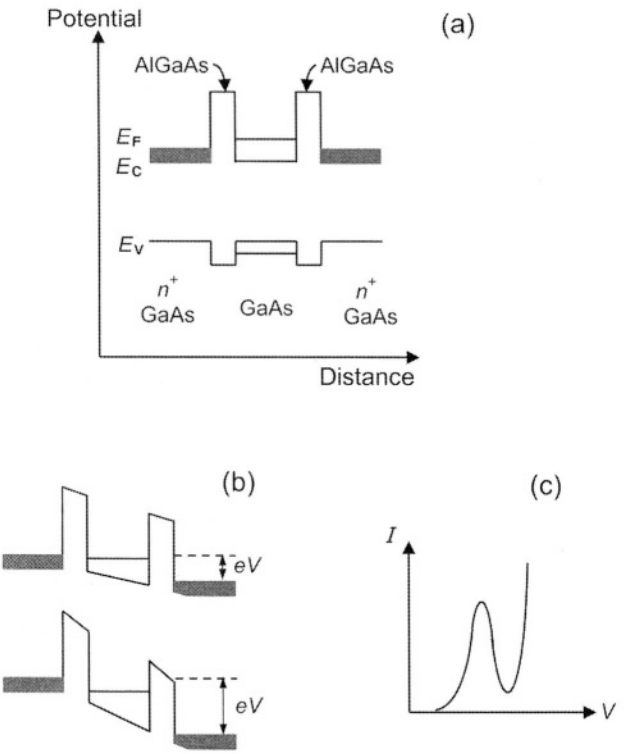

FIGURE 6.15. Illustration of resonant tunneling: (a) a QW structure consisting of a narrower energy-gap GaAs sandwiched between wider energy-gap AlGaAs barriers, (b) QW structure under applied voltage, and (c) current–voltage characteristic of the RTD, showing the region of negative differential resistance.

The major issues related to semiconductor nanostructures include (i) control of size, composition and assembly of nanostructures, (ii) elucidation of the role of surfaces and interfaces, (iii) advances in nanofabrication and nanocharacterization techniques, (iv) reproducible preparation of nanostructures, (v) stability (e.g., chemical and thermal) of nanostructures, and (vi) control in statistically driven processes.

For a general overview of low-dimensional semiconductors, see, e.g., Davies (1998), Kelly (1995), and Weisbuch and Vinter (1991) in Bibliography Section B2.

6.12. CHOICES OF SEMICONDUCTORS FOR SPECIFIC APPLICATIONS

In choosing a semiconductor for a specific application, one has to consider the relevant semiconductor properties. The band-structure parameters and transport properties of a semiconductor have crucial effect on the operation of

semiconductor devices. These parameters are, e.g., the energy gap and type (i.e., direct or indirect), the effective mass of carriers, the intrinsic conductivity, and the (low-field) carrier mobility. Other important materials properties are, e.g., the saturation electron drift velocity, the breakdown electric field, the thermal conductivity, and the mechanical strength.

Although for many applications semiconductors must be grown as nearly perfect single crystals, in some cases polycrystalline and amorphous semiconductors can also be utilized. Thus, the cost of the devices can be substantially reduced, and, as in the case of amorphous semiconductors, large-area devices (e.g., solar cells, displays, and imaging devices) on various substrates can be produced.

An ability to control the electrical conductivity (by such means as doping, excess charge carrier injection, and optical excitation) over orders of magnitude in semiconductors offers unique applications of these materials in various electronic devices. Also, by "engineering" the band structure, or by "tuning" the energy gap by alloying various semiconductors, or simply by choosing a semiconductor with desired properties (e.g., energy gap and doping), one can design various semiconductor devices with specific characteristics and applications.

Examples of some of the most common applications of selected semiconductors are listed in Table 6.4, which refers to the applications that are facilitated by the present state of technology of both the materials synthesis and device fabrication. Note that in many applications, semiconductors are grown as thin films on various substrates.

As mentioned in Chapter 1, Si and GaAs are some of the most important materials for electronic devices due to the advanced Si-related fabrication technology and some specific properties (e.g., high carrier mobility and direct energy gap) of GaAs. It should be emphasized, however, that, in general, in order to evaluate a semiconductor for specific device applications, one has to consider all relevant materials and device parameters. As noted above, in addition to the parameters mentioned (e.g., the energy gap and type, the intrinsic conductivity, and the carrier mobility), parameters such as the breakdown electric field strength and thermal conductivity are of great importance. For example, comparing Si and GaAs again, in some applications involving small devices, an efficient heat dissipation (i.e., high thermal conductivity, which is about three times larger in Si than in GaAs) may become of paramount importance. The same applies to other materials that may not be as developed as Si, e.g., from the view of materials purity or defects, but other superior properties (or combination of application-relevant properties) make them more advantageous, as compared to either Si or GaAs, for specific applications.

6.13. SUMMARY

There is a wide range of semiconductor types. These can be classified on the basis of various criteria, such as the grouping of elements in the periodic table, or on the basis of the magnitude of the energy gap that defines many applications of a

TABLE 6.4. Common applications of selected semiconductors[a]

Material	E_g (eV)	Major applications (the present technology)
Ge	0.67	Photodetectors, substrate for active devices
Si	1.12	Integrated circuits, photodetectors, solar cells, substrate for active devices
a-Si:H	1.7	Applications in large area and/or flexible substrates, photoreceptor for electrophotography, solar cells, thin-film transistor in liquid crystal displays, imaging devices
SiC (W, 4H)	3.26	High-power devices, high-frequency power devices,
SiC (W, 6H)	3.0	High-temperature devices, optoelectronic devices (Blue light emitting diodes, UV photodetectors)
InSb	0.17	Infrared photodetectors
InAs	0.36	Infrared photodetectors
InP	1.35	Optoelectronic devices, microwave devices, substrate in heterostructures
GaAs	1.42	High-speed devices, optoelectronic devices, solar cells, microwave devices, substrate for active devices
GaP	2.27	Light-emitting diodes (red, yellow, and green)
GaN	3.44	Light-emitting diodes (blue and green)
CdTe	1.56	Solar cells
CdS	2.42	Solar cells
ZnSe	2.7	Blue light-emitting devices, protective windows, optical components
ZnS	3.68	Infrared windows, optical elements, phosphors in lighting and various display applications (CRT and flat panel displays)
PbSe	0.28	Infrared optoelectronics (diode lasers and photodetectors)
PbS	0.41	Infrared optoelectronics (diode lasers and photodetectors)
$Hg_{1-x}Cd_xTe$	0–1.56	Infrared photodetectors (2–16 μm), thermal imaging systems
$Al_xGa_{1-x}As/GaAs$	1.42–2.16	Lasers, component in modulated-doped field-effect transistor (MODFET), high-electron-mobility transistor (HEMT), heterojunction bipolar transistor (HBT), Solar cells
$GaAs_{1-x}P_x$	1.42–2.27	Light-emitting diodes (red through to green)
$Ga_xIn_{1-x}As$	0.36–1.42	Lasers, photodetectors, fiber-optic communication (at 1.55 μm)
$Ga_xIn_{1-x}As_yP_{1-y}/InP$	0.75–1.35	Lasers, photodetectors, fiber-optic communication (at 1.3 and 1.55 μm)

$Al_xGa_{1-x}As$: $E_{g,\,dir} = 1.424 + 1.247x$ $(0 < x < 0.45)$
$GaAs_{1-x}P_x$: $E_{g,\,dir} = 1.424 + 1.150x + 0.176x^2$ $(x < 0.45)$
$Ga_xIn_{1-x}As$: $E_{g,\,dir} = 0.360 + 1.064x$

Note that in some cases the lattice matching of these ternary and quaternary compounds to their corresponding substrates is possible only for specific x (and y), as shown in Fig. 6.3.

[a] E_g is the energy gap (300 K). In many applications, semiconductors are grown as thin films on various substrates. Note that in the case of *a*-Si:H, the energy gap depends on the concentration of hydrogen. In ternary $Al_xGa_{1-x}As$ and $GaAs_{1-x}P_x$ and quaternary $Ga_xIn_{1-x}As_yP_{1-y}$ semiconductors, the energy gap is "tunable" to a desired value depending on x and y; these ternary and quaternary semiconductors are often grown as epitaxial layers in heterojunction systems on substrates such as GaAs or InP, e.g., $Al_xGa_{1-x}As$ on GaAs, or $Ga_xIn_{1-x}As_yP_{1-y}$ on InP. The $Hg_{1-x}Cd_xTe$ epitaxial layers can be grown on CdTe or lattice-matched $Cd_{1-x}Zn_xTe$ substrates. For specific details on ternary and quaternary compounds, see Madelung (1996) and Berger (1997) in Bibliography Section B2

particular semiconductor, or on the basis of the structure (e.g., crystalline, polycrystalline, or amorphous).

The choice of a semiconductor for a specific application depends on various parameters, such as (i) the magnitude of the energy gap and its type (i.e., direct or indirect), (ii) the effective mass of carriers, (iii) the intrinsic conductivity, (iv) the (low-field) carrier mobility, (v) the saturation electron drift velocity, (vi) the breakdown electric field, (vii) the thermal conductivity, and (viii) the mechanical strength. In addition, issues such as the cost and the need for large-area devices also determine the particular choice of the material. For example, although in many cases single crystalline semiconductors are required, in some applications polycrystalline and amorphous semiconductors can be employed, resulting in substantial cost reductions.

Various semiconductor compounds are commonly employed in optoelectronic applications, such as light-emitting devices and radiation detectors. In such applications, it is the energy gap of a semiconductor that determines the energy of the emitted or absorbed electromagnetic radiation. In this context, it should be re-emphasized that the choice of an appropriate semiconductor for a specific application depends on other properties as well, and also on the availability of a synthesis technique that would yield a material of sufficient quality and quantity. It should also be noted that, in the ternary and quaternary alloys, the energy gap is "tunable" by alloying various semiconductors; this offers the flexibility of producing materials with desired properties by varying the composition of the alloy. The energy-band-gap engineering can also be realized by employing low-dimensional semiconductor structures, such as QWs, quantum wires and QDs; in these cases, the materials employed remain the same, and the semiconductor properties are modified by varying their size.

The basic properties of a wide range of semiconductors are summarized by Madelung (1996) and Berger (1997) (see Bibliography Section B2.)

PROBLEMS

6.1. Describe the factors that make Si the most attractive material in electronic device applications.

6.2. Outline the main advantages and applications of wide energy-gap semiconductors.

6.3. Discuss the main advantages and disadvantages of using amorphous semiconductors in various electronic device applications.

6.4. Outline some basic criteria for selecting a semiconductor for specific device applications.

7

Characterization of Semiconductors

7.1. INTRODUCTION

The developments and applications of various materials and device characterization techniques have contributed greatly to the continuing advances in semiconductor technology. The objective of this chapter is to outline some of the commonly employed techniques for the characterization of semiconductors and semiconductor devices. For details on a wide range of characterization techniques, which cannot be covered in the limited format of this book, the reader is referred to Bibliography Section B3. (For comprehensive reviews on various techniques, see, e.g., Grasserbauer and Werner, 1991; Brundle *et al.,* 1992; Schroder, 1998.)

Any adequate description of a semiconductor should in principle include:
(a) the electronic band structure;
(b) the chemical composition;
(c) the crystallographic structure;
(d) electrical and optical properties;
(e) possible presence of various defects.

For detailed analyses of these properties, a wide variety of techniques have been developed. These techniques provide complementary information related to the material's physical, structural, and device properties, and the different types of information obtained from the same sample can be used for interpreting the relationship between the synthesis and processing, the properties, and the applications of the material.

It is of great importance to determine the variations of the properties of interest throughout the material with highest spatial resolution possible. Some analytical techniques permit analysis of various properties as a function of depth (in some cases, with very high spatial resolution, as well as high depth resolution). The practical usefulness of each technique depends mainly on the sensitivity and resolution (both spatial and depth).

A selection of any specific technique for the characterization of semiconductors is based on several criteria. These include the following:

(a) the type of information that is obtainable;
(b) the sensitivity;
(c) quantifiability of the analysis;
(d) the depth of analysis (i.e., surface, subsurface, or bulk analysis);
(e) the spatial and depth resolutions;
(f) the data acquisition and analysis time;
(g) whether the analysis method is destructive or nondestructive;
(h) whether the method is contactless or requires processing (e.g., metallization);
(i) the cost.

In general, the characterization of materials is based on the analysis of effects produced by excitation of materials using methods such as irradiation with various beams (e.g., photons, electrons, ions, or neutrons), or application of fields to solid surfaces. The effects, or responses, produced by the interaction of the excitation probe with the solid, provide useful information about a material. The signals containing the information of interest about a semiconductor under excitation may be formed by various backscattered and secondary photons and particles, diffracted waves, or processes occurring in the bulk of a semiconductor (see Fig. 7.1); note that this is a general diagram, and usually only one type of excitation is used, and not all the signals are produced by a particular beam.

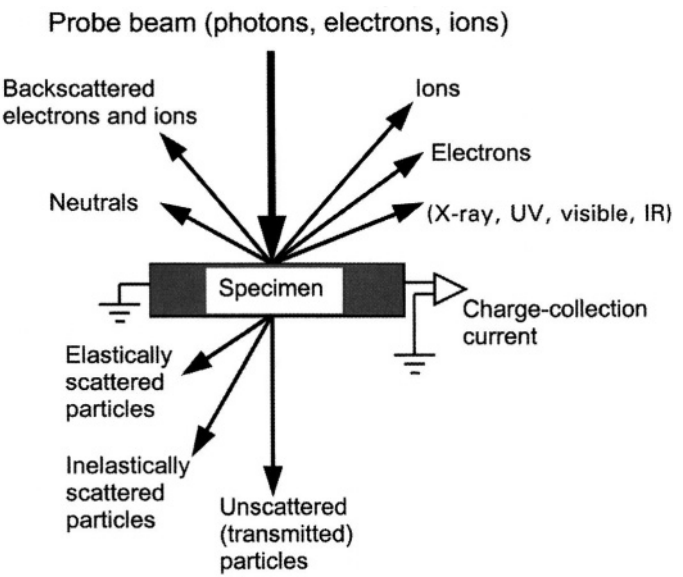

FIGURE 7.1. Schematic diagram of types of major signals produced as a result of the interaction of the primary beam (e.g., photon, electron, or ion) with a semiconductor.

To summarize briefly, the *excitation* of a semiconductor produces a *secondary effect* that can be analysed by a detector system monitoring a *specific variable.* Some examples of the excitation and secondary effect are photon, electron, ion, X-ray, atom, field, sound, and heat, whereas the variables monitored are intensity, energy (or wavelength), time, angle, mass, position, and temperature.

It should be noted that various characterization techniques are of great importance in failure analysis and quality control of semiconductor materials and devices. In general, two basic failure modes are associated with the manufacturing process-induced defects and device degradation in service, which in some cases may be interdependent. The fabrication (and processing)-induced defects may be related, e.g., to the mask misalignment, mechanical damage, and contamination. The device degradation and failure in service may be associated, e.g., with high current densities, localized high voltages, thermal cycling and vibration, corrosion, and high-temperature operation.

The term used in the semiconductor industry for measurements employed in process control is *metrology* (defined as the science of measurement). The term *characterization* is often used interchangeably with the term metrology. In this context, another term that is often used is *critical dimension* measurement (e.g., for monitoring lithographic process).

Some characterization techniques can be employed *in situ* (i.e., during the materials preparation, or device processing, or device operation), whereas others can only be used *ex situ* (i.e., analysis is performed after the growth or preparation of the materials or devices). It should be emphasized that in certain cases it is of great advantage to use *in situ* techniques, since they offer the analysis of a semiconductor material during its preparation or of a semiconductor device during its operation.

There is also an important category of characterization techniques, i.e., *microscopy techniques,* which provide a means of *microcharacterization* (and, in some cases, *nanocharacterization* with spatial resolution in the range between about 0.1 and 10 nm) for materials and devices. The demand, development, and applications of such microscopy techniques with high spatial resolution are inevitable, since increasingly complex submicron structures and devices are being produced and investigated. The spatial resolution of a technique is often related to the beam (or probe) size of the excitation source. There are, however, several factors affecting the spatial resolution. For some techniques, e.g., the beam-sample interactions may produce the so-called generation (or excitation) volume that may be substantially larger than the excitation probe size (i.e., spot at surface of irradiation) and from which the signal of interest is generated.

Another important category of characterization techniques is that related to *spectroscopy techniques.* The modern generation of spectroscopy techniques use various forms of energy to produce spectra that can be employed for both qualitative and quantitative analyses of a wide range of materials properties. Although the early spectroscopy measurements referred to studies with various types of electromagnetic radiation, other types of spectroscopy, such as electron and mass spectroscopies, have been developed in recent decades. Optical spectroscopy techniques that are widely used in the analysis of semiconductor properties include

absorption and emission spectroscopies, and Raman scattering. Other methods, such as electron and mass spectroscopies, are of great importance in the analysis of a wide range of compositional and chemical properties of materials.

In this section, first, general characterization categories will be outlined briefly, together with a comparison of some advantages and disadvantages of different techniques based on the excitation sources. This will be followed in the subsequent sections by the description of specific characterization methods, such as electrical, optical, microscopy, structural, and surface techniques.

Electrical measurements on semiconductors and semiconductor devices are routinely performed for the analysis of a wide range of semiconductor properties. Transport properties of semiconductors are characterized using techniques such as *resistivity* (*conductivity*), the *Hall effect*, and *capacitance-voltage* measurements. One important distinction between these measurements and some spectroscopic techniques is that the electrical measurements can often provide an indirect means of determination of the desired properties (e.g., related to defects), whereas the microscopic and spectroscopic techniques provide a direct measurement of semiconductor characteristics such as defect and doping concentrations.

