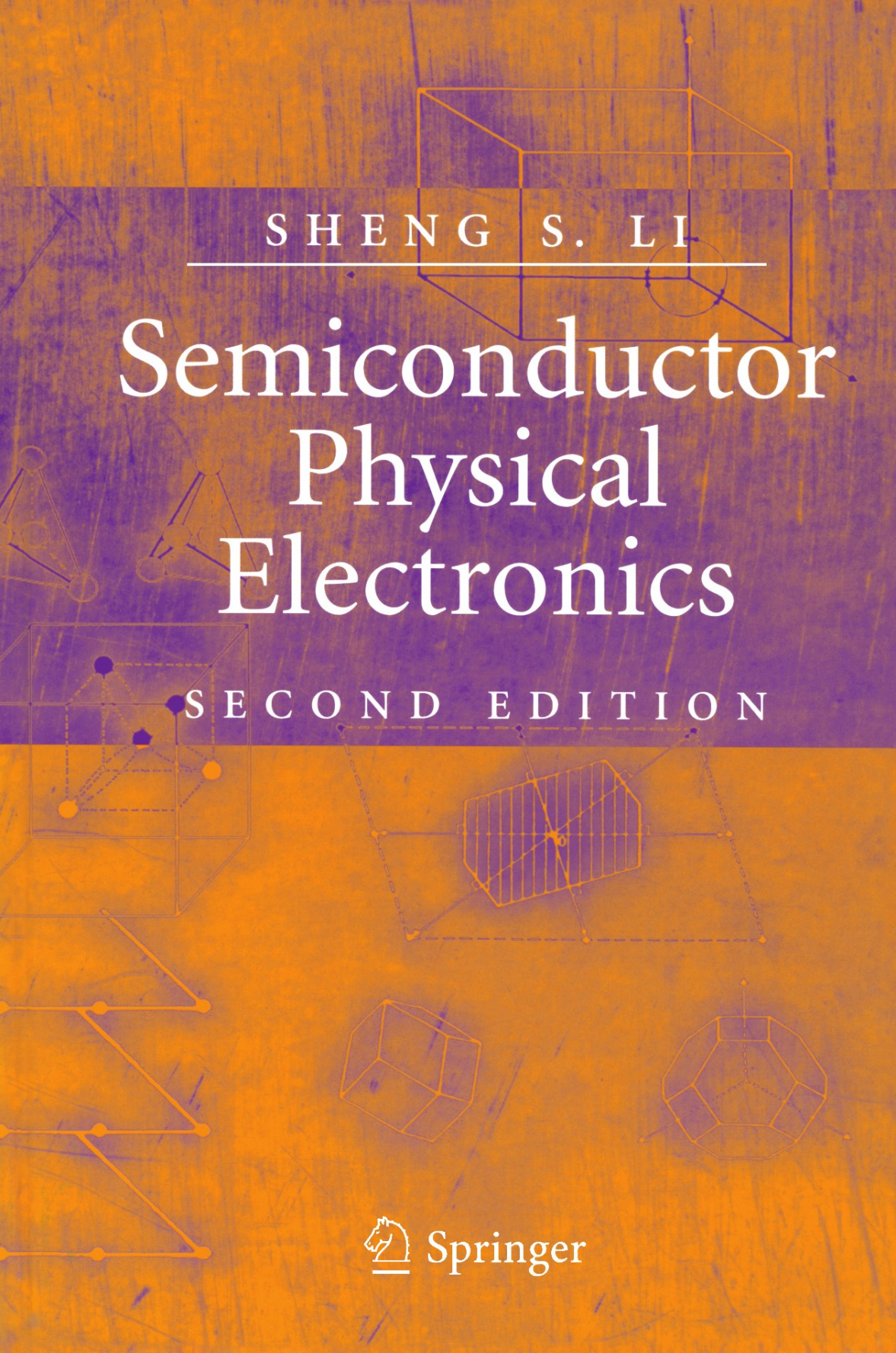
SHENG S. LI

Semiconductor Physical Electronics

SECOND EDITION



Springer

Semiconductor Physical Electronics

Sheng S. Li

Semiconductor Physical Electronics

Second Edition

With 230 Figures

Springer

Sheng S. Li
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32611–6130
USA

Library of Congress Control Number: 2005932828

ISBN 10: 0-387-28893-7
ISBN 13: 978-0387-28893-2

Printed on acid-free paper.