Optical characterization methods employ the interaction of electromagnetic radiation with the solid that leads to the formation of various signals that result from processes such as absorption, reflection, emission, and scattering. A wide variety of possible configurations of both light sources and optical instruments are available for the characterization of semiconductors. The main advantages of optical techniques, as compared to those employing charged particle excitation beams, include the ability to analyze specimens in air and the absence of charging in insulating materials.

Electron-beam excitation of a semiconductor produces a variety of signals that can be used in deriving information on structural, compositional, and electronic properties of the material. The signals produced as a result of the interaction of primary (i.e., incident) electrons with the solid provide the basis for the various modes in a *scanning electron microscope* (SEM); in thin specimens, the signals produced by transmitted electrons provide modes in a *transmission electron microscope* (TEM); and the *Auger electrons* provide the basis for *Auger electron spectroscopy* (AES). One of the major advantages of using an electron beam is the easiness of obtaining finely focused probes by using electromagnetic lenses. Some of the disadvantages of the electron beam techniques are the need for a vacuum environment, the sample preparation, the charging of insulating specimens, and potential electron beam-induced heating and damage of samples. In addition, prolonged electron bombardment of the specimen may result in surface contamination, i.e., the deposition of a layer consisting primarily of polymerized hydrocarbons; this is due to the migration, on the specimen surface, of the adsorbed hydrocarbon molecules (from the pump oil, or from previous sample preparation steps) to the irradiated region where these molecules are polymerized by electron bombardment.

Ion beam techniques are distinguished between those that employ the excitation ion beam energies in the keV range and those that employ the ion beams in the MeV range. An important technique that is extensively employed in the analysis of

semiconductors is *secondary ion mass spectrometry* (SIMS). In this technique, secondary ions emitted as a result of the incident ion beam bombardment (with energies up to about 20 keV) are identified using a mass-spectrometer, and the ion beam sputtering of the sample allows depth profiling. High-energy ion beam techniques employ ion beams in the MeV range for excitation of the material. Among these, *Rutherford backscattering spectrometry* (RBS) is a nondestructive depth profiling technique that employs the measurement of the backscattered ion energy, providing information about the mass and the depth of the target nucleus (i.e., the composition and depth profile) of elements in the solid.

In contrast to the methods employing the irradiation of the material with excitation probes that were mentioned above, a variety of techniques employing the *tunneling of electrons* from the material are used in the nanocharacterization of a wide range of materials. These include *scanning tunneling microscopy* (STM), which is capable of imaging objects on the atomic scale. The development of the STM has catalyzed the evolution of other similar techniques, and currently there is a battery of techniques, i.e., *scanning probe microscopies* that are used in the analysis of a variety of materials properties.

7.2. ELECTRICAL CHARACTERIZATION

Electrical measurements on semiconductors and semiconductor devices are routinely performed for the analysis of a wide range of semiconductor properties. Important semiconductor material and device properties that can be derived using various electrical measurements are: (i) the electrical resistivity (or conductivity), (ii) the energy gap and the separation of an impurity level from the band edge, (iii) the majority carrier concentration, (iv) the mobility of electrons and holes, (v) lifetime and diffusion length of minority carriers, (vi) surface recombination velocity of carriers, and (vii) deep-impurity levels, as well as device parameters such as, e.g., barrier height, contact resistance, interface state densities, junction depth, and channel length and width.

For details on a wide range of electrical characterization methods, see Orton and Blood (1990), Grasserbauer and Werner (1991), Blood and Orton (1992), and Schroder (1998) in Bibliography Section B3.

7.2.1. Resistivity (Conductivity) and the Hall Effect

Some of the routine methods of electrical characterization of semiconductors are related to the measurements of the electrical resistivity (conductivity), carrier mobility, and carrier type.

One of the most important characteristics of a semiconductor is its electrical resistivity, which can be measured by employing the *four-point probe, spreading resistance,* and *Hall–Van der Pauw* methods. The four-point probe technique is essentially a collinear four-point probe array, in which the two outer points carry constant current I through the layer, whereas across the two inner points a voltage V is monitored (see Fig. 7.2). In practice, in such systems, the probe spacing is of

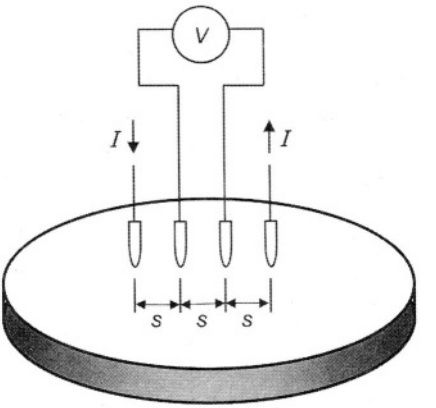

FIGURE 7.2. Schematic illustration of a collinear four-point probe measurement.

the order of 1 mm. For the probe with equidistant spacing s between the points and a semi-infinite sample, the resistivity can be expressed as $\rho = 2\pi s\,(V/I)$ (Ω cm). For finite geometries, this equation must incorporate a geometrical correction factor F, which is related to the sample thickness, the probe spacing, and edge effects. Thus, the resistivity can be expressed as

$$\rho = 2\pi s F (V/I) \qquad (\Omega\text{-cm}) \tag{7.2.1}$$

The correction factors are given in the literature (see, e.g., Schroder, 1998). For wafer (or semiconductor layer) thickness d much lower than s, the resistivity can be expressed as

$$\rho = (\pi/\ln 2) d (V/I) = 4.53\, d (V/I) \qquad (\Omega\text{-cm}) \tag{7.2.2}$$

For very thin semiconductor layers, it is preferable to use a *sheet resistance*, $\rho_s = \rho/d$, measured in units of ohms per square; i.e., ρ_s is the resistance of a given size square of the layer, and from the above equation, it can be expressed as $\rho_s = 4.53\ (V/I)$ in units of ohms per square. Caution has to be exercised in such measurements in thin layers; erroneous resistances may be obtained as a result of the carrier injection occurring, for high current density at the probe tips, or due to probe pressure-induced damage under the probe.

The spreading resistance is employed for measurement of resistivity variations across a sample with a spatial resolution in the range between about 20 and 50 μm, and it is also used for in-depth resistivity derived from measurements on angle-lapped samples. From the Hall–Van der Pauw measurements, one can derive the resistivity, the net-carrier concentration and carrier, mobility; from such measurements at various temperatures, one can also derive the energy levels and concentrations of impurities contributing to the net-carrier concentration.

In semiconductors, both electrons and holes contribute to the current, and thus, the bulk conductivity σ ($\sigma = 1/\rho$, where ρ is the resistivity) can be expressed as (see Chapter 4)

$$\sigma = ne\mu_e + pe\mu_h \tag{7.2.3}$$

In the case of an intrinsic semiconductor (see Section 4.4),

$$\sigma = \sigma_0 \exp(-E_g/2k_B T) \tag{7.2.4}$$

Thus, by ignoring the temperature variation of E_g, and by plotting $\ln \sigma$ as a function of $1/T$ (which yields a straight line), the energy gap E_g can be derived from the slope $-E_g/2k_B$.

In the case of an extrinsic semiconductor, for $n \gg p$, the material is a n-type semiconductor, and for $p \gg n$, the material is a p-type semiconductor. In order to determine the carrier type, in addition to the conductivity (resistivity) measurements, a complementary measurement, such as Hall effect (see below), is required. From the temperature dependence of the conductivity it is possible to determine the impurity levels in semiconductors (see Section 4.5).

Hall effect measurements on semiconductors are routinely employed for determining the density and sign of majority carriers. In this measurement (see Fig. 7.3), the magnetic field B is applied (along the z-direction) perpendicular to the current flow direction (x-direction). The charge carriers, moving through a semiconductor, are deflected due to a Lorentz force exerted by an applied magnetic field. As a result of this deflection, a potential difference is established across the side of the semiconductor that is transverse to the magnetic field and the current direction. The Hall voltage V_y and electric field $\mathcal{E}_y$ appear according to ($V_y = y\mathcal{E}_y$, where y is the sample width)

$$\mathcal{E}_y = R_H J_x B_z \tag{7.2.5}$$

where $J_x = I_x/yz$ is the current density; R_H is the Hall coefficient, which is negative for n-type semiconductors, and it is positive for p-type semiconductors. In other

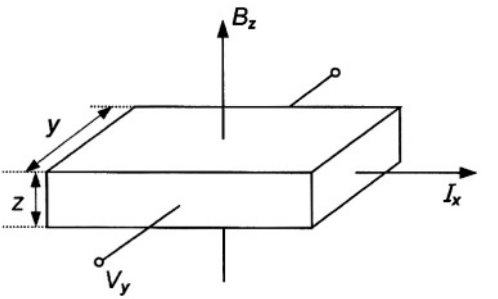

FIGURE 7.3. Schematic illustration of a Hall effect measurement.

words, for *n*-type semiconductors $R_H = -(ne)^{-1}$, whereas for the case of *p-type* semiconductors $R_H = (pe)^{-1}$. Thus, from measured R_H, the density of carriers in the semiconductor can be determined, and from the field polarity, the conductivity type can be ascertained. If the conductivity of a semiconductor is known, the (Hall) mobility can be also determined from $\mu_H = \sigma R_H$. Thus, from combined conductivity (resistivity) and Hall effect measurements, the carrier density, type, and mobility can be determined. From these measurements as a function of temperature, one can also derive the temperature variations of the carrier density and mobility. The doping range sensitivity of this method is between about 10^{14} and 10^{19} cm^{-3}. It should be noted that one should distinguish between the Hall mobility μ_H and drift mobility μ; the ratio of μ_H over μ can typically vary in the range between 1 and 2 depending on the dominant scattering mechanisms, which in turn depend on temperature and doping concentration.

Another method for measuring the carrier mobility is the *Haynes–Shockley experiment,* which provides the mobility of minority carriers. This experiment is based on the production of the photogenerated electron–hole pairs by a pulsed illumination of a small semiconductor region. A semiconductor bar with three contacts is used. Two ohmic contacts are employed for applying an electric field to the semiconductor, whereas the third contact is nonohmic (as a result of forming a rectifying junction in a small semiconductor region located at a given distance from the illuminated region) that is used (in reverse bias) to ensure that only minority carriers are detected. Electrons and holes drift in opposite directions in an applied electric field, and the minority carriers (detected by a reverse–biased contact) are observed using an oscilloscope that allows determination of the delay time that depends on the distance traveled. Thus, the drift velocity and the drift mobility of minority carriers can be determined. The *Haynes–Shockley* experiment can also be used to determine the diffusion constant *D,* as well as the carrier lifetime τ. Other variations of such an experiment (also referred to as the *time-of-flight method*) include an optical injection of minority carriers followed by optical detection (using electron–hole pair recombination emission), as well as a method employing electrical injection of minority carriers followed by electrical detection. The information on the diffusion and recombination of minority carriers is derived from the shape of the pulse moving along the bar due to the drift (i.e., the amplitude of the pulse decreases and it becomes wider due to the minority carrier diffusion and the pulse area is reduced due to the recombination).

The *conductivity type* in semiconductors can also be determined by using the *hot probe* (or *thermoelectric probe*) method. In this case, two probes (one hot and one cold) contact the semiconductor surface. As a result, a temperature gradient is established and the conductivity type is determined by the sign of the generated voltage.

7.2.2. Capacitance–Voltage Measurements

There are different categories of techniques that are based on the capacitance–voltage (*C–V*) measurements on various diode structures (e.g., *p–n* junction, Schottky barrier, and metal–insulator–semiconductor diodes). The capacitance

measurement in these cases can be related to the charge distributions within the semiconductor depletion region. Since the depletion layer width depends on the applied voltage bias, capacitance measurement as a function of the voltage bias (i.e., *C–V* measurement) can be used for determining doping concentration profiles in semiconductors. Two major *C–V* methods are (i) experiments that involve steady-state equilibrium conditions and (ii) experiments involving nonequilibrium conditions and measurements of transient signals. In the former case, it is possible to derive the ionized dopant concentration in a semiconductor, and the barrier height of the junction, as well as interface state densities. In the case of experiments involving transient measurements, the nonequilibrium conditions are realized by employing voltage bias pulses (it is also possible to use optical beam pulses); such methods (e.g., *deep-level transient spectroscopy,* DLTS) can be used to derive information on the states, introduced by various defects and impurities, in the energy gap of the semiconductor.

The *p–n* junction *C–V* measurement is based on a relationship between the doping concentration, the junction capacitance *C*, junction area *A*, and an applied voltage *V*, which can be expressed as (see Section 5.2)

$$C = A\left[\frac{e\varepsilon N_a N_d}{2(N_a + N_d)}\right]^{1/2} \frac{1}{(V_{bi} - V)^{1/2}} \tag{7.2.6}$$

This equation can be rearranged as follows:

$$\frac{1}{C^2} = \left[\frac{2(N_a + N_d)}{e\varepsilon A^2 N_a N_d}\right](V_{bi} - V) \tag{7.2.7}$$

Thus, plotting $1/C^2$ as a function of the bias voltage V results in a straight line (see Fig. 7.4), from which one can obtain the value of V_{bi} from the intercept on the V-axis. From the slope of the straight line, one can determine the value of the expression in the square brackets, and subsequently extract one of the parameters of interest (such as doping concentration) if the others are known. For specific

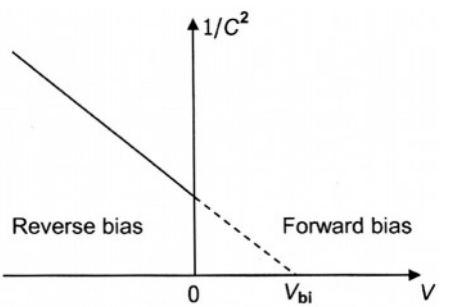

FIGURE 7.4. Typical plot of the dependence of $1/C^2$ on applied voltage.

cases, such as for $N_a \gg N_d$ (i.e., a one-sided abrupt p–n junction), Eq. (7.2.7) can be simplified to

$$\frac{1}{C^2} = \left(\frac{2}{e\varepsilon A^2 N_d}\right)(V_{bi} - V) \tag{7.2.8}$$

Thus, from the slope of a straight line, one can determine N_d, which corresponds to the doping concentration of the lower-doped side of the junction. In principle, the doping concentration of the higher-doped side of the junction (i.e., N_a in this case) can also be derived from the known V_{bi} and calculated doping concentration of the lower-doped side.

The information obtained from DLTS measurements is related to deep level impurities and point defects; it is a nondestructive technique with good accuracy and detection limit. The DLTS method involves repetitive filling and emptying of traps in the semiconductor depletion region and simultaneous measurement (monitoring) of the junction capacitance with a fast meter. Such a process results in capacitance transients due to the release of trapped carriers from deep levels within the depletion region. The DLTS measurements are performed on semiconductor devices, such as p–n junction or Schottky barrier diodes, and metal–oxide–semiconductor (MOS) and metal–insulator–semiconductor (MIS) capacitors. As mentioned above, DLTS measurements are based on the capture and thermal release of carriers at traps (typically, a voltage bias pulse is employed to fill the traps). Following each excitation pulse, the deep levels are in a nonequilibrium state; the thermal emission of captured carriers restores the thermal equilibrium. If the levels are located in the space-charge region of, e.g., a p–n junction or Schottky barrier, the relaxation process will result in a measurable current transient or capacitance transient, with the rate of decay depending on the energy of the deep level and also on temperature. Thus, by monitoring the time constant of the transients as a function of the excitation pulse repetition rate, or the rate window, at different temperatures, the energy level and the concentration of deep levels can be derived. It is important to emphasize that the capacitance transient allows differentiation between electron and hole traps from the sign of the signal, which is independent of the rate window. (Note that in the case of current transients the sign of the signal depends on the rate window and is the same for electrons and holes.)

7.2.3. Photoconductivity

In photoconductivity measurements, irradiation with photons, whose energies are greater than the semiconductor energy gap, produces photogenerated electron–hole pairs that can contribute to the conductivity (this is the case of *intrinsic photoconductivity*). At photon energies lower than the energy gap of the semiconductor, photoconductivity can still be produced by excitation of carriers from impurity levels in the energy gap (in this case, only one kind of carriers is generated, and this is the case of *extrinsic photoconductivity*). For the photoconductivity measurements, an external bias is applied across the electrodes on opposite ends of

the semiconductor sample. When electron–hole pairs are generated, the *dark conductivity* $\sigma_0 = e(n_0\mu_e + p_0\mu_h)$ is increased by $\Delta\sigma = e(\Delta n\mu_e + \Delta p\mu_h)$, where Δn and Δp are densities of photogenerated electrons and holes, respectively. In an intrinsic semiconductor, $\Delta n = \Delta p$, and, hence, $\Delta\sigma = e\Delta n\,(\mu_e + \mu_h)$. An important parameter, which characterizes the photoconductive response of a semiconductor, is the *photogeneration rate G*, defined as the number of carriers collected at the electrodes for each absorbed photon. For simplicity, in the case of the photogeneration of only one kind of carriers and neglecting diffusion, the continuity equation can be expressed as

$$\frac{d\Delta n}{dt} = G - \frac{\Delta n}{\tau_n} \tag{7.2.9}$$

where τ_n is the electron lifetime. For the steady state (i.e., equilibrium), $\Delta n = G\,\tau_n$. In the absence of traps, the time constant associated with the photoconductivity, which increases to the steady state (when the photoexcitation at low intensity is switched on) or decreases to the dark value (when the photoexcitation is switched off), can be determined from the continuity equation (for details, see, e.g., Bube, 1992, in Bibliography Section B2). Thus, the rise curve (i.e., photoconductivity increase) can be expressed as

$$\Delta n(t) = G\tau_n[1 - \exp(-t/\tau_n)] \tag{7.2.10}$$

whereas the decay curve is

$$\Delta n(t) = G\tau_n \exp(-t/\tau_n) \tag{7.2.11}$$

In the case of uniform absorption of photons, $G = \alpha\eta I/h\nu$, where α is the absorption coefficient, η is the quantum efficiency, and I is the intensity of the photon flux.

Two basic methods of photoconductivity measurements that are employed for the characterization of semiconductors are the continuous and pulsed illumination methods. In the pulsed illumination case, when the illumination is switched off, the current decays as a function of time, and from the decay curve one can derive the minority carrier lifetime.

The major factors that may affect the photoconductivity are related to carrier trapping, and the presence of space charges in the material and localized electric fields at the contacts. In such cases, the description of the photoconductivity phenomena is more complex than outlined above. For example, more detailed analysis, including the presence of traps, indicates that traps lead to the increase of both the response time and the sensitivity of a photoconductor. This is due to the fact that as more carriers of one type are trapped, more carriers of the opposite type stay in the conduction or valence band (depending on whether these are electrons or holes, respectively) because of charge neutrality, and thus the conductivity is higher. The relaxation of photoconductivity (i.e., the conductivity drop with illumination switched off) is also affected, and it may now be composed

of two recovery time constants, i.e., fast time constant and longer time constant associated with thermal release of carriers from traps. Further complications may arise due to the presence of different types of traps having distribution of energy levels. In this context, the panchromatic excitation, containing longer-wavelength light, may empty the traps (without generating electron–hole pairs), and this would result in the reduction of both the longer time constant and photosensitivity. (For more details on photoconductivity phenomena in semiconductors, see, e.g., Bube, 1992, in Bibliography Section B2.)

7.3. OPTICAL CHARACTERIZATION METHODS

Optical properties of semiconductors are determined by the interaction of electromagnetic radiation with the material that leads to the formation of various signals as a result of processes such as absorption, reflection, emission, and scattering (see Fig. 7.5). Optical characterization methods employ these processes. A wide variety of possible configurations of both light sources and optical instruments are available for the optical characterization of semiconductors. As mentioned earlier, the main advantages of optical techniques, as compared to those employing charged particle excitation beams, include the ability to analyze specimens in air and the absence of charging in insulating materials. Optical characterization techniques are, in most cases, inherently contactless and nondestructive, and thus they are well suited for *in situ* characterization and process control (such as thin-film deposition processes). The major optical techniques for characterization of semiconductors include optical microscopy,

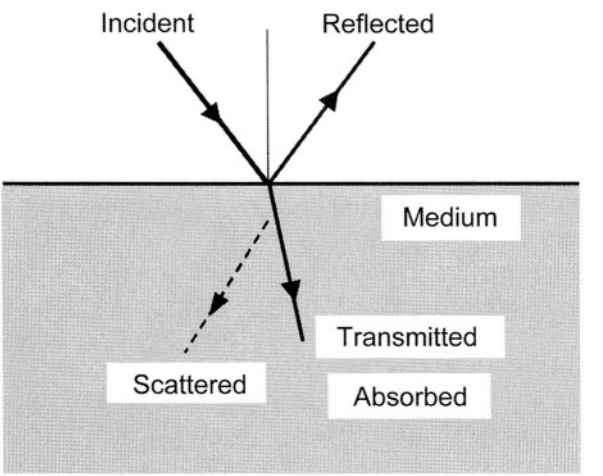

FIGURE 7.5. Schematic representation of various optical processes occurring as a result of the interaction of the incident radiation with a medium. Scattering processes include Brillouin and Raman scattering. Some of the absorbed radiation may also be emitted leading to photoluminescence.

optical absorption, photoluminescence, the Raman scattering, ellipsometry, and modulation techniques.

7.3.1. Optical Absorption

Optical absorption spectroscopy can be performed from the infrared to the ultraviolet ranges. In the near-ultraviolet, visible, and near-infrared ranges, semiconductors (depending on the energy gap) absorb electromagnetic radiation strongly through a mechanism of the generation of electron–hole pairs at photon energies greater than the fundamental energy gap (hence, fundamental absorption edge). Thus, the measurement of the fundamental absorption edge facilitates the determination of the energy gap of a semiconductor. Optical absorption may also occur at various energies (lower than the fundamental energy gap) due to various impurities, defects, and vibrational bonds.

The optical absorption is described by an absorption coefficient α, which can be derived from transmission measurements (see Section 4.7). If I_0 is an incident light intensity, and I is the transmitted light intensity, the transmission, $T = I/I_0$, can be written as (neglecting reflection and interference effects) $I = I_0 \exp(-\alpha d)$.

Infrared absorption spectroscopy can be used to determine molecular bonding and coordination, and various electronic states in the energy gap of a semiconductor. Molecular bonds, described as spring oscillators, have resonant frequencies related to stretching or bending vibrational modes. Typical values of the resonant frequencies are of the order of 10^{14} s^{-1}, which corresponds to a wavelength λ of about 3–10 μm, i.e., in the mid-infrared. In infrared spectroscopy, a more convenient quantity used in the analysis is a wavenumber, which is the reciprocal of the wavelength, expressed in units of cm^{-1}. The characteristic frequencies of various bonds can be identified from the tabulated data given in the infrared-spectroscopy literature (databases). Typically, the spectral ranges of infrared spectrometers are between about 200 and 4000 cm^{-1} (corresponding to wavelengths between about 50 and 2.5 μm). Some examples of the infrared absorption spectroscopy employed in the analysis of semiconductors include detection of interstitial oxygen (at 1105 cm^{-1}) and substitutional carbon (at 607 cm^{-1}) in Si (at room temperature); in these cases, the impurity contents in the range between about 5×10^{15} and 2×10^{18} cm^{-3} can be derived using infrared absorption spectroscopy.

Another IR absorption mechanism, as mentioned above, is that related to the electronic bands of shallow donors and acceptors. In this case, for sufficient spectral resolution, the measurements (e.g., in silicon) have to be performed at cryogenic temperatures ($T \leq 30$ K), and thus, the main phosphorus and boron bands at 316 and 320 cm^{-1}, respectively, can be observed.

In general, the infrared spectrophotometers are employed for the determination of variations in the beam intensity of infrared radiation (as a function of wavelength or frequency) due to its interaction with the sample of interest. In this case, the spectrophotometer is employed for dispersing the light from a source producing a broad range of infrared wavelengths, and for obtaining the infrared

spectrum (i.e., the dependence of the ratio of the intensity of light before and after its interaction with the sample as a function of frequency). In such infrared spectrophotometers, the light is dispersed spatially into spectral components. The high sensitivity in the infrared absorption measurements is achieved by using the *Fourier transform infrared spectroscopy* (FTIR) technique, which typically employs the Michelson interferometer. The main components of an FTIR system include (i) the source producing light over a broad range of infrared wavelengths, (ii) the interferometer, and (iii) the infrared detector. In the interferometer the light passes through a beamsplitter that directs one beam to a stationary mirror and back to the beamsplitter, whereas the other beam is directed to a moving mirror that causes the variable total path length in comparison with that related to the beam directed to a stationary mirror. Upon recombination of the two beams at the beamsplitter, the difference in path lengths produces constructive and destructive interference pattern, or an interferogram. Subsequently, the recombined beam passes through the sample, which absorbs the light at characteristic wavelengths that are subtracted from the interferogram, which contains the essential information on intensities and frequencies related to the spectrum, but that information is not readily inferred. In this case, although the frequencies of light (emitted by the interferometer) follow the same optical path, their emission is time-dependent, i.e., such a system is temporally dispersive. The interferometer output recorded in terms of intensity vs. time can be converted into intensity vs. frequency spectrum by employing a mathematical function referred to as Fourier transform. The main advantages of this technique include an improved signal-to-noise ratio due to the fact that all of the source energy impinges on the sample, and the ability to perform mathematical refinement of the data using the computer programs.

7.3.2. Photoluminescence

Photoluminescence (PL) spectroscopy is a very sensitive (contactless and nondestructive) technique for the characterization of various properties of semiconductors, and especially for the analysis of various dopant and impurity levels present in the energy gap of a semiconductor. The details of recombination and luminescence processes were outlined in Chapter 4.

Photoluminescence experimental arrangement (see Fig. 7.6) basically requires the following:

(a) a source of optical excitation (usually a laser with appropriate emission characteristics; if tunability is required, a dye or Ti:sapphire laser is used);
(b) a spectrometer (preferably double monochromator for high spectral resolution);
(c) an appropriate detector (a photomultiplier or a solid-state detector with a suitable wavelength response characteristics; photomultiplier tubes offer good sensitivity in the visible range, whereas Si, Ge, or other photodiodes are used in the near infrared range; in addition, photodiode arrays can be employed to reduce measurement times).

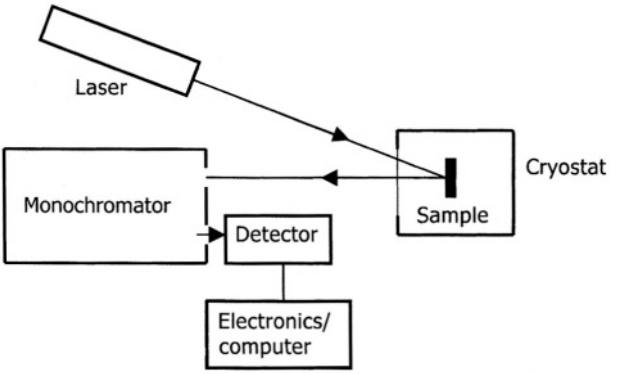

FIGURE 7.6. Schematic diagram of a typical PL measurement setup.

It should be mentioned that the optical systems required for absorption, reflectance, and Raman scattering measurements are similar to this optical system. (Note that in the presence of substantial scattered light, a double or triple monochromator may be required.) Usually, in order to distinguish various transitions and obtain more detailed spectra, measurements at cryogenic temperatures ($\leq$ 77 K) are required, and thus the photoluminescence setup usually incorporates a cryostat. This is important, since (i) at cryogenic temperatures, dopant centers are no longer ionized (note that the thermal ionization of these centers prevents their detection, which can only be deduced from carrier mobility scattering effects) and (ii) the lattice-vibration-induced spectral broadening of luminescence bands is significantly reduced. It should be emphasized that in order to eliminate any strain-induced effects that influence luminescence spectra, it is essential to employ appropriate sample mounting methods (such as V-groove or free suspension). With appropriate optical arrangement, the excitation probe can be focused to about 1 μm, and thus spatially-resolved information can be obtained, as well as PL mapping.

The different PL configurations employed in the analysis of semiconductors include (i) PL spectral analysis, (ii) time-resolved PL analysis, and (iii) PL excitation spectroscopy. In the PL spectral analysis mode, the wavelength of the incident radiation is fixed, and the PL intensity over a specific range of the emission wavelengths is monitored. On the other hand, in PL excitation spectroscopy, the wavelength of the incident radiation is varied, and the wavelength of the analyzing spectrometer is fixed. Time-resolved PL involves measuring the PL intensity (at a given wavelength) as a function of time delay following the excitation pulse; such a measurement allows deriving the carrier lifetime in semiconductors. These various PL measurements can be related to electronic transitions in the sample, and thus they provide information on the electronic band structure (e.g., the energy gap), impurity levels, carrier lifetimes, recombination mechanisms, the quality of semiconductor layers and interfaces, disorder present in the material, and characteristics of quantum wells. One of the important applications of PL is identifying various impurities and defect centers in semiconductors. PL can also be

employed for providing the alloy composition of compounds such as $Al_xGa_{1-x}As$. (In such cases, a calibration curve of the energy gap E_g as a function of composition is required.) Other major applications of PL (i.e., its emission spectra and efficiency) are in (i) the analysis of dopant levels in various semiconductors, (ii) monitoring surface damage and passivation, and (iii) examining various properties of low-dimensional structures (e.g., determining the width of quantum well structures). The PL spectral features that are employed for deriving useful information are the PL peak energy, peak width, and peak intensity. From the peak energy, one can derive information on (i) energy gap and electronic levels, (ii) alloy composition, (iii) binding energies of impurities (and excitons), (iv) internal strain, and (v) quantum well width. The peak width can be used to determine (i) doping concentration, (ii) interface roughness in quantum wells, and (iii) structural quality. The peak intensity can be used in the analysis of (i) doping or defect concentration and their relative quantities, (ii) surface damage and passivation, and (iii) radiative efficiency. As mentioned above, the excitation probe can be focused to about 1 μm, allowing one to obtain spatially-resolved information by measuring the spatial variations in PL peak energy, intensity, or peak width.

One of the major strengths of PL is its sensitivity, down to parts per billion (ppb) level, depending on a specific impurity and host material. However, the quantitative determination of impurity concentration is difficult; thus, PL is mostly a qualitative characterization technique. The reason for this, as discussed in Section 4.8, is related to the fact that the rate of luminescence emission is proportional to the radiative recombination efficiency defined as $\eta = \tau/\tau_r = [1 + (\tau_r/\tau_{nr})]^{-1}$, where τ_r and τ_{nr} are the radiative and nonradiative recombination lifetimes, respectively. In general, η depends on various parameters, such as temperature, the particular dopants and their concentrations, and the presence of various defects. Thus, in the observed luminescence intensity one cannot distinguish between radiative and nonradiative processes in a quantitative manner. The information in the peak of the edge emission band (i.e., corresponding to intrinsic luminescence at ambient temperatures) can be related to the energy gap; however, the information in the broad extrinsic luminescence bands (i.e., those which arise from transitions that start and/or finish on localized states of impurities in the energy gap) observed at room temperature is difficult to interpret in the absence of any generally applicable theory for the wide variety of possible types of luminescence centers and radiative recombination mechanisms. Experimentally, thermal broadening can be minimized by using cryogenic temperatures (e.g., liquid helium temperatures) at which luminescence spectra generally consist of a series of sharp lines corresponding to transitions between well-defined energy levels that can be associated with excitons, phonon replicas, and donor-acceptor pairs. Nevertheless, the quantitative information can be derived on the basis of correlations between dopant concentrations and spectral features such as peak energy and peak width, and relating the measurements to the appropriate standards.

The main advantages of PL spectroscopy include (i) very high sensitivity, (ii) the fact that it is contactless and nondestructive method, and (iii) relative simplicity of measurement. The main disadvantages are (i) the difficulties involved with the

interpretation of accompanying complex competitive radiative and nonradiative processes, surface recombination, and the presence of often unknown concentrations of various defects, (ii) the lack of appropriate standards, and (iii) for sufficient spectral resolution, i.e., to resolve the emission lines, measurements at cryogenic temperatures (preferably at liquid helium temperatures) are required.

7.3.3. Raman Spectroscopy

The energy of scattered light, in general, corresponds to the same value as the incident light. In some cases, however, the energy of scattered light is at a different energy due to the inelastic scattering of light (resulting from the molecule changing its motions), called the *Raman scattering.* In this process, only a small portion (about 10^{-5} % of light) is scattered at energies (usually) lower than the energy of the incident photons. This change in energy is associated with a change in vibrational, rotational or electronic energy of a molecule. In other words, the difference in energy between the incident photon and the scattered photon, i.e., the *Raman shift*, is equal to the energy of a vibration of the scattering molecule. It should be noted that, in general, photons could interact with both the optical and the acoustical phonons (see Section 2.5). The former case corresponds to Raman scattering, whereas the interaction with the acoustical phonons corresponds to *Brillouin scattering.*

Raman spectroscopy is based on measuring the energy shift of the incident photon beam that is inelastically scattered off the material. As mentioned above, the intensity of light corresponding to Raman scattering process is weak, and thus, in Raman spectroscopy measurements, the specimen is illuminated with intense monochromatic light generated by a laser. The energy shift during such a scattering process is due to either the photon energy transfer to the lattice (i.e., phonon emission), or the absorption of a phonon by the photon. In the case of phonon emission, the reduction in photon energy is called the *Stokes shift,* and in the case of scattered photon emerging at a higher energy, it is called the *anti-Stokes shift.* The latter, i.e., the emission of a more energetic photon, occurs when there is a substantial density of phonons present. In this process, the energy and momentum conservation laws must be satisfied, i.e.,

$$\hbar\omega_{\text{scattered}} = \hbar\omega_{\text{incident}} \pm \hbar\omega_{\text{phonon}} \tag{7.3.1}$$

$$k_{\text{scattered}} = k_{\text{incident}} \pm k_{\text{phonon}} \tag{7.3.2}$$

In this process, the vibrational energy eventually dissipates as heat. However, due to low intensity of Raman scattering, the heat dissipation does not result in measurable temperature increase in a sample. In Raman measurements at room temperature, the thermal population of vibrational excited states is relatively low, and typically, the scattered photon has lower energy than the incident photon. This (Stokes) shift to lower energy is the typically observed effect in Raman spectroscopy. Nevertheless, due to a small fraction of the molecules in

vibrationally excited states, Raman scattering in this case results in the scattered photon having higher energy (i.e., the anti-Stokes shift). Note that the intensity in this case is weaker as compared with the Stokes-shifted spectrum. Based on these arguments, the intensity ratio of anti-Stokes to Stokes spectra can be used as a measure of sample temperature.