Printed in the United States of America. (TB/EB)

9 8 7 6 5 4 3 2 1

springer.com

Preface

The purpose of the second edition of this book is to update the developments in various semiconductor and photonic devices since the first edition was published in 1993. Due to the advances in semiconductor technologies over the past decade, many new semiconductor devices have emerged and entered the marketplace. As a result, a significant portion of the material covered in the original book has been revised and updated. The intent of this book is to provide the reader with a self-contained treatment of the fundamental physics of semiconductor materials and devices. The author has used this book for a one-year graduate course sequence taught for many years in the Department of Electrical and Computer Engineering of the University of Florida. It is intended for first-year graduate students who majored in electrical engineering. However, many students from other disciplines and backgrounds such as chemical engineering, materials science and engineering, and physics have also taken this course sequence. This book may also be used as a general reference for processing and device engineers working in the semiconductor industry.

The present volume covers relevant topics of basic solid-state physics and fundamentals of semiconductor materials and devices and their applications. The main subjects covered include crystal structures, lattice dynamics, semiconductor statistics, one-electron energy band theory, excess carrier phenomena and recombination mechanisms, carrier transport and scattering mechanisms, optical properties, photoelectric effects, metal–semiconductor contacts and devices, p-n junction diodes, bipolar junction transistors (BJTs), heterojunction bipolar transistors (HBTs), MOS devices (MOSFETs, CCDs), photonic devices (solar cells, LEDs, and LDs), quantum-effect devices (QWIPs, QDIPs, QW-LDs), and high-speed III-V semiconductor devices (MESFETs, HEMTs, HETs, RTDs, TEDs). The text presents a unified and balanced treatment of the physics of semiconductor materials and devices. It is intended to provide physicists and materials scientists with more background on device applications, and device engineers with a broader knowledge of fundamental semiconductor physics.

The contents of the book are divided into two parts. In Part I (Chapters 1–9), the subjects of fundamental solid-state and semiconductor physics that are essential for understanding the physical, optical, and electronic properties of semiconductor

materials are presented. Part II (Chapters 10–16) deals with the basic device physics, device structures, operation principles, general characteristics, and applications of various semiconductor and photonic devices.

Chapter 1 presents the classification of solids, crystal structures, concept of reciprocal lattice and Brillouin zone, Miller indices, crystal bindings, and defects in solids. Chapter 2 deals with the thermal properties and lattice dynamics of crystalline solids. The lattice-specific heat, the dispersion relation of lattice vibrations, and the concept of phonons are also described. Chapter 3 is concerned with the three basic semiconductor statistics. Derivation of Maxwell–Boltzmann (M-B), Bose–Einstein (B-E), and Fermi–Dirac (F-D) distribution functions are given in this chapter. Chapter 4 describes the elements of quantum concepts and wave mechanics, the one-electron energy band theory, the effective mass concept for electrons and holes in a semiconductor, the energy band structures for elemental and compound semiconductors, and the density-of-states functions for bulk semiconductors and low-dimensional systems such as superlattices, quantum wells, and dots. Chapter 5 deals with the equilibrium properties of intrinsic and extrinsic semiconductors. Derivation of general expressions for electron and hole densities, and discussion of the shallow- and deep-level impurities in semiconductors are given in this chapter. Chapter 6 presents the recombination mechanisms and excess carrier phenomenon in a semiconductor. The basic semiconductor equations, which govern the transport of excess carriers in a semiconductor, are described in this chapter. Chapter 7 deals with the derivation of transport coefficients using the Boltzmann equation and relaxation time approximation. The low-field galvanomagnetic, thermoelectric, and thermomagnetic effects in n-type semiconductors are described in this chapter. Chapter 8 is concerned with the scattering mechanisms and the derivation of electron mobility in n-type semiconductors. The relaxation time and mobility expressions for the ionized and neutral impurity scatterings and acoustical and optical phonon scatterings are derived. Chapter 9 presents the optical properties and photoelectric effects in semiconductors. The fundamental optical absorption and free-carrier absorption processes as well as the photoelectric effects such as photoconductive, photovoltaic, and photomagnetoelectric effects in a semiconductor are depicted. Chapter 10 deals with the basic theories and relevant electronic properties of metal–semiconductor contacts and their applications. The current conduction in a Schottky barrier diode, methods of determining and enhancing the barrier heights in a Schottky contact, and ohmic contacts in a semiconductor are presented. Chapter 11 presents the basic device theories and characteristics of a p-n junction diode. The p-n heterojunction diodes and junction-field effect transistors (JFETs) are also discussed. Chapter 12 is concerned with the device physics, device structures, and characteristics of various photovoltaic devices (solar cells), photodetectors, and their applications. The solid-state light-emitting devices, which include the light-emitting diodes (LEDs) and semiconductor diode lasers (LDs) are described in Chapter 13. Recent advances in LEDs and LDs and their applications are given in this chapter. Chapter 14 deals with the basic device physics, modeling, and electrical characteristics of bipolar junction transistors (BJTs), p-n-p-n four-layer devices (SCRs, thyristers), and heterojunction bipolar transistors (HBTs).