It should be emphasized that Raman scattering is a powerful means for elucidating the phonon spectra of molecular systems. The Raman spectral peaks, which correspond to the frequencies of the vibrational modes, can, in principle, be attributed to a particular molecular group or phase in the material studied. Such spectra typically provide a means for the elucidation of the structure and bonding from the examination of the position, relative intensity and symmetry of individual peaks.

The coupling of an optical microscope with the Raman system also allows obtaining spatially-resolved Raman measurements, i.e., *micro Raman spectroscopy,* with a resolution of about 1 μm. In this case, the illumination of the sample by a laser is through an optical microscope coupled with the monochromator.

One of the major applications of the technique includes analysis of the crystal structure and composition. One can distinguish between crystalline and amorphous semiconductors from the variations in the shift, the width, and the symmetry of the Stokes line. Raman spectroscopy can provide important information on such issues as defects, induced damage, stresses, and semiconductor processing. This technique can also be used in analyzing temperature-induced effects in semiconductors by employing the time-resolved Raman spectroscopy measurements at high temperatures. Some of the main advantages include the fact that this is a nondestructive technique that requires no vacuum.

7.3.4. Ellipsometry

Ellipsometry is a powerful contactless and nondestructive method for the analysis of semiconductors. In this method, the change in the state of polarization of the electromagnetic wave upon reflection (or transmission) at an interface between two dielectric media is measured. Such a measurement of change in the polarization involves evaluating the amplitude ratio and phase variation between polarization components. The advantages of ellipsometry are (i) its high sensitivity to the presence of very thin layers (down to an atomic monolayer) and (ii) the fact that it depends on measurements of angles, i.e., parameters that are independent of light intensity, detector sensitivity, and total reflectance. The change in the state of polarization (upon reflection) can be represented as the ratio of the two complex reflection coefficients for an electromagnetic wave polarized parallel and perpendicular to the plane of incidence (i.e., $\rho = R_p/R_s$), and it can be expressed as $\rho = \tan(\psi)\exp(j\Delta)$, where ψ and Δ are the measured ellipsometric angles, which determine the refractive index and absorption coefficient of the reflecting medium.

One of the most common ellipsometric arrangements is that incorporating a Polarizer, Compensator, Specimen, and Analyzer (i.e., PCSA).

Some important applications include determining the optical constants and the thickness of thin films, monitoring the changes due to ion implantation-

induced damage in semiconductors, real-time monitoring of thin film growth, surface roughness evaluation, and studies of oxidation of semiconductors. It should be noted that the thickness of thin films and their optical parameters are not directly measured, but are derived from the measurements.

7.3.5. Optical Modulation Techniques

Optical modulation spectroscopy provides a powerful method for both *in situ* and *ex situ* analysis of important semiconductor properties. The basic principle of modulation spectroscopy is measuring the optical spectral response (i.e., optical reflectance or transmittance) of a semiconductor. The optical response is modified by applying a repetitive perturbation, which results in sharp spectral features that are analogous to taking the derivative of the spectrum. This is accomplished by applying a repetitive perturbation such as an electric field (hence, electromodulation), or stress (piezomodulation), or heat (thermomodulation). In addition to the above external modulation, there is also the internal modulation method in which the wavelength is modulated or optical spectral response of the sample is compared to a reference sample. Such modulation methods in essence accentuate specific electronic transitions from largely featureless spectra. Thus, modulation spectroscopy allows one to measure the energies of the interband transitions with high accuracy at room temperature. These energy positions are sensitive to parameters such as composition, strain, temperature, and electric field. In addition, the line widths associated with the electronic transitions depend on various parameters, such as crystalline quality of the material and dopant (impurity) concentration. Thus, this technique can be used to evaluate the effect of various growth and processing procedures on crystal quality. Typically, $\Delta R/R$ is measured, and the most useful external modulation is electromodulation, since it provides the sharpest features. The most commonly employed internal modulation technique is differential reflection spectroscopy, which involves the comparison of the sample reflectance to that of a standard material. The fitting of the sharp features derived in such measurements can be related to the properties (e.g., the energy gap) of the semiconductor. The electromodulation spectra derived for high built-in electric fields may also exhibit oscillations (called Franz–Keldysh oscillations) above the energy gap. The analysis of such oscillations yields direct measurement of the built-in electric field.

The main advantages of modulation spectroscopy techniques are (i) the operation at room temperature and (ii) their potential usefulness for the *in situ* monitoring and control of thin-film growth and processing.

7.4. MICROSCOPY TECHNIQUES

Microscopy techniques are indispensable tools in the characterization of semiconductors and semiconductor devices, since increasingly complex and miniaturized materials and devices are being developed. In semiconductor technology, it is essential to control the undesirable defect densities present in

the material. The main objectives in this case are (i) to detect defects and measure their densities, (ii) to identify them, and (iii) to establish their origin. This is the reason why various types of microscopy are so crucial in the understanding of semiconductors. The microscopy techniques that are commonly employed in semiconductor characterization include various forms of optical microscopy, electron microscopy, and scanning probe microscopy (SPM).

Scanning instruments provide a powerful means of microcharacterization. The general concept of scanning microscopy is based on the serial formation of an image point by point. This can be realized either by electronically scanning a beam (e.g., the electron beam in a scanning electron microscope), or mechanically scanning a probe (e.g., the tip in a scanning tunneling microscope, STM), or mechanically scanning the specimen and keeping the probe stationary. It should be emphasized that in the development of various characterization methods, each technique did not replace the others, but complemented them.

Important factors in considering a particular analysis technique include the ambient (vacuum) required, the sample preparation and size, the magnification range, irradiation-induced effects (damage), the ease of operation and throughput, and quantifiability.

7.4.1. Optical Microscopy

One of the most versatile and indispensable tools in this category is optical microscopy. The main advantages of the optical microscopic methods, in comparison with the electron-beam techniques, are (i) the versatility of optical examination, (ii) relative simplicity of sample preparation, (iii) no vacuum requirement, and (iv) the absence of specimen charging. These have been further enhanced by recent advances in the developments of (*confocal*) *scanning optical microscopy* and of computer-based image analysis systems. The main features of a scanning optical microscope include (i) laser illumination, (ii) the detection of the transmitted, or reflected, light by a photodetector, and (iii) the display of the video image on a cathode-ray tube (CRT). In the confocal configuration, both the resolution and the depth of field are improved. In this case, only one point (column) of the sample is illuminated at a time and the light is collected from just one (confocal) object point in that column, and therefore the contrast and depth resolution are improved. In a scanning optical microscope, either the beam or the sample is scanned, and a raster image can be displayed on a CRT and stored in a computer for image processing. One of the important applications of the scanning optical microscope is in the microcharacterization of defects in semiconductors and of electronic devices using the *optical beam-induced current* (OBIC) and scanning photoluminescence. These techniques are analogous to *electron beam-induced current* (EBIC) and *cathodoluminescence* (CL) modes available in the SEM (see Section 7.4.2). The main advantages of these scanning optical microscopic methods, in comparison with the analogous electron-beam techniques, are no vacuum requirement and the absence of specimen charging.

7.4.2. Electron Beam Techniques

7.4.2.1 Scanning Electron Microscopy The presence of several different modes, providing complementary information on physical, compositional, and structural properties of solid-state materials and devices, constitutes one of the main advantages of this technique. In addition, SEM is capable of accommodating (i.e., examining) macroscopic specimens (e.g., semiconductor wafers and devices) with no special sample preparation steps, in general.

The electron-optical column of the SEM includes the electron gun, the electromagnetic lenses, and the specimen chamber (see Fig. 7.7). A finely focused electron beam, produced by the electron-optical column, is scanned in a raster fashion over the specimen surface. (Scanning coils are used to deflect the electron

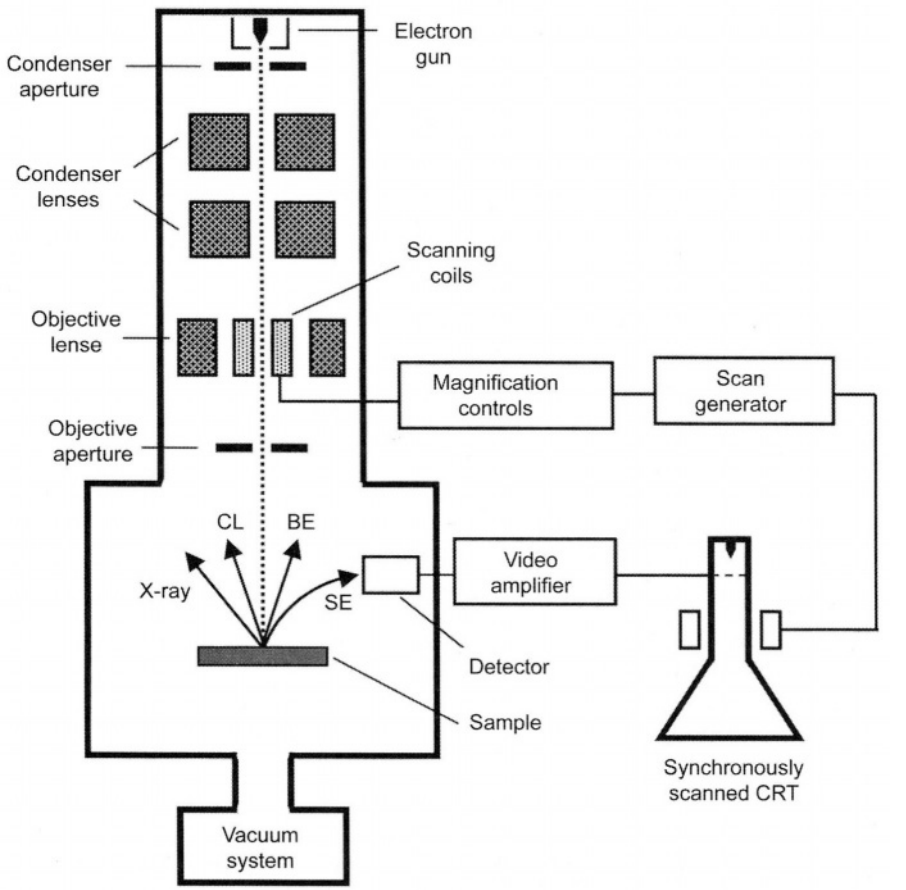

FIGURE 7.7. Schematic diagram showing the basic components of a SEM. SE and BE denote secondary electrons and backscattered electrons, respectively, and CL denotes cathodoluminescence.

beam in such a manner that its spot scans line by line a square raster over the sample surface.) The microscopic images are formed using various signals that are produced as a result of the dissipation of the electron-beam energy in the material into other forms of energy. A particular signal is detected and converted into an electrical signal that is amplified and fed to the grid of the synchronously scanned display CRT. The raster scan of the electron-beam spot over the specimen surface results in a one-to-one correspondence between picture points on the display CRT screen and points on the specimen. The amplified signal modulates the brightness of the CRT, and as a result, the variations in the strength of the particular signal being detected cause variations in brightness on the CRT screen, i.e., contrast on the micrograph and an image of the specimen. The SEM image magnification, which can be continuously varied between about 10× and 500,000×, is the ratio of the size of the square area scanned on the display CRT screen to the size of the area scanned on the specimen. In addition to its high spatial resolution for secondary electron imaging, the SEM also has a relatively large depth of field, as compared to an optical microscope. (The depth of field is the thickness on the specimen surface for which the magnified image is in focus.) Such a relatively large depth of field results in a three-dimensional appearance of the SEM images.

Interaction of incident electrons with solids, electron-beam energy dissipation, and the generation of electron–hole pairs in the semiconductor are important for understanding the effects produced in electron-irradiated materials and in interpreting SEM observations. Two types of scattering mechanisms (i.e., elastic and inelastic) have to be considered. The elastic scattering of the incident electrons by the nuclei of the atoms gives rise to high-energy backscattered electrons and, in the SEM, to atomic number contrast and channeling effects in the *emissive mode.* Inelastic interaction processes result in a variety of signals, such as the emission of secondary electrons, Auger electrons, characteristic X-rays, the generation of electron–hole pairs, CL, and thermal effects. As a result of the interaction between the incident electrons and the solid, the primary electrons undergo a successive series of elastic and inelastic scattering events in the material, and the original trajectories of the incident electrons are randomized. The range, R_e, of electron penetration depends on the electron-beam energy E_b as $R_e = (k/\rho)E_b^{\alpha}$, where k and α depend on the atomic number of the material and on E_b, and ρ is the density of the material. From this one can estimate the *generation* (or *excitation) volume* in the material. The *generation factor* (i.e., the number of electron–hole pairs generated per incident beam electron) is given by $G = E_b(1 - \gamma)/E_i$, where E_i is the ionization energy (i.e., the energy required for the formation of an electron–hole pair) and γ represents the fractional electron-beam energy loss due to the backscattered electrons. The ionization energy E_i is related to the energy gap of the material (for practical purposes, if unknown, it is reasonable to assume that $E_i \approx 3E_g$). For example, for Si, $E_i = 3.63$; thus, one incident 30 keV electron can generate in Si several thousand electron–hole pairs in the excitation volume that is several microns in diameter (for 30 keV electrons). Note that in bulk samples, various signals actually originate from different effective depths in the material. Secondary electrons originate from material within about 100 Å of the specimen surface, whereas backscattered electrons are emitted from within about the upper

one-half of the excitation volume, and the signals in CL and *charge-collection* (CC) modes originate from within the whole excitation volume.

Among the various SEM modes, the most routinely used in practice is *secondary electron imaging* (SEI) for characterizing topographic features of solid surfaces. In addition, the SEI mode can be used to analyze electric and magnetic fields that are present in the material, and that affect the energy and direction of emission of these low-energy electrons. Since the emission of the secondary electrons is dependent on the electrical potential present at the electron-beam impact point, it is possible to image and measure variations in the electrostatic potential in different regions of a sample. Thus, this so-called *voltage contrast* can be employed in the characterization of semiconductor devices. Voltage contrast with high-frequency stroboscopy is especially useful in the analysis of integrated circuits. In such measurements, a special holder is employed for external control of the integrated circuit operation. Using such a stroboscopic voltage contrast in the SEM, an integrated circuit device can be monitored operating at its normal frequency, which allows locating the faulty regions in the working device. *Backscattered electrons* provide atomic number contrast, and thus can provide qualitative information on compositional uniformity. Under appropriate operating conditions, backscattered electrons can also provide crystallographic information due to electron channeling, which is based on the variation of the backscattered electron yield as the angle of incidence of the scanned electron beam passes through the Bragg angle to crystal lattice planes. In this case, SEM pictures consisting of series of bands and fine lines and called *electron channeling patterns* (ECPs) are produced, and these depend on the crystal structure and orientation of the sample. *Characteristic X-rays*, emitted due to electronic transitions between inner-core levels, can be used in the analysis of the particular chemical element present, and thus of the materials composition.

The CC and CL modes provide microcharacterization of electronic properties of semiconductors and semiconductor devices. CL offers a contactless and nondestructive characterization tool in the microanalysis of luminescent materials, whereas, in the CC mode, electrical contacts applied to the semiconductor device allow monitoring of electrical signal in the external circuit.

In the *electron acoustic mode*, the chopped electron-beam-induced intermittent heating, and thermal expansion and contraction, result in the propagation of ultrasonic waves that can be detected by piezoelectric transducers; this technique is useful in detection of subsurface defects in solid-state materials.

An additional SEM mode, which complements CL spectroscopy for the assessment of nonradiative centers in the semiconductor, is *scanning deep-level transient spectroscopy* (SDLTS). This mode, which is based on the capture and thermal release of carriers at traps, allows one to determine the energy levels and the spatial distribution of deep states in semiconductors. (For more details on microcharacterization of semiconductors using various SEM modes, see Holt and Joy, 1989, in Bibliography Section B3.)

Among these SEM modes, CC (which is often referred to as EBIC) and CL modes of SEM are especially valuable for the microcharacterization of electronic properties of defects, since these defects often have such a profound effect on both

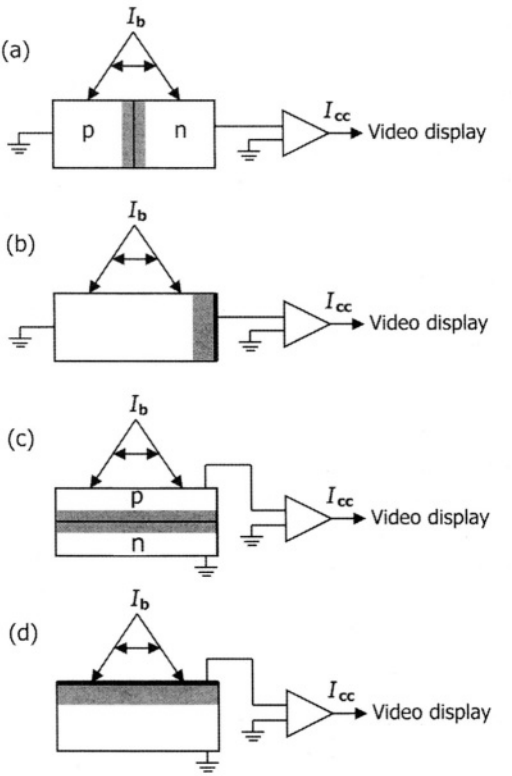

FIGURE 7.8. Schematic illustration of CC geometries; (a) and (b) illustrate perpendicular p–n junction and Schottky barrier geometries, respectively; (c) and (d) show planar p–n junction and Schottky barrier geometries, respectively.

(i) the electronic and optoelectronic properties of semiconductors and (ii) the performance of semiconductor devices.

In the measurement of *electron-beam-induced current* (EBIC), electron–hole pairs generated in the depletion region, or within minority-carrier diffusion range of it, are separated by the built-in electric field and the CC current is monitored in the external circuit (see Fig. 7.8). The EBIC technique is routinely employed in the evaluation of p–n junction and Schottky barrier characteristics and in the analysis of various defects, such as dislocations. The EBIC contrast in an SEM image is due to variations in charge-collection efficiency that may arise from recombination at various defects. In such measurements, regions with a high carrier recombination efficiency will appear darker in comparison with regions with low carrier recombination that will appear brighter, providing a means for direct imaging of electrically active defects in semiconductor devices. Some major applications of the EBIC mode include (i) microcharacterization of the electrical junction characteristics of semiconductor devices, (ii) microcharacterization of the

concentration and distribution of electrically active defects and detecting subsurface defects and damage, (iii) measuring the minority-carrier diffusion length and lifetime, the surface recombination velocity, and the width and depth of depletion zones, (iv) measuring the Schottky barrier height, and (v) failure analysis of semiconductor devices.

Cathodoluminescence (CL) analysis is based on the emission of photons in the ultraviolet, visible and near-infrared ranges due to recombination of electron–hole pairs generated by an incident electron beam. Two types of signal can be obtained to produce micrographs (i.e., CL microscopy) or spectra (i.e., CL spectroscopy). In CL microscopy, luminescent images of regions of interest are displayed on the CRT by using a monochromatic or panchromatic imaging, whereas in CL spectroscopy (with the electron beam fixed at a given area) luminescence spectra from selected areas of the sample are obtained. The CL signal is generated by detecting photons that are emitted as a result of electronic transitions between the conduction band, and/or levels (due to impurities and defects) located in the energy gap, and the valence band. These electronic transitions are affected by many factors, and thus, no universal law can be applied to interpret CL spectra. The description of the formation of the CL signal involves the analysis of the generation, diffusion, and recombination of minority carriers (see Section 4.8). The CL intensity L_{CL} can be derived from the overall recombination rate $\Delta n(r)/\tau$ by noting that only a fraction $\Delta n(r)\eta/\tau$ recombines radiatively and assuming a linear dependence of the CL intensity L_{CL} on Δn. By deriving the excess carrier density $\Delta n(r)$ from the solution of the differential equation of continuity for the diffusion of the excess minority carriers, and assuming the simplified case of a point source or a sphere of uniform generation, one can obtain $L_{CL} = f\eta G I_b/e$, where f is a function containing correction parameters of the CL detection system and factors that account for optical absorption and internal reflection losses, I_b is the electron beam current and e is the electronic charge. However, since the rate of CL emission is proportional to $\eta = \tau/\tau_r = [1 + (\tau_r/\tau_{nr})]^{-1}$, in the observed CL intensity one cannot distinguish between radiative and nonradiative processes in a quantitative manner. (In general, η depends on temperature, the particular dopants and their concentrations, and the presence of various defects.)

As mentioned in Chapter 4, the information in broad luminescence bands observed above liquid nitrogen temperatures is relatively difficult to interpret. At low temperatures (e.g., liquid helium temperatures), the thermal broadening effects are minimized, and luminescence spectra in general become much sharper and more intense, allowing a more unambiguous identification of luminescence centers. The near-energy-gap emission (i.e., edge emission) at liquid helium temperatures is often resolved into emission lines, which can be due to excitons, free-carrier to donor (or acceptor) transitions and their phonon replicas, and/or donor–acceptor pair (DAP) lines. Thus, it should be emphasized that, for detailed studies of impurities and defects in semiconductors, it is necessary to employ systems having the capability of sample cooling (preferably to liquid helium temperatures). The CL mode is a powerful tool in the analysis of semiconductors due to the following features: (i) CL is a contactless method that provides analysis of electronic properties of semiconductors with a spatial resolution of less than 1 μm, (ii) the

detection limit for impurity concentrations can be as low as 10^{14} cm^{-3}, and (iii) CL is especially useful in the analysis of electronic properties of wide energy-gap materials (for which optical excitation sources are not readily available), and also in the microcharacterization of optoelectronic materials and devices in which cases it is the luminescence properties that are of practical importance. Some applications of CL microcharacterization technique include (i) deriving information on the electronic band structure related to the fundamental energy gap, (ii) uniformity analysis of semiconductors, i.e., obtaining distribution of various defects and impurities and recombination studies in the vicinity of defects, (iii) measurements of the minority-carrier diffusion length and lifetime, and of the dopant concentration, (iv) degradation studies of optoelectronic devices, (v) depth-resolved studies of defects in ion-implanted samples and of interface states in heterojunctions, and (vi) microcharacterization of stress distributions in epitaxial layers. Note that the CL depth profiling can be performed by varying the range of incident electron penetration that depends on the electron-beam energy. (The excitation depth can be varied from about 10 nm to several micrometers for electron-beam energies in the range between about 1 and 30 keV, respectively.) The minority carrier lifetime can be obtained from time-resolved CL measurements employing a beam blanking system and a fast detector. (For more details on EBIC and CL modes, see Holt and Joy, 1989, in Bibliography Section B3.)

The mechanisms of signal formation (related to the details of electron beam–solid interactions, electronic excitations, and materials properties) are different for the SEM modes. Consequently, their spatial resolution, quantifiability, and sensitivity vary substantially.

7.4.2.2 Transmission Electron Microscopy Transmission electron microscopy (TEM), which employs transmitted electrons that can be detected and analyzed in thin samples (of the order of 1000 Å and less), provides information related to the crystal structure and defects in solid-state materials. In this technique, a high-energy electron beam (typically between about 100 and 1000 keV) is employed, and the electrons transmitted through the thin sample are focused by electromagnetic lenses. The electron-optical column contains condenser lenses to control the electron illumination of the sample, and objective, intermediate, and projector lenses that produce either the micrograph or the diffraction pattern on the fluorescent screen. The thin sample, mounted in a special holder, is inserted, using a vacuum interlock system, into the TEM column and positioned below the condenser lenses. In the case of a *conventional transmission electron microscope* (CTEM), a photographic camera, positioned under the removable fluorescent screen, allows one to record a magnified image or diffraction pattern of the sample. In the *scanning transmission electron microscope* (STEM), which is equipped with scanning coils, the electron beam spot scans a square raster over the sample surface; a particular signal is detected, amplified and fed to the grid of the synchronously scanned display CRT. Thus, the amplified signal modulates the brightness of the CRT, variations of which result in contrast on the micrograph and an image of the specimen.

In the TEM, the spatial resolution is limited by diffraction and spherical aberration of the objective lens. Under certain approximation, the minimum

resolvable distance Δx is related to the spherical aberration constant C_s and the electron wavelength λ as

$$\Delta x = 0.6\, C_s^{1/4}\, \lambda^{3/4} \qquad (\text{Å}) \tag{7.4.1}$$

The de Broglie wavelength associated with electrons can be written as $\lambda_e = h/(2meV)^{1/2}$, where h is Planck's constant, m and e are the mass and charge of the electron, respectively, and V is the accelerating voltage. For a typical operating voltage of 100 keV in TEM, the incident electrons have a wavelength of 0.037 Å (which is several orders of magnitude smaller than the wavelength of visible light employed in optical microscopes). Thus, compared to optical microscopy, much smaller details can be resolved in the TEM; a spatial resolution attainable according to Eq. (7.4.1) for Δx is 2 Å, allowing atomic resolution imaging. The images observed in the TEM are due to the local brightness variations (i.e., the contrast). The contrast mechanisms leading to image formation include diffraction contrast, mass-thickness contrast, and phase contrast. In the diffraction contrast, the structural defects, such as dislocations, can be imaged by using either the bright-field mode or dark-field mode. In the bright-field mode, a small aperture inserted in the back focal plane of the objective lens blocks the diffracted beam, and thus the image is formed by the direct beam only. For example, a strong local diffraction may occur on the distorted lattice planes near a dislocation, and thus these diffracted electrons are excluded from the image. Thus, in the bright-field image, the features (e.g., dislocations) in the image are observed as regions of dark contrast relative to the bright background. When the incident electron beam is tilted to allow the diffracted electrons to travel along the optic axis, the dark-field image is obtained (in this case, the image is formed by the diffracted electrons), and in this case the diffracting features in the image are observed as regions of bright contrast relative to the dark background. Another TEM imaging technique, based on phase contrast, is *high-resolution transmission electron microscopy* (HRTEM), in which the transmitted and diffracted beams, passed through the objective aperture, are recombined to form an image. The phase interference between these beams results in periodic intensity fringes. Contrast in the image is due to these periodic fringes, which are related to the Bragg diffracting planes; such images provide the capability of the *lattice imaging.*

The crystalline structure can be ascertained by employing *transmission electron diffraction* (TED) pattern, which can be formed by inserting the *selected area diffraction* (SAD) aperture in the image plane of the objective lens. Diffraction patterns can also be obtained with a high spatial resolution by means of the convergent-beam diffraction technique. In this case, the size of the electron-beam probe, incident on the sample, determines the region for analysis. The electron diffraction pattern is very useful in the analysis of crystallographic structures of materials, including the determination of lattice spacing. In single crystals, the electron diffraction pattern consists of sharp spots. In polycrystalline specimens, the diffraction pattern typically consists of a series of sharp concentric rings, which are essentially due to numerous single crystal diffraction patterns from the randomly-oriented grains. (Note that in large-grain polycrystalline specimens, the pattern consists of superimposed spots and rings.) In amorphous materials,

the absence of long-range order results in a breakdown of the Bragg scattering condition, and the electron diffraction pattern consists of a series of diffuse concentric rings. In crystalline materials, the basic information that can be obtained from the electron diffraction patterns are the d_{hkl} -spacing, the crystal lattice type and lattice parameters.

In the STEM, it is possible to incorporate various detectors, in analogy with the SEM, for employing various modes, such as secondary electrons, backscattered electrons, X-rays, and CL. In addition, for electron energies greater than about 100 keV, energy losses of the transmitted electrons are characteristic of the elements present in the material; this is a basis for *electron energy-loss spectroscopy* (EELS) that can provide chemical (compositional) and structural information. These different modes make STEM a truly analytical tool for the examination of structural and physical properties; together with high-resolution imaging of various defects, this allows making important direct correlations between different materials properties. The advantage of such an analytical system is that, due to the negligible lateral beam spreading in the thin sample, the spatial resolution can be greatly improved compared to bulk SEM modes. This is, however, limited by the inverse relation of spatial-to-signal resolution. The major disadvantages in this case include (i) the low signal levels compared to the bulk SEM analysis, and (ii) the fact that, since the volumes analyzed are relatively small, the results may be highly atypical, unless these volumes are carefully selected or the observations are repeated in different regions.

In TEM, it is also possible (with appropriate attachments and sample holders) to study *in situ* processes and structural changes due to annealing. If the heating sample holder is equipped with additional external probes, such as electrical contacts, it is also possible to correlate *in situ* between the annealing steps and the electrical behavior of a semiconductor or a semiconductor device.

The need for thin samples in TEM observations requires a tedious sample preparation procedure, usually by chemical etching or by low-energy (of the order of 10 keV) ion milling. Typically, two sample configurations are employed. These are plan-view samples (i.e., parallel to the sample surface) and cross-sectional samples (i.e., perpendicular to the sample surface). If a sample is prepared for cross-sectional TEM (i.e., XTEM) observations, it is possible to observe distributions of defects as functions of depth (e.g., the propagation of dislocations in epitaxial heterostructures), and it is also possible to determine the sharpness of the interfaces in multilayer structures.

7.4.3. Scanning Probe Microscopy

Unlike the techniques that employ effects produced by the irradiation of the material with excitation probes such as photon beam, electron beam, or ion beam, a different approach is used in the SPM. The development of this class of techniques (or microscopy techniques without lenses) was catalyzed in the last decade by the discovery of the *scanning tunneling microscope* (STM), which is capable of imaging surfaces on the atomic scale. The basic components of SPM are (i) a fine probe (or tip) which is mechanically scanned in very close proximity to

the specimen and (ii) a feedback circuit which controls the distance between the tip and the sample surface. A specific signal, which is generated as a result of an interaction between a fine probe and a specimen, is used to produce an image. The resolution in this case is determined by the tip diameter and the tip-to-sample separation.

The basic principle of the operation of the STM is the measurement of the quantum-mechanical electron tunneling current between an ultrasharp tip and the sample (see Fig. 7.9). The tip is made of conductive material (e.g., tungsten) and it can be moved in three dimensions by employing piezoelectric elements for x, y, and z translators. The tip is positioned in close proximity of about 10 Å to the sample surface, so that at a low operating voltage of the order of millivolts, the tunneling current of about 1 nA is detected. The tunneling current depends exponentially on the distance between the tip and the sample surface and it is very sensitive to that distance. Two operating modes can be used in the measurements. These are *constant current* and *constant height* methods. In the constant current mode, by using a feedback circuit, which is employed to change the tip height z (or tip separation from the sample surface) by applying the voltage to the z-controlling piezoelectric element, the tunneling current is kept constant at each point. As the tip is scanned across the specimen surface in the x and y directions, the tunneling current is monitored and the voltage that controls the tip height is recorded, and these allow obtaining an image that reveals the surface topography with atomic resolution. In the constant height mode, the tip travels in a horizontal plane above

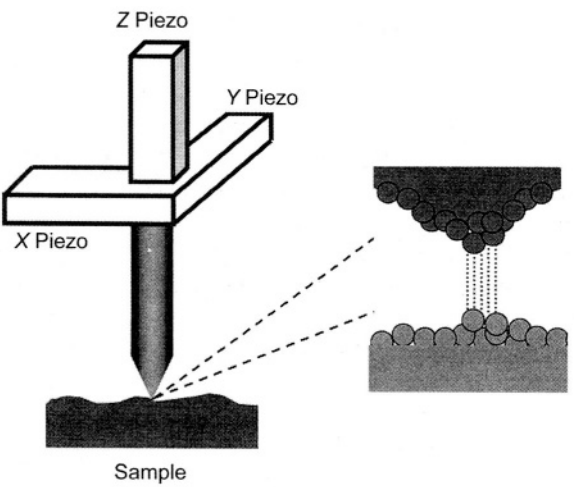

FIGURE 7.9. Schematic diagram of STM tip–sample interaction. The piezoelectric elements control the spacing between the tip and the sample surface (*Z* Piezo), and they also provide a tip movement in a raster pattern (*X* Piezo and *Y* Piezo) across the sample surface as the tunneling current is monitored to produce an image. An enlarged view of the tip–sample region reveals naturally occurring protrusions, which effectively reduce the area related to the tunneling current and thus facilitate the atomic resolution.

the sample, and the tunneling current varies depending on the topography and local surface electronic properties of sample. The constant current mode is more suitable for irregular surfaces, but the measurement requires longer time, whereas the constant height mode is faster (since it does not have to move the scanner up and down), but this mode requires smooth surfaces.

The tunneling current, which depends strongly on barrier width s (i.e. the gap spacing between the tip and the sample surface), and also on the applied voltage V and the effective work function Φ, can be expressed as

$$I = C(V/s)\exp(-As\,\Phi^{1/2}) \tag{7.4.2}$$

where C is a constant, $A \approx 1.025\ (\text{eV})^{-1/2}\ \text{Å}^{-1}$, and Φ is an effective work function [$\Phi = (\Phi_1 + \Phi_2)/2$, where Φ_1 and Φ_2 are the work functions of the tip and sample]. For Φ of a few electron-volts, the tunneling current changes by about an order of magnitude for each Angstrom change in s. The surface electronic structure can be analyzed using *scanning tunneling spectroscopy* (STS), which essentially involves determining peaks in dI/dV (obtained from the tunneling current–voltage spectra), which can be related to the surface electronic density of states. Thus, since the tunneling current depends on local electronic structure, spectroscopic information on an atomic scale can be derived. Also, by studying the dependence of current on distance (barrier width) variations, Φ can be derived. Thus, in addition to the atomic-scale analysis of surface structure, any changes in Φ as a function of lateral position on the sample surface can be determined and related to a chemical bond of adsorbed species (i.e., providing information on local surface contamination).

Another example of the STM capabilities is the measurement of the electric potential distribution with nanometer-scale resolution. This technique, i.e., *scanning tunneling potentiometry* involves the application of a voltage across the sample and the detection of the potential gradient variations. Thus, local potential nonuniformities related to semiconductor structures and devices, or defects (e.g., grain boundaries) can be determined with high spatial resolution. (In this case, it is important to differentiate between the types of information related to surface topography and electric potential distribution.)

These basic principles of SPM can be applied in various configurations for the analysis of a wide range of materials properties. For the STM analysis, the surface of the material has to be electrically conductive. *Atomic force microscopy* (AFM), on the other hand, can be employed in the analysis of conductors as well as insulators. In this SPM method, a tip, such as a diamond crystal fragment, is attached to a flexible cantilever that is deflected due to the interaction force between the tip and the sample surface. The interaction force, such as interatomic forces, experienced by the tip can be derived from the deflection of the cantilever that can be measured employing electron-tunneling detection or optical detection. In one example of the optical-detection method, a diode laser beam is reflected off the cantilever to a position-sensitive photodetector. The deflection, and thus the interaction force, is controlled by a feedback system, which allows one to record the topography of the sample surface employing the constant-force mode of operation that is analogous to the constant current-mode in STM. The common

interatomic force that contributes to the deflection of the cantilever is the Van der Waals force, which is repulsive in a contact mode (i.e., tip-to-sample separation is less than a few Angstroms) and attractive in non-contact mode (for tip-to-sample separation in the range between about 10 and 100 Å).