Chapter 15 presents the silicon-based metal-oxide-semiconductor (MOS) devices. The device physics and characteristics for both metal-oxide-semiconductor field-effect transistors (MOSFETs) and charge-coupled devices (CCDs) are described. Finally, high-speed and high-frequency devices using GaAs and other III-V compound semiconductors are discussed in Chapter 16. The GaAs-based metal–semiconductor field-effect transistors (MESFETs), high-electron-mobility transistors (HEMTs), hot-electron transistors (HETs), resonant tunneling diodes (NTDs) and transferred electron devices (TEDs) are described in this chapter.

Throughout the text, the author stresses the importance of basic semiconductor physics and its relation to the properties and performance of various semiconductor devices. Without a good grasp of the physical concepts and a good understanding of the underlying device physics, it would be difficult to tackle the problems encountered in material growth, device processing and fabrication, device characterization, and modeling. The materials presented in this book should provide a solid foundation for understanding the fundamental limitations of various semiconductor materials and devices. This book is especially useful for those who are interested in strengthening and broadening their basic knowledge of solid-state and semiconductor device physics.

The author would like to acknowledge his wife, “Julie” Wen-Fu Shih, for her support, love, and encouragement during the course of preparing this second edition.

Contents

1
Classification of Solids and Crystal Structure

1.1. Introduction

Classification of solids can be based on atomic arrangement, binding energy, physical and chemical properties, or the geometrical aspects of the crystalline structure. In one class, the atoms in a solid are set in an irregular manner, without any short- or long-range order in their atomic arrangement. This class of solids is commonly known as noncrystalline or amorphous materials. In another class, the atoms or group of atoms in the solid are arranged in a regular order. These solids are referred to as the crystalline solids. The crystalline solids can be further divided into two categories: the single-crystalline and the polycrystalline solids. In a single-crystalline solid, the regular order extends over the entire crystal. In a polycrystalline solid, however, the regular order exists only over a small region of the crystal, with grain size ranging from a few hundred angstroms to a few centimeters. A polycrystalline solid contains many of these small single-crystalline regions surrounded by the grain boundaries. Distinction between these two classes of solids—amorphous and crystalline—can be made through the use of X-ray or electron diffraction techniques.

Classification of solids can also be made according to their electrical conductivity. For example, while the electrical conductivity of an insulator is usually less than $10^{-8}\ \Omega^{-1}$/cm, the electrical conductivity of a metal is on the order of $10^{6}\ \Omega^{-1}$/cm at room temperature. As for a semiconductor, the room-temperature electrical conductivity may vary from 10^{-4} to $10^{4}\ \Omega^{-1}$/cm, depending on the doping impurity density in the semiconductor. Furthermore, the temperature behavior of a semiconductor can be quite different from that of a metal. For example, the electrical conductivity of a metal is nearly independent of the temperature over a wide range of temperatures (except at very low temperatures), while the electrical conductivity of a semiconductor is in general a strong function of the temperature.