Additional SPM techniques, relevant to semiconductors and semiconductor devices, include *ballistic-electron-emission microscopy* (BEEM) for the nondestructive analysis of subsurface interfaces (e.g., metal–semiconductor interface), and *scanning near-field optical microscopy* (SNOM).

The BEEM technique is employed in the nondestructive analysis of subsurface interfaces, and it allows direct imaging of interfaces with nanometer-scale resolution. In BEEM, three electrodes are used; these are the STM emitting tip, a biasing electrode, and the collecting electrode, which allow the analysis of the behavior of ballistic electrons traversing the specimen surface layer to the interface region. This technique can provide direct imaging of subsurface interface electronic structure by scanning the STM tip across the specimen surface and simultaneously measuring the collector current. In addition, by measuring the collector current as a function of an applied specimen-tip voltage, spectroscopic information can also be obtained. This technique allows one to obtain spatially-resolved measurements of the barrier height for various junctions, as well as information on interface electronic density of states and heterojunction band offsets.

Basic principles of SPM have also facilitated the development of new methods for the analysis of the optical properties of materials. One such technique is a SNOM, which allows achieving a spatial resolution exceeding the far-field diffraction limit by removing lenses completely from the imaging system and employing the near-field collimation of light. The basic principle of this method is the utilization of a subwavelength-size aperture (for light emission or collection), which is scanned above the sample surface in the near-field of an object (see Fig. 7.10). For the scanning and fine positioning of the aperture, piezoelectric transducers are used. The microscopic image is obtained when the aperture is mechanically scanned (in a raster pattern) at a constant distance relative to the sample and the light transmitted through the sample is collected with a microscope objective in the far-field. Different modes of operation (i.e., illumination, collection, and reflection modes) have been developed for SNOM (see Fig. 7.11). In the illumination mode, a collimated beam emanating from a subwavelength-size aperture in an opaque screen is brought to a small distance (less than the diameter of the aperture) to the sample. In this case, the image is obtained when the aperture is mechanically scanned (in a raster pattern) at a constant distance relative to the sample and the light transmitted through the sample is collected with a microscope objective in the far-field. The resolution (determined by the aperture size) of less than 50 nm can be obtained. In the collection mode, the light is focused on to the sample using a microscope objective, and a fraction of radiation transmitted through the sample is collected by the aperture in the near-field. In the reflection mode, a subwavelength-size aperture is used for both the illumination and collection of light reflected from the sample surface. Examples of applications include (i) semiconductor imaging and defect analysis, (ii) investigation of nanostructures, (iii) optical spectroscopy (luminescence, Raman), and (iv) localized photoconductivity.

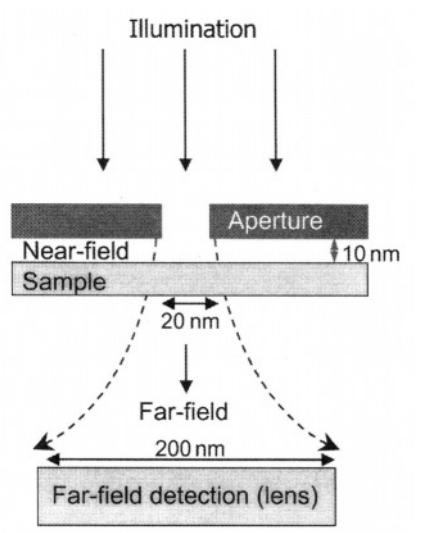

FIGURE 7.10. Schematic diagram of the principle of operation of a SNOM.

There are also some possible applications employing a STM in combination with optical detectors and probes. For example, it is possible to measure luminescence spectra by using *photon emission spectroscopy* and *microscopy*; in this case, electrons injected by tunneling into the bulk conduction band of a semiconductor may recombine radiatively and lead to the emission of characteristic luminescence that can provide local information on electronic properties of the material with the nanometer-scale spatial resolution. Using *scanning tunneling optical spectroscopy*, it is also possible to determine the semiconductor energy gap

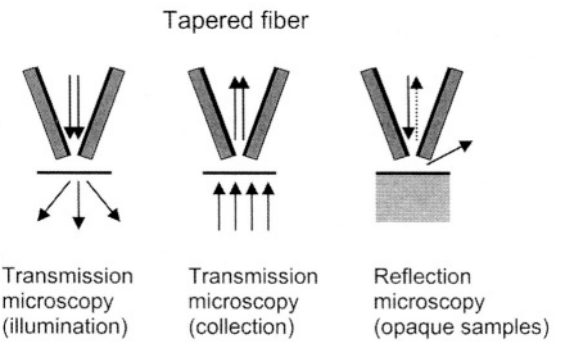

FIGURE 7.11. Schematic diagram of various modes of operation (i.e., illumination, collection, and reflection modes) for SNOM.

TABLE 7.1. Main features related to electron microscopy (including SEM and TEM), and SPM (including STM and AFM)

	SEM	TEM	SPM
Operating environment	Vacuum	Vacuum	Ambient, liquid, vacuum
Sample preparation needed	Little	Thinning	None
Spatial resolution	5 nm–1 μm	0.2 nm	0.1 nm (STM) 0.1–1 nm (AFM)
Magnification range	10×–500,000×	50×–10^6×	500×–10^8×

Note that the typical operating magnification ranges are in between the extreme values shown. Also note that in the SEM, the spatial resolution depends on the specific mode.

with the nanometer-scale spatial resolution by employing the measurement of photoenhanced scanning tunneling current in the semiconductor sample illuminated with appropriate monochromatic light.

For more details on these and other emerging SPM techniques, which will certainly become more important with the continuing trend towards miniaturization of semiconductor structures and devices, see Wiesendanger (1994) and Wiesendanger and Güntherodt (1995) in Bibliography Section B3.

The main features of the three main microscopy techniques are summarized in Table 7.1. It should be emphasized that microscopy techniques typically provide very small sampling volumes and/or fields of view, resulting in a need to ensure (by performing multiple measurements in different regions of the sample) that the measurements are not atypical for a given analysis area.

7.5. STRUCTURAL ANALYSIS

The main objective of structural characterization is the description of the three-dimensional arrangement of the atoms in a solid. In practice, the main objectives include measuring the lengths and angles in the unit cell, i.e., the lattice parameters, and determining the arrangement of the atoms in the unit cell. Some structural features of interest are of a macroscopic nature (e.g., cracks), and some are of microscopic nature (e.g., vacancies).

The standard techniques for structural characterization of semiconductors are X-ray diffraction, neutron diffraction analysis, electron microscopy and electron diffraction, and Raman spectroscopy. These techniques are used for deriving structural information, such as crystalline state (i.e., whether the material is crystalline, polycrystalline, or amorphous), crystalline defects, phase transformations, and stresses present in the material.

7.5.1. X-ray Diffraction

One of the most powerful techniques for the characterization of the structural properties of semiconductors is the *X-ray diffraction* (XRD). In this technique, a

sample is irradiated with a collimated beam of X-rays (with wavelengths between about 0.5 and 2 Å) and the scattered X-rays are detected with an appropriate detector. Depending on factors such as the orientations of the sample and detector, and on the specific crystal structure of the sample material, XRD pattern can be recorded. Such a pattern consists of peaks in the scattered X-ray intensity plotted as a function of scattering angle. (Note that the peaks are due to constructive interference of the scattered X-rays.) In XRD, the X-rays are diffracted by the crystalline material according to Bragg's law, i.e., $n\lambda = 2d\sin\theta$ (see Chapter 2). One can obtain information on phases present, crystal structure, defects (from defect imaging), crystallite sizes, crystal orientation, and strain. The phase identification is one of the routine applications of the XRD, and it involves comparing the derived d_{hkl}-spacing (and peak intensities) from the diffraction spectra with those for known standards given in the literature. Preferred grain orientation can be derived from the relative peak intensities of the crystallographic directions, whereas the strain can be characterized by the position and width of the diffraction peaks, and crystallite size can be determined from the width of the diffraction peaks. The main advantages of the XRD include the facts that (i) analysis can be performed at ambient conditions, (ii) large area samples with little preparation can be used, and (iii) the nondestructive nature of the analysis.

7.5.2. Electron Diffraction

As outlined earlier, the crystalline structure can be analyzed in TEM by employing a TED pattern that can be obtained by inserting a SAD aperture in the image plane of the objective lens. The TEM-related techniques, however, are destructive (i.e., they involve thinning of the sample) and they also require vacuum. One of the advantages of the TED, compared to XRD, is high intensity for electron diffraction, which is orders of magnitude greater than in the case of XRD.

The concept of a reciprocal lattice, which was discussed in Chapter 2, facilitates the interpretation of diffraction patterns, since it allows identifying the sets of reflecting planes that cause a diffraction spot in a diffraction pattern. As mentioned in Chapter 2, the reciprocal lattice can be constructed by plotting a normal to each set of planes in the direct lattice and specifying points along these normals at distances $1/d$ from the origin; the combined set of all these points produce the basic array of the reciprocal lattice. For the analysis of the diffraction patterns, it is possible to select a two-dimensional section out of the three-dimensional reciprocal lattice; the resulting specific array of reciprocal lattice points corresponds to the array of diffraction spots in the diffraction pattern. Such a construction also facilitates labeling of the individual spots with appropriate Miller indices. The frequently encountered sections of the three-dimensional reciprocal lattices of the commonly observed crystal structures are available in the literature; thus, a specific diffraction pattern can be indexed by selecting the reciprocal lattice section with such an array of points that corresponds to the diffraction spots in the diffraction pattern.

7.5.3. Structural Analysis of Surfaces

For structural analysis of surfaces, *low-energy electron diffraction* (LEED) and *reflection high-energy electron diffraction* (RHEED) can be employed. In LEED, a collimated monoenergetic electron beam is diffracted by the sample surface. The electron-beam energies employed are in the range between about 10 and 1000 eV (corresponding to about 4.0 and 0.4 Å, respectively). The main applications of LEED measurements are in the analysis of properties such as the state of surface cleanliness, surface crystallography, and surface microstructure. For the analysis of surfaces in electron probe instruments, it is also possible to employ reflected electrons for RHEED. In RHEED, a high-energy electron beam, striking the sample at a grazing angle of about 1–5°, is scattered by the sample surface. The electron-beam energies employed in this case are in the range between about 5 and 50 keV. In RHEED measurements, the pattern produced on a phosphor screen, positioned opposite the electron gun, can be monitored or recorded by various methods; from the features of such RHEED patterns (i.e., from the spacing and symmetry of the features) one can derive information on the lattice constant, surface structure, and surface symmetry. This technique is extensively employed as an *in situ* monitoring tool in molecular beam epitaxy (MBE).

7.6. SURFACE ANALYSIS METHODS

A detailed knowledge about surfaces and surface states is crucial in semiconductor technology. In general, the properties of semiconductor surfaces (and interfaces) vary widely depending on the processing, or exposure to air. Thus, prior to metallization it is essential to determine the presence of any specific oxides or contaminants on semiconductor surfaces exposed to air. This necessitates the use of surface-sensitive characterization techniques. The most commonly used surface characterization techniques are *Auger electron spectroscopy* (AES) and *scanning Auger electron microscopy* (SAEM), *X-ray photoelectron spectroscopy* (XPS), *secondary ion mass spectrometry* (SIMS), and *Rutherford backscattering spectrometry* (RBS). These techniques, having advantages and disadvantages for different applications, provide very important information about surface (and subsurface) and interface properties of semiconductors. In most cases, these surface analytical techniques probe the top several monolayers of the material (with subsequent *in situ* etching of the surface with ion beam, one can also derive bulk information and depth profiling of the property of interest), and they provide information about the composition, concentration and distribution of various species in the material, and also chemical information. In addition to these techniques, scanning probe microscopies (e.g., STM and AFM), as well as SEM are widely used for the analysis of surface properties.

The major applications of surface analysis methods involve (i) investigations of surface composition (including surface contamination) in correlation with other materials and device properties, (ii) depth profiling of thin films and thin-film multilayer structures and devices, and (iii) obtaining the microscopic description of surface, such as surface topography and various types of inhomogeneities.

7.6.1. Auger Electron Spectroscopy

AES is based on the detection of the electron that has been ejected due to the re-arrangement of core electrons in the atom as a result of primary electron beam bombardment with typical energies of about 5 keV. In this process, after the ionization of an atomic core level by an electron bombardment, the filling of unoccupied level (e.g., *K* level) by an electron from a higher energy level (e.g., one of the *L* levels) is accompanied by the emission of a photon (e.g., characteristic X-ray). Another likely event, however, is the energy transfer to another electron (in the same level or close to it) that carries off the energy gained by the first electron; such an escaping electron is called an *Auger electron.* The Auger process is designated in terms of three levels involved, which are (i) the level in which the electron-beam-excitation-induced vacancy is created, (ii) the level associated with the electron filling the vacancy, and (iii) the level from which the Auger electron is emitted. The principal Auger transitions (and energies) involve electrons in neighboring orbitals, such as KLL, LMM, and MNN (see Fig. 7.12, illustrating the process for $KL_1L_{2,3}$ transition). The charts of the principal Auger electron energies as a function of the atomic number for the above transitions (i.e., KLL, LMM, and MNN) allow elemental identification. Thus, using this technique, binding energies of core electrons in the atom can be deduced and the chemical elements can be identified. All elements above helium can be analyzed with detection limits in the range between about 0.01 and 1 at.%. (The detection of hydrogen and helium is not possible, since the Auger process involves three electrons.) The dominant Auger electron transitions can be related to the atomic

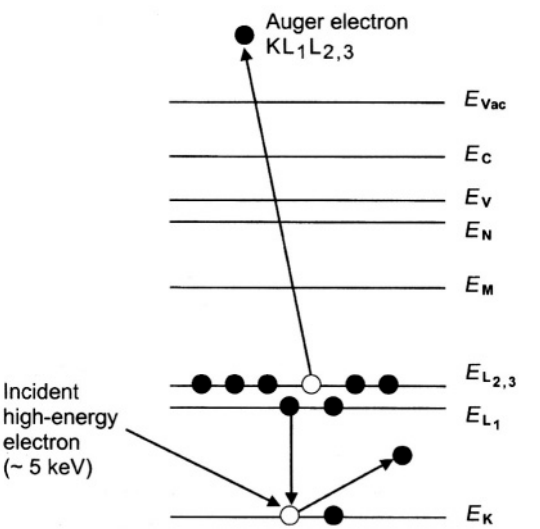

FIGURE 7.12. Schematic depiction of electronic processes involved in Auger electron spectroscopy, showing the generation of Auger electrons (through $KL_1L_{2,3}$ transition process) as a result of incident electron bombardment.

number of the element being analyzed as follows: KLL transitions are dominant for $3<Z<14$, LMM transitions for $14<Z<40$, and MNN for $40<Z<82$.

The analysis can be quantified. Although, in principle, Auger electron peak heights are directly proportional to elemental concentrations, a number of both instrumental factors and materials properties affect these peak heights. These typically include incident beam energy, the characteristics of the analyzer, sample orientation, and the chemical states of elements in the sample. Thus, to summarize briefly, in Auger spectroscopy, the kinetic energies of the emitted electrons are measured, with each element in a given sample producing a characteristic spectrum of peaks at various kinetic energies. Note that Auger spectra are commonly presented in a differentiated form, which offers improved detection sensitivity. In this case, the derivative of the electron energy distribution $N(E)$ is obtained and plotted as $dN(E)/dE$ vs. E.

Only those Auger electrons that are generated close to the surface (from about the top 10 Å of the material) can escape from the material. (Those generated deep within the solid are reabsorbed.) Because of such a very short escape distance of electrons, the techniques based on the detection of such electrons are inherently surface-sensitive. Thus, AES, providing analysis of surface compositional variations, constitutes an important surface analytical technique in ultrahigh-vacuum electron probe instruments.

Auger electron spectroscopy is often combined with the slow removal of outermost atomic layers of the material by sputtering employing Argon ions; thus, AES can provide depth profiling of chemical constituents. Spatial and depth resolutions of about 100 and 10 Å, respectively, can be obtained. By employing scanning Auger electron microscopy (i.e., the excitation electron beam is focused and scanned across a sample surface), it is also possible to obtain compositional images with high spatial resolution. In addition, in the scanning Auger electron microscopy system one can also obtain secondary electron images (or other images if appropriate detectors are available), allowing a comparison between the compositional images and surface morphology. The major advantages of AES are (i) its high surface sensitivity (to a few monolayers), (ii) elemental sensitivity and range, (iii) excellent spatial and depth resolutions, and (iv) the capability of obtaining elemental mapping. Some major disadvantages include (i) possible electron-beam-induced charging and damage, (ii) possible effect of beam-induced artifacts on chemical state information, and (iii) sensitivity limitations for quantitative detection of low-level concentrations of less than about 0.01%.