In this chapter, we are concerned with the classification of crystalline solids based on their geometrical aspects and binding energies. Section 1.2 presents the seven crystal systems and fourteen Bravais lattices. The crystal structure, the concept of reciprocal lattice and Brillouin zone, and the definition of Miller indices are described in Sections 1.3, 1.4, and 1.5, respectively. Section 1.6 presents the

classification of solids according to the binding energy of their crystalline structure. Of particular interest is the fact that many important physical properties of a solid can be understood, at least qualitatively, in terms of its binding energy. Finally, defects in a semiconductor including vacancies, interstitials, impurities, dislocations, and grain boundaries are described in Section 1.7. It should be mentioned that these defects play an important role in influencing the physical and electrical properties of a semiconductor.

1.2. The Bravais Lattice

In a crystalline solid, the atoms or groups of atoms are arranged in an orderly or periodic pattern. It can be distinguished from all other aggregates of atoms by the three-dimensional (3-D) periodicity of the atomic arrangement. Thus, by properly choosing a small polyhedron as a basic building block, it is possible to construct the entire crystal by repeatedly displacing this basic building block along the three noncoplanar directions of the crystal lattice by translational operation. The suitable geometrical shapes for the basic building blocks are a regular cubic dodecahedron, a truncated octahedron, and any arbitrary parallelepiped. The basic building block of a crystal is called the unit cell. Although a variety of unit cells may be chosen for a particular crystalline structure, there is generally one that is both the most convenient and the most descriptive of the structure. If the shape of the unit cell is specified, and the arrangement of all the atoms within the unit cell is known, then one has a complete geometrical description of the crystal lattice. This is due to the fact that there is only one way in which the unit cells can be stacked to fill the entire space of the crystal.

The various possible arrangements of unit cells in a crystalline solid can be readily achieved by means of the space lattice, a concept introduced by Bravais. The space lattice is an arrangement of lattice points in space such that the placement of points at any given point in space is the same for all points of the space lattice. In general, the periodic translational symmetry of a space lattice may be described in terms of three noncoplanar basis vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ defined in such a way that any lattice point $r(n_1, n_2, n_3)$ can be generated from any other lattice point $r(0, 0, 0)$ in space by the translational operation

$$r(n_1, n_2, n_3) = r(0, 0, 0) + \boldsymbol{R}, \tag{1.1}$$

where

$$\boldsymbol{R} = n_1\boldsymbol{b}_1 + n_2\boldsymbol{b}_2 + n_3\boldsymbol{b}_3 \tag{1.2}$$

is the translational basis vector and n_1, n_2, n_3 are arbitrary integers. A lattice generated by such a translational operation is called the simple Bravais lattice, and the parallelepiped spanned by the three basis vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ is called the unit cell of the Bravais lattice. Figure 1.1 shows a parallelepiped unit cell defined by the length of three basis vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ and three angles α, β, and γ. This is the conventional unit cell for the Bravais lattice.

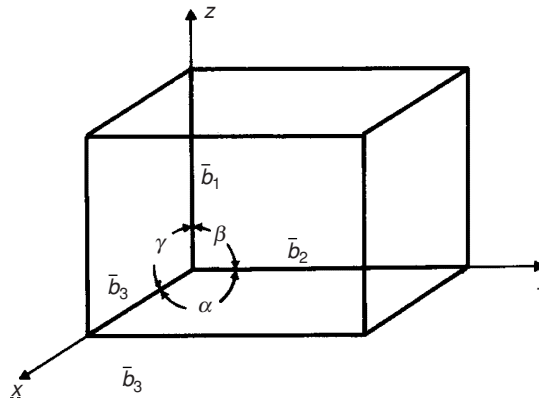


FIGURE 1.1. A parallelepiped unit cell for a Bravais lattice.

Depending on the lengths of the three noncoplanar basis vectors, the angles between them, and the number of lattice points in a unit cell, the space lattice may be divided into seven lattice systems and fourteen Bravais lattices. The seven lattice systems and fourteen Bravais lattices generated by (1.1), are shown in Table 1.1 and Figure 1.2, respectively. The fourteen Bravais lattices, which are all that are known to exist in nature, are constructed using all possible arrangements of lattice points in each unit cell. Each of the Bravais lattices is unique in that it cannot be generated by any of the other thirteen Bravais lattices. Rather, they are generated through any combination of a simple lattice, a base-centered lattice, a face-centered lattice, and a body-centered lattice. All of the Bravais lattices have different symmetry properties. A simple Bravais lattice contains lattice points only at the vertices of a parallelepiped. A base-centered Bravais lattice has lattice points located at the centers of the top and bottom faces as well as at the vertices of the unit cell. In a face-centered Bravais lattice, in addition to the vertex lattice points,

TABLE 1.1. Seven lattice systems and fourteen Bravais lattices.

Lattice systems	Angles and basis vectors	Bravais lattices
Triclinic	$b_1 \neq b_2 \neq b_3$ $\alpha \neq \beta \neq \gamma$	Simple
Monoclinic	$b_1 \neq b_2 \neq b_3$ $\alpha = \beta = 90^\circ \neq \gamma$	Simple, base-centered
Orthorhombic	$b_1 \neq b_2 \neq b_3$ $\alpha = \beta = \gamma = 90^\circ$	Simple, base-centered, body-centered, face-centered
Tetragonal	$b_1 = b_2 \neq b_3$ $\alpha = \beta = \gamma = 90^\circ$	Simple, body-centered
Trigonal	$b_1 = b_2 = b_3$ $\alpha = \beta = \gamma \neq 90^\circ$	Simple
Hexagonal	$b_1 = b_2 \neq b_3$ $\alpha = \beta = 90^\circ, \gamma = 120^\circ$	Simple
Cubic	$b_1 = b_2 = b_3$ $\alpha = \beta = \gamma = 90^\circ$	Simple, body-centered, face-centered

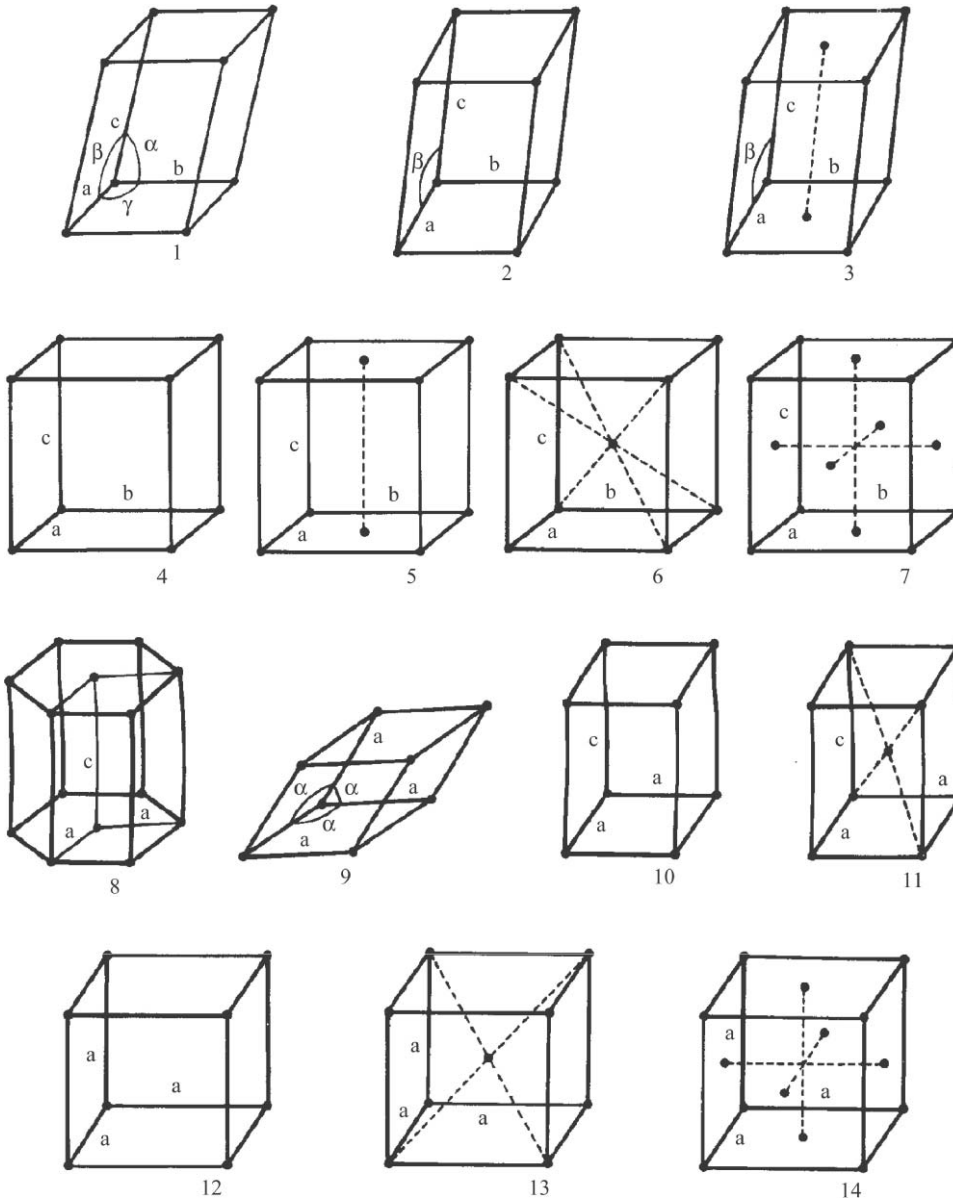
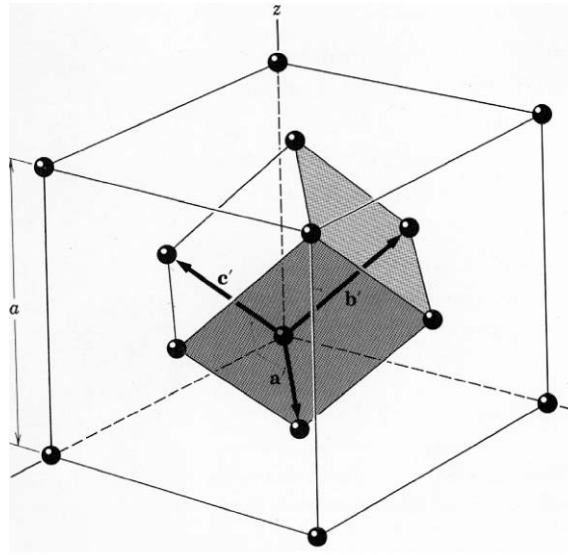