7.6.2. Photoelectron Spectroscopy

X-ray photoelectron spectroscopy, XPS (also referred to as Electron Spectroscopy for Chemical Analysis, ESCA), which employs X-rays for the excitation of the solid and the detection of emitted photoelectrons with characteristic energies, can provide chemical information about the material. In general, the kinetic energy E_k of the photoelectron depends on the photon energy $h\nu$ (in this case, X-ray) following the Einstein photoelectric relation, i.e., $E_k = h\nu - E_B$, where E_B is the binding energy of the specific electron to a particular atom. (The binding energy is a measure of the

energy required to just remove an electron from its initial level to the vacuum level; since the electron binding energies in solids are typically measured relative to the Fermi level, rather than the vacuum level, a small correction to the above equation is made to account for the work function of the solid.) The photoelectrons that have sufficient kinetic energy can escape from the surface of the sample by overcoming its work function. Thus, from the measurement of the photoelectron kinetic energy, one can determine the electron binding energy, which is characteristic to the particular atom, and thus the corresponding atom can be identified. Due to quantized energy levels in atoms, the photoelectron kinetic energy distribution is comprised of a series of discrete bands. Since the energies of photoelectrons are much less than 1 keV, the escape depth, and thus the depth resolution, is within about 20 Å of the surface. Thus, in this technique, the energy spectrum of the photoelectrons, which are emitted from the sample, provides nondestructive elemental and chemical analysis of the surface. Main applications of photoelectron spectroscopy are in determining binding energies, in the analysis of the band structure of solids, and in detecting particular elements present at the surface of the material. Since the atomic environment influences the binding energy of an electron, it is possible to obtain information on chemical bonding of a particular element from the *chemical shift,* which enables identification of the compounds. It is also possible to derive nondestructive depth information from *angle-resolved* XPS.

The main advantages of the XPS are (i) its sensitivity, (ii) nondestructive nature of the analysis, (iii) minimal sample charging and beam damage, and (iv) the capability of the analysis of chemical shifts from the same element in different compounds. Some major disadvantages include (i) relatively poor spatial resolution, (ii) difficulties involved with depth profiling due to excitation probe size, (iii) some relative difficulties with interpretation, and (iv) sensitivity limitations for detecting low-level concentrations of less than about 0.1%.

Another photoelectron spectroscopy technique, i.e., *UV photoelectron spectroscopy* (UPS), employs UV photons (with energies $\leqslant$50 eV) that result in lower kinetic energies for photoelectrons, and this allows investigation of the valence band states only.

To summarize, (i) photoelectron spectroscopy employs photo-ionization and energy-dispersive analysis of the emitted photoelectrons to derive the composition and electronic state of the surface region of a sample, (ii) two techniques are distinguished (according to the source of exciting radiation), as XPS (using X-ray excitation to study core-levels), and UPS (using vacuum UV radiation from discharge lamps to investigate valence levels), and (iii) each element has a characteristic binding energy associated with each core atomic orbital, and, thus, the presence of a specific element in the sample can be determined, with the intensity of the characteristic peaks being correlated with the concentration of that specific element in the sampled region.

7.6.3. Ion-Beam Techniques

As mentioned earlier, ion beam techniques are distinguished between those that employ the excitation ion beam energies in the keV range and those that

employ the ion beams in the MeV range. In *secondary ion mass spectrometry* (SIMS), secondary ions emitted as a result of the incident ion-beam bombardment (with energies up to about 20 keV) are identified using a mass-spectrometer, and the ion-beam sputtering of the sample allows obtaining depth profiling. In *Rutherford backscattering spectrometry* (RBS), high-energy ion beams in the MeV range are employed for the excitation of the material.

7.6.3.1. Ion Beam–Solid Interaction Processes In ion beam–solid interaction processes, the characteristics of energy transferred depend mainly on (i) the energy and angle of incidence of the incoming (incident) ion, and (ii) the masses of the incident ion (M_1) and of the target nuclei (M_2). Several limits of the incident ion energy should be considered in order to understand different processes associated with the interactions between the high-energy ions and a solid. Incident ions with energies between about 1 and 20 keV transfer energy to the target's nuclei, leading to sputtering (i.e., sequential removal of surface layers), which is often employed in the sputtering of the target in the deposition of various thin films and also in various analytical techniques in order to obtain depth-resolved information. At incident ion energies in the range between about 30 and 100 keV, ions can be implanted into the target material to a depth of about 0.1 μm. Such an ion implantation is employed to control the electrical properties of semiconductors. High-energy (MeV) ions (e.g., H or He ions) can penetrate deeply (several microns) into the target. At the initial stages of penetration, incident ions loose small amounts of energy through electronic collisions (the high-energy ion excites electrons from target atoms) until, at lower ion energies, nuclear collisions occur and energy is lost in displacements of the target atoms. It is important to note that, since incident ions with initial high energies cause insignificant damage to the target, the backscattered ions can be used for the analysis, such as RBS (see Section 7.6.3.3).

7.6.3.2. Secondary Ion Mass Spectrometry In this technique, the primary (incident) ion bombardment of the specimen results in the emission of secondary ions that are analyzed in a mass-spectrometer for their identification. The primary beam of ions (e.g., O_2^+, Cs^+, or Ar^+) with energies of up to 20 keV is employed. Ion beam sputtering of the specimen allows obtaining depth profiles. However, the secondary ion yield in this technique varies widely for different elements, and it also depends strongly on the *matrix effect* (i.e., given combination of element and a matrix), which has to be taken into account in quantitative analysis, which is possible by using appropriate standards. Two variations of this technique are the *static* SIMS and *dynamic* SIMS. The main difference between them is in the primary ion dose used. During static SIMS measurements, the total primary ion dose is sufficiently low for the analysis of the surface monolayer only (however, signal levels are correspondingly low), whereas in dynamic SIMS the primary ion dose is sufficiently high in order to maximize the signal for trace element analysis. In such a case, due to high primary ion doses, the sample surface is rapidly eroded and that provides depth-resolved information. The quantitative information can be obtained by converting the ion count rates to atomic concentrations by using

the so-called *relative sensitivity factor,* derived from measurements on standards of known composition. (Such standards are usually obtained by ion implantation of given elements into a matrix of interest.) It should be noted that the relative sensitivity factors depend strongly on particular materials and specific SIMS operating conditions.

There are two methods for obtaining SIMS images. In a direct imaging mode in the ion microscope (analogous to a TEM), it is possible to display secondary ion images, which provide information on the lateral distribution of elements of interest in the sample. In a scanning imaging method, the secondary ion intensity is monitored as a function of the (lateral) position of a finely focused scanning ion beam. By obtaining a series of secondary ion images as a function of depth, it is also possible to derive three-dimensional compositional information.

SIMS is widely employed in the depth-resolved analysis of dopants in diffused or implanted layers of semiconductors. The major advantages of this technique are (i) its high sensitivity for all elements, including hydrogen, with detection limits (in atomic fraction) in the range between about 10^{-6} and 10^{-9}, (ii) an excellent depth resolution, and (iii) the capability of obtaining secondary ion images. Some major disadvantages include (i) the difficulty involved in obtaining quantitative information, (ii) the destructive nature of the analysis, and (iii) the difficulty related to the identification of the elements due to mass interferences.

7.6.3.3. Rutherford Backscattering Spectrometry Rutherford backscattering spectrometry (RBS) is a depth profiling technique, which allows obtaining quantitative information on elemental composition or impurity concentration without the need for standards. RBS employs high-energy ion (typically He ions with energies of the order of 1 MeV) bombardment of the materials surface and the measurement of the energy of the backscattered ions. Since the energy loss during the penetration of the material is proportional to the penetration distance, a depth scale (the depth range is up to about 1 μm) can be related to the backscattered ion energy spectra. Also, since measured yield of backscattered ions is proportional to the scattering cross-section, quantitative information on the impurity concentration can be obtained, and the composition depth profile can be determined from the energy loss and cross-section measurements. As mentioned in Section 7.6.3.1, in the ion beam–solid interaction processes, the characteristics of energy transferred depend mainly on (i) the energy and angle of incidence of the incident ion and (ii) the masses of the incident ion M_1 and of the target nuclei M_2. In this process, some of the incident high-energy ions are elastically backscattered, transferring some energy to the stationary atom of a target. The laws of conservation of energy and momentum can describe the interaction in such an elastic scattering process. An important parameter for such a process is the *kinematic factor K*, which is defined as the ratio of the energy of the incident ion after the collision (E_1) to its energy before the collision (E_0). Thus, the *kinematic factor* is a measure of the energy loss by incident ions, and the energy of the incident ion after the collision is determined by the masses of the incident ion and of the target atom, and the scattering angle. By applying

the conservation of energy and momentum, the kinematic factor can be expressed, for 180° backscattering angle, as (for $M_1 < M_2$)

$$K = E_1/E_0 = (M_2 - M_1)^2/(M_2 + M_1)^2 \qquad (7.6.1)$$

where M_1 is the mass of the incident ion (e.g., He or H) and M_2 is the mass of the target atom. Using these conditions, one can obtain the energy spectrum of backscattered ion. Note that for large M_2 of the target atom, the energy spectrum will be somewhat shifted, whereas for smaller M_2 the energy shift will be relatively greater. This indicates the high sensitivity of the RBS for detecting heavy elements in light-element host matrices. In such a case, the heavy-element impurity signal appears at higher energies of the RBS spectrum (see Fig. 7.13). From the shape of this high-energy peak, one can derive the depth distribution and concentration of the impurities.

An important feature of this technique is the *channeling* phenomenon, which can provide information about the structure of the material, as well as the specific location of the impurity atoms, i.e., whether they occupy lattice or interstitial sites (see Fig. 7.14). In the channeling measurement, the interaction of the incident beam with a crystalline sample is monitored employing the backscattering mode. The sample is oriented with one of its crystallographic axes parallel to the incident beam, and the alignment of the incident beam with the axis results in a significant decrease in backscattering counts. The relevant information related to defects is derived from the *dechanneling,* i.e., deterioration of channeling due to defects in crystalline materials. In principle, most defect types have an effect on channeling, and thus this technique provides useful means for examining various defects. The effect of defects, such as dislocations and interstitials, on the channeling spectrum is shown in Fig. 7.15.

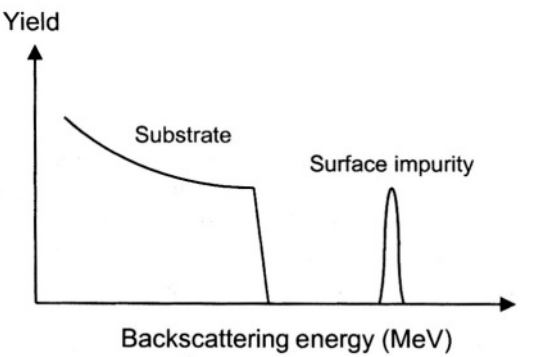

FIGURE 7.13. Schematic illustration of an RBS spectrum of a light-element target including a heavy-element surface impurity. (Note that this is a simplified depiction presented as illustration only, and the actual RBS spectra may contain additional features.)

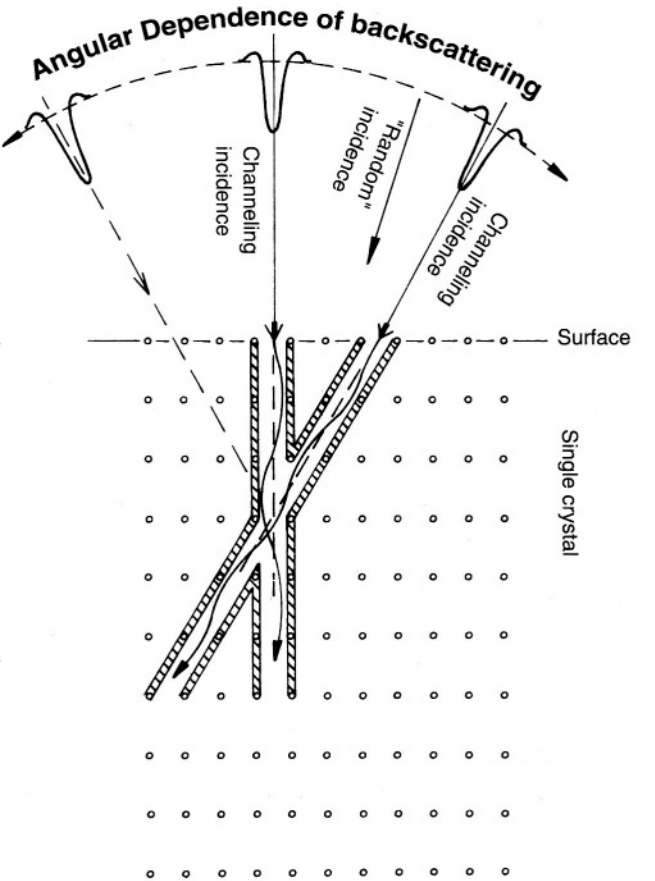

FIGURE 7.14. Schematic representation of the channeling measurement; the minima in the backscattering yield correspond to the direction of the incident beam being aligned parallel with one of the crystallographic axes (see Revesz and Li 1994, in *Microanalysis of Solids* in Bibliography Section B3).

The capabilities for analyzing depth profiling of impurity concentration, possible damage, and specific lattice site location are especially useful for the (i) examination (quantitative depth profiling) of dopants and defects in thin-film structures (including superlattices) and ion-implanted semiconductors and (ii) investigations related to the near-surface processing of semiconductor devices. Specifically, applications include analysis of interface contamination and of interdiffusion kinetics. The major advantages of this technique are (i) its ability to obtain quantitative information, such as elemental composition, without standards (RBS is used to calibrate spectra obtained by SIMS and AES), (ii) depth profiling in a nondestructive manner and (iii) detection limits (in atomic fraction) in the range between about 10^{-1} and 10^{-4} depending on the atomic

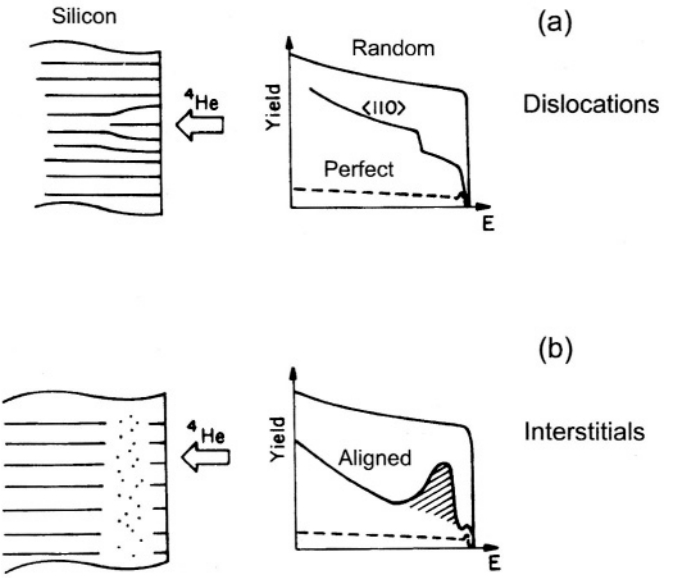

FIGURE 7.15. (a) The effect of dislocations on the channeling spectrum; (b) The effect of interstitials on the channeling spectrum. (see Revesz and Li 1994, in *Microanalysis of Solids* in Bibliography Section B3).

number. Some disadvantages include (i) poor spatial resolution and no lateral imaging capability, (ii) its limitations for detecting elements with low atomic number, and (iii) the limitation related to the analysis of materials with more than a few elements.

7.6.4. Comparison of Surface Analytical Techniques

The general comparison of surface analytical techniques indicates that, on the basis of overall performance (e.g., elemental resolution, spatial resolution, depth resolution, relative simplicity of operation, equipment availability, analysis time, and availability of extensive data base), AES is the most useful and routinely employed technique. The main advantage of XPS is in its ability to provide detailed chemical binding information. Among these techniques, SIMS is the most sensitive technique for the analysis of low concentrations of dopants and impurities (although for quantitative analysis, calibrated standards are required). The strength of RBS is in its ability to obtain quantitative information without standards. Taking into account both the advantages and limitations of each technique, in general, the effectiveness of each individual technique can be further augmented by applying different techniques to the analysis of the same sample.

The basic characteristics of commonly employed surface analysis techniques are summarized in Table 7.2.

TABLE 7.2. Summary of the basic features of commonly employed surface analysis techniques

Features/technique	AES	XPS	SIMS	RBS
Incident (excitation) particles	Electrons	X-rays	Ions	Ions
Excitation energy (keV)	1–20	1–3	0.5–20	500–3000
Emitted particles	Electrons	Electrons	Secondary ions	Primary ions
Spatial resolution (μm)	0.01–1	10–100	0.1–10	10–200
Depth resolution (Å)	5–50	5–50	10–100	30–200
Detection limit (at. fraction)	10^{-2}–10^{-3}	10^{-2}–10^{-3}	10^{-6}–10^{-9}	10^{-1}–10^{-4}
Detectable elements	> Li	> Li	> H	> Li

Note that the values (or ranges of values) for various characteristics are typical ones obtained in routine measurements, and these depend strongly on the material, and especially on the element that is being analyzed, and also on the excitation conditions. In some specific cases, however, reports indicate improved characteristics; for example, although the typical spatial resolution for RBS is given as about 1 mm, probes down to about 1 μm have been obtained.