FIGURE 1.2. The fourteen Bravais lattices: (1) triclinic, simple; (2) monoclinic, simple; (3) monoclinic, base-centered; (4) orthorhombic, simple; (5) orthorhombic, base-centered; (6) orthorhomic, body-centered; (7) orthorhombic, face-centered; (8) hexagonal; (9) rhombohedral; (10) tetragonal, simple; (11) tetragonal, body-centered; (12) cubic, simple; (13) cubic, body-centered; (14) cubic, face-centered.

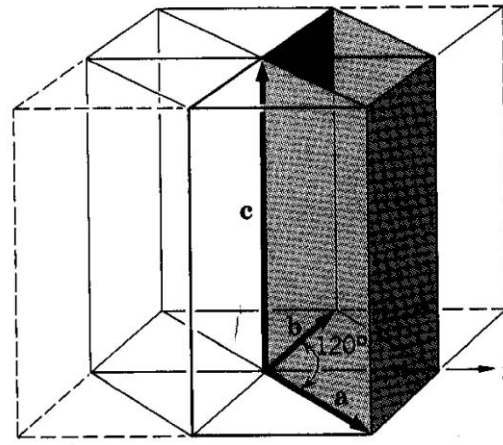
it has lattice points located at the centers of all six faces. Finally, a body-centered Bravais lattice has an extra lattice point located at the volume center of the unit cell. It should be noted that the parallelepiped unit cell shown in Figure 1.1 does not need to be the smallest unit cell for a Braivais lattice. A primitive cell is constructed using three noncoplanar primitive basis vectors with lattice points located only at the vertices of the parallelepiped; it is the smallest unit cell (in volume) in a Bravais lattice. Figure 1.3 shows (a) the primitive cells for a face-centered cubic (FCC) lattice and (b) the relation of the primitive cell in the hexagonal lattice (heavy lines) to a prism of hexagonal symmetry. The rhombohedral primitive cell of the FCC lattice is formed by three primitive translation basis vectors a', b', and c' connecting the lattice point at the origin with lattice points at the three face-center lattice points. The volume of the primitive cell is only one-fourth of the volume of the conventional parallelepiped unit cell for the FCC lattice (see Problem 1.4).