TABLE 7.3. Selected examples for the choice of common characterization methods

Property	Characterization method
Resistivity (conductivity)	Four-point probe, van der Pauw method, spreading resistance, Hall effect
Conductivity type	Hall effect, thermoelectric probe, rectification method
Carrier concentration (density)	Capacitance–voltage, spreading resistance, Hall effect
Carrier mobility	Resistivity and Hall effect, current–voltage
Minority-carrier lifetime	Time-resolved photoconductivity, open circuit voltage decay, surface photovoltage, junction current–voltage, PL, CL
Minority-carrier diffusion length	Surface photovoltage, OBIC, EBIC, junction photocurrent, PL, CL
Surface recombination velocity	PL, time-resolved photoconductivity, CL, surface photovoltage
Deep-level impurities	DLTS and SDLTS, PL, Hall effect, FTIR
Energy gap (and alloy composition)	Optical absorption, reflectivity, PL, CL, temperature dependence of electrical conductivity
Radiative transitions	PL, CL
Elemental composition	Electron probe microanalysis, RBS, AES, XRD, SIMS
Crystallographic structure	XRD, TEM, Raman spectroscopy
Crystalline quality	XRD, TEM, RBS
Structural defects (imaging)	TEM, STM, EBIC, CL–SEM
Structural defects (spectroscopy)	PL, CL, optical absorption, Raman spectroscopy, RBS, DLTS
Surface morphology	Optical microscopy, SEM, STM, TEM
Elemental surface composition	AES, XPS, SIMS, RBS
Interface lattice structure	HRTEM
Dislocation density	TEM, EBIC, CL-SEM, etch-pit density
Strain or stress	XRD, Raman spectroscopy, PL, CL

7.7. SUMMARY

In summary, Table 7.3 lists some examples for the choice of typical characterization methods. This chapter has outlined the techniques that are commonly employed in the analysis of various properties of semiconductors. There is also a wide range of characterization techniques that are more specific and not as routinely employed, and there are those that are being continuously developed for specific applications. For general overviews and more details on various techniques, see Grasserbauer and Werner (1991), Brundle *et al.* (1992), and Schroder (1998). In addition to being outlined in these books, some specific characterization techniques are also discussed in more detail in the dedicated sources listed in Bibliography Section B3.

PROBLEMS

7.1. Describe the methods for measuring the energy gap of a semiconductor; discuss their advantages and disadvantages.

7.2. If an unknown semiconductor is given, describe at least two methods for its unambiguous identification.

7.3. Describe the fundamental differences between photon-beam and electron-beam excitation of semiconductors.

7.4. Outline the problems related to the quantification of luminescence analysis.

7.5. Discuss why Auger electrons are suitable for surface analysis of materials.

This page intentionally left blank

APPENDIX A

Physical Constants

Avogadro's constant	$N_A = 6.022 \times 10^{23}$ molecules mol^{-1}
Boltzmann's constant	$k_B = 1.381 \times 10^{-23}$ J $K^{-1} = 8.617 \times 10^{-5}$ eV K^{-1}
Bohr radius	$a_0 = 5.292 \times 10^{-11}$ m
Electron charge	$e = 1.602 \times 10^{-19}$ C
Electron rest mass	$m_0 = 9.11 \times 10^{-31}$ kg
Permittivity of vacuum	$\varepsilon_0 = 8.85 \times 10^{-12}$ F m^{-1}
Planck's constant	$h = 6.626 \times 10^{-34}$ J s $= 4.136 \times 10^{-15}$ eV s
Speed of light in vacuum	$c = 2.998 \times 10^{8}$ m s^{-1}

Useful Conversion Factors

1 eV $= 1.6 \times 10^{-19}$ J
Thermal energy, $k_B T = 0.0259$ eV (at 300 K)
The wavelength λ (μm) of a photon is related to the photon energy E (eV) as $\lambda = 1.2398/E$
1 Å $= 10^{-4}$ μm $= 10^{-8}$ cm
1 nm $= 10$ Å
1 μm $= 10^{3}$ nm

Prefixes

atto (a)	$\times 10^{-18}$
femto (f)	$\times 10^{-15}$
pico (p)	$\times 10^{-12}$
nano (n)	$\times 10^{-9}$
micro (μ)	$\times 10^{-6}$
milli (m)	$\times 10^{-3}$
kilo (k)	$\times 10^{3}$
mega (M)	$\times 10^{6}$
giga (G)	$\times 10^{9}$
tera (T)	$\times 10^{12}$
peta (P)	$\times 10^{15}$
exa (E)	$\times 10^{18}$

APPENDIX B

TABLE B1. Properties of common semiconductors at room temperature (300 K)

Material (structure)	E_g (eV)	Transition	Doping	ε_s	μ_e [cm^2 (V s)$^{-1}$]	μ_h [cm^2 (V s)$^{-1}$]	a (Å)	Density (g cm^{-3})	Melting point (K)
Ge (D)	0.67	i	n, p	16	3900	1900	5.646	5.327	1230
Si (D)	1.12	i	n, p	11.9	1700	600	5.431	2.329	1685
C (D)	5.47	i	p	5.7	2200	1600	3.567	3.515	4100
SiC (Z)	2.30	i	n, p	9.7	1000	40	4.360	3.166	3100
InSb (Z)	0.17	d	n, p	17.7	77000	1000	6.479	5.775	798
InAs (Z)	0.36	d	n, p	15.2	33000	460	6.058	5.66	1215
GaSb (Z)	0.72	d	n, p	15.7	5000	1000	6.096	5.619	980
InP (Z)	1.35	d	n, p	12.6	4600	150	5.869	4.787	1335
GaAs (Z)	1.42	d	n, p	12.9	8500	400	5.653	5.318	1510
AlSb (Z)	1.58	i	n, p	11.6	900	400	6.136	4.218	1330
AlAs (Z)	2.16	i	n, p	10.9	1200	400	5.662	3.76	1870
GaP (Z)	2.27	i	n, p	11.1	300	150	5.451	4.138	1730
AlP (Z)	2.45	i	n, p	9.8	80		5.464	2.40	2823
GaN (W)	3.44	d	n	10.0	1000	30	3.189	6.15	1920
AlN (W)	6.28	d	n	9.1	135	14	3.11	3.23	3000
CdTe (Z)	1.56	d	n, p	10.2	1050	100	6.482	6.20	1365
CdSe (W)	1.70	d	n	9.5	720	75	4.30	5.81	1530
ZnTe (Z)	2.28	d	p	9.3	340	100	6.101	5.64	1510
CdS (W)	2.42	d	n	8.9	300	50	4.136	4.82	1750
ZnSe (Z)	2.70	d	n	9.2	600	80	5.668	5.27	1790
ZnS (Z)	3.68	d	n	8.9	165	40	5.410	4.075	2100
PbSe (R)	0.28	d	n, p	250	1000	900	6.117	8.26	1355
PbTe (R)	0.31	d	n, p	412	1800	900	6.462	8.22	1197
PbS (R)	0.41	d	n, p	170	600	700	5.936	7.61	1383

The energy gap, E_g; Transition (d or i) indicates whether it is a direct or indirect energy gap; Doping (*n* and *p*) indicates commonly observed doping type (note that in some materials, one of the doping types, although achievable and reported in the literature, is not readily obtainable); (relative) static dielectric constant, ε_s; mobility of electrons, μ_e, and holes, μ_h (note that these values depend strongly on the purity of the material measured, as well as on the temperature); lattice constant, *a;* density; and melting point. Lattice: D , diamond; W, wurtzite (hexagonal); Z, zincblende (cubic); R, rocksalt. Note that some compounds (e.g., SiC, ZnS) can be grown in either W or Z structures, and they can also exhibit various polytypic structures; in the case of the wurtzite structure, for complete description a lattice constant *c* is also required (see Table 2.3). For more details on a wide range of properties of various semiconductors, see Madelung (1996) and Berger (1997) in Bibliography Section B2.

BIBLIOGRAPHY

B1. Solid State Physics

Ashcroft, N. W. and N. D. Mermin (1976), *Solid State Physics* (Saunders, Philadelphia) [an advanced text].

Blakemore, J. S. (1985), *Solid State Physics* (Cambridge University Press, Cambridge) [an introductory text].

Bube, R. H. (1988), *Electrons in Solids: An Introductory Survey* (Academic Press, San Diego).

Burns, G. (1985), *Solid State Physics* (Academic Press, Orlando) [an introductory text to basic concepts of solid-state physics].

Hummel, R. E. (2001), *Electronic Properties of Materials* (Springer-Verlag, New York).

Kittel, C. (1986), *Introduction to Solid State Physics* (Wiley, New York).

Myers, H. P. (1990), *Introductory Solid State Physics* (Taylor & Francis, London).

Rosenberg, H. M. (1988), *The Solid State* (Oxford University Press, Oxford) [an introductory text].

Rudden, M. N. and J. Wilson (1980), *Elements of Solid State Physics* (Wiley, New York) [an introductory text].

B2. Semiconductors and Their Applications

Balkanski, M. and R. F. Wallis (2000), *Semiconductor Physics and Applications* (Oxford University Press, New York).

Bar-Lev, A. (1993), *Semiconductors and Electronic Devices* (Prentice-Hall, New York) [an introductory text].

Berger, L. I. (1997), *Semiconductor Materials* (CRC Press, Boca Raton, FL) [contains useful information on a wide range of properties of various semiconductors].

Bhattacharya, P. (1997), *Semiconductor Optoelectronic Devices* (Prentice-Hall, Upper Saddle River, NJ).

Böer, K. W. (1990), *Survey of Semiconductor Physics; Volume I: Electrons and Other Particles in Bulk Semiconductors; Volume II: Barriers, Junctions, Surfaces, and Devices* (Van Nostrand Reinhold, New York) [an extensive introductory overview].

Bube, R. H. (1992), *Photoelectronic Properties of Semiconductors* (Cambridge University Press, Cambridge).

Capasso, F. (Ed.) (1990), *Physics of Quantum Electron Devices* (Springer-Verlag, New York).

Chuang, S. L. (1995), *Physics of Optoelectronic Devices* (Wiley, New York).

Cohen, M. L. and J. R. Chelikowsky (1989), *Electronic Structure and Optical Properties of Semiconductors* (Springer-Verlag, New York).

Cooke, M. J. (1990), *Semiconductor Devices* (Prentice Hall International, New York).

Dalven, R. (1980), *Introduction to Applied Solid State Physics* (Plenum Press, New York) [an introductory text].

Davies, J. H. (1998), *The Physics of Low-Dimensional Semiconductors: An Introduction* (Cambridge University Press, Cambridge).

Farges, J. P. (Ed.) (1994), *Organic Conductors* (Marcel Dekker, New York) [several chapters discuss organic semiconductors and their applications].

Ferry, D. K, (1991), *Semiconductors* (Macmillan, New York) [an advanced text with more thorough treatment of transport properties].

Fraser, D. A. (1983), *The Physics of Semiconductor Devices* (Clarendon Press, Oxford) [an introductory text].

Furdyna, J. K. and J. Kossut (Eds.) (1988), *Diluted Magnetic Semiconductors* (Academic Press, San Diego) [this is volume 25 in the series *Semiconductors and Semimetals* (Willardson, R. K., Beer, A. C., and Weber, E. R., Eds.)].

Grovenor, C. R. M. (1989), *Microelectronic Materials* (Adam Hilger, Bristol).

Harrison, P. (2000), *Quantum Wells, Wires and Dots: Theoretical and Computational Physics* (Wiley, New York).

Jackson, K. A. and W. Schröter (2000), *Handbook of Semiconductor Technology* (Wiley, New York).

Jaeger, R. C. (1988), *Introduction to Microelectronic Fabrication* (Addison-Wesley, Reading, MA) [an introductory text].

Jaros, M. (1989), *Physics and Applications of Semiconductor Microstructures* (Clarendon Press, Oxford).

Kano, K. (1998), *Semiconductor Devices* (Prentice-Hall, Upper Saddle River, NJ).

Kasap, S. O. (2002), *Principles of Electronic Materials and Devices* (McGraw-Hill, New York).

Kelly, M. J. (1995), *Low-Dimensional Semiconductors: Materials, Physics, Technology, Devices* (Clarendon Press, Oxford).

Klingshirn, C. F. (1997), *Semiconductor Optics* (Springer-Verlag, New York).

Madelung, O. (Ed.) (1996), *Semiconductors—Basic Data* (Springer-Verlag, New York).

Mahajan, S. and L. C. Kimerling (Eds.) (1992), *Concise Encyclopedia of Semiconducting Materials and Related Technologies* (Pergamon Press, Oxford).

Mayer, J. W. and S. S. Lau (1990), *Electronic Materials Science: For Integrated Circuits in Si and GaAs* (Macmillan, New York) [an introductory text including such topics as properties of semiconductors, and processing, fabrication, and operation of semiconductor devices].

McKelvey, J. P (1986), *Solid State and Semiconductor Physics* (Krieger Publishing, Malabar, FL) [a detailed introductory account of the basic properties of semiconductors].

Moss, T. S. (Ed.) (1992–1994), *Handbook on Semiconductors* (North Holland, Amsterdam), Vols. 1–4. [includes extensive chapters on various topics in semiconductor science (an advanced level)].

Navon, D. H. (1986), *Semiconductor Microdevices and Materials,* (Holt, Rinehart, and Winston Publishing, New York).

Ng, K. K. (1995), *Complete Guide to Semiconductor Devices* (McGraw-Hill, New York).

Pankove, J. I. (1971), *Optical Processes in Semiconductors* (Dover, New York).

Phillips, J. C. (1973), *Bonds and Bands in Semiconductors* (Academic Press, New York).

Pierret, R. F. and G. W. Neudeck (Eds.) (1989), *Modular Series on Solid State Devices* (Addison-Wesley, Reading, MA) [an introductory text including volumes on semiconductor fundamentals, microelectronic fabrication, and semiconductor devices].

Sapoval, B. and C. Hermann (1995), *Physics of Semiconductors* (Springer-Verlag, New York).

Seeger, K. (1999), *Semiconductor Physics: An Introduction* (Springer-Verlag, New York) [an introductory text].

Shur, M. (1996), *Introduction to Electronic Devices* (Wiley, New York) [an introductory text].

Singh, J. (1994), *Semiconductor Devices* (McGraw-Hill, New York).

Smith, R. A. (1978), *Semiconductors* (Cambridge University Press, Cambridge) [an introductory account of the basic properties of semiconductors].

Solymar, L. and D. Walsh (1998), *Electrical Properties of Materials* (Oxford University Press, Oxford).

Street, R. A. (1991), *Hydrogenated Amorphous Silicon* (Cambridge University Press, Cambridge).

Streetman, B. G. (1995), *Solid State Electronic Devices* (Prentice-Hall, Englewood Cliffs, NJ) [an introductory text].

Sze, S. M. (1981), *Physics of Semiconductor Devices* (Wiley, New York) [a comprehensive advanced text].

Sze, S. M. (Ed.) (1998), *Modern Semiconductor Device Physics* (Wiley, New York).

Tyagi, M. S. (1991), *Introduction to Semiconductor Materials and Devices* (Wiley, New York).

Wang, S. (1989), *Fundamentals of Semiconductor Theory and Device Physics* (Prentice-Hall, Englewood Cliffs, NJ) [a comprehensive description of solid-state devices].

Weisbuch, C. and B. Vinter (1991), *Quantum Semiconductor Structures: Fundamentals and Applications* (Academic Press, Boston).

Wenckebach, W. T. (1999), *Essentials of Semiconductor Physics* (Wiley, New York).

Wilson, J. and J. F. B. Hawkes (1998). *Optoelectronics: An Introduction* (Prentice-Hall, Englewood Cliffs, NJ).

Wolf, H. F. (1971), *Semiconductors* (Wiley, New York).

Wolfe, C. M., N. Holonyak, and G. E. Stillman (1989), *Physical Properties of Semiconductors* (Prentice-Hall, Englewood Cliffs, NJ).

Yang, E. S. (1978), *Fundamentals of Semiconductor Devices* (McGraw-Hill, New York).

Yu, P. Y. and M. Cardona (1996), *Fundamentals of Semiconductors: Physics and Materials Properties* (Springer, New York).

B3. Characterization of Semiconductor Materials and Devices

Blood, P. and J. W. Orton (1992), *The Electrical Characterization of Semiconductors: Majority Carriers and Electron States* (Academic Press, London).

Brundle, C. R., C. A. Evans, Jr., and S. Wilson (Eds.) (1992), *Encyclopedia of Materials Characterization* (Butterworth-Heinemann, Boston) [a comprehensive source containing reviews on a wide range of materials characterization techniques].

Grasserbauer, M. and H. W. Werner (Eds.) (1991), *Analysis of Microelectronic Materials and Devices* (Wiley, New York) [a comprehensive source containing extensive reviews on a wide range of characterization methods with emphasis on electronic materials and devices].

Güntherodt, H.-J. and R. Wiesendanger (Eds.) (1994), *Scanning Tunneling Microscopy I* (Springer-Verlag, New York).

Holt, D. B. and D. C. Joy (Eds.) (1989), *SEM Microcharacterization of Semiconductors* (Academic Press, London).

Orton, J. W. and P. Blood (1990), *The Electrical Characterization of Semiconductors: Measurement of Minority Carrier Properties* (Academic Press, London).

Perkowitz, S. (1993), *Optical Characterization of Semiconductors: Infrared, Raman, and Photoluminescence Spectroscopy* (Academic Press, London).

Runyan, W. R. and T. J. Shaffner (1998), *Semiconductor Measurements and Instrumentation* (McGraw-Hill, New York).

Schroder, D. K. (1998), *Semiconductor Material and Device Characterization,* 2nd edn. (Wiley, New York) [a text that covers a wide range of characterization techniques and provides useful summaries on strengths and weaknesses of various characterization techniques].

Stradling, R. A. and P. C. Klipstein (Eds.) (1990), *Growth and Characterization of Semiconductors* (Hilger, New York).

Wiesendanger, R. (1994), *Scanning Probe Microscopy and Spectroscopy* (Cambridge University Press, Cambridge).

Wiesendanger, R. and H.-J. Güntherodt (Eds.) (1995), *Scanning Tunneling Microscopy II*(Springer-Verlag, New York).

Yacobi, B. G., D. B. Holt, and L. L. Kazmerski (Eds.) (1994), *Microanalysis of Solids* (Plenum Press, New York).

This page intentionally left blank

Index