Symmetry is a very important consideration in crystalline solids because many of the physical, electrical, magnetic, elastic, and thermal properties of the solids are strongly dependent on the symmetry properties of their crystal lattice. For example, the electrical conductivity of a cubic crystal is isotropic and independent of its crystalline orientations, while the electrical conductivity of a trigonal crystal can be highly anisotropic along different crystalline axes. The symmetry of a real crystal is determined by the symmetry of its basis and of the Bravais lattice to which the crystal belongs. In addition to the translational symmetry, each Bravais lattice may have different degrees of rotational, reflectional, and inversional symmetry. The rotational symmetry of a crystal lattice is obtained when rotation about a certain crystal axis through an angle of $2\pi/n$ radians leaves the lattice invariant. The lattice is said to have an n-fold rotational axis. Due to the requirements of translational symmetry, the possible values of n are limited to 1, 2, 3, 4, and 6. There are no five- and sevenfold rotational symmetries in a Bravais lattice. Examples of rotation axes can be seen in the (100) axes of a cubic crystal, which has fourfold rotational symmetry, and in the body diagonal (111) axis, which has threefold rotational symmetry. Another type of symmetry, known as reflectional symmetry, is possessed by a crystal lattice when it is invariant under reflection in a plane through the lattice. For example, the six faces of a cubic lattice are the reflection planes for that lattice.

Finally, it is noted that all Bravais lattices possess inversional symmetry. A crystal lattice with inversion symmetry will remain invariant if the lattice point at the coordinate $r = x, y, z$ is replaced by the lattice point at $r = -x, -y, -z$. Although all monatomic crystals have a center of inversion, this type of symmetry is not a general property of crystals. The different types of symmetry that a crystalline solid possesses can be identified through the use of X-ray diffraction techniques. Among the Bravais lattices, the cubic lattice possesses the most symmetry properties, and most semiconductors have the structure of a cubic lattice. Table 1.2 lists some of the important characteristics of cubic lattices.



(a)



(b)

FIGURE 1.3. (a) The primitive cells for a face-centered cubic (FCC) lattice and (b) the relation of the primitive cell in the hexagonal lattice (heavy lines) to a prism of hexagonal symmetry.

1.3. The Crystal Structure

The Bravais lattice discussed in the preceding section is a mathematical abstraction that describes the periodic arrangement of lattice points in space. In general, a real

TABLE 1.2. Characteristics of cubic lattices.

	Simple cubic lattice	Body-centered cubic	Face-centered cubic
Volume, unit cell	a^3	a^3	a^3
Lattice points per cell	1	2	4
Volume, primitive cell	a^3	$a^3/2$	$a^3/4$
Lattice points per unit volume	$1/a^3$	$2/a^3$	$4/a^3$
Number of nearest neighbors	6	8	12
Nearest-neighbor distance	a	$\sqrt{3}a/2$	$a/\sqrt{2}$
Number of second neighbors	12	6	6
Second neighbor distance	$\sqrt{2}a$	a	a

crystal is not a perfect replica of a Bravais lattice, with identical atoms at every lattice point. In fact, there is generally a set of atoms, whose internal symmetry is restricted only by the requirement of translational periodicity, that must be associated with each lattice point of the corresponding Bravais lattice. This set of atoms is known as the basis, and each basis of a particular crystal is identical in composition, arrangement, and orientation. A crystalline structure is formed when a basis of atoms is attached to each lattice point in the Bravais lattice. Figure 1.4 shows the distinction between a space lattice and a crystal structure. Many metals and semiconductors have a simple crystal structure with high degrees of symmetry. For example, alkali metals such as lithium, sodium, and potassium have the face-centered cubic (FCC) structure, while elemental and compound semiconductors have either the diamond, zinc-blende, or wurtzite structure. Figure 1.5 shows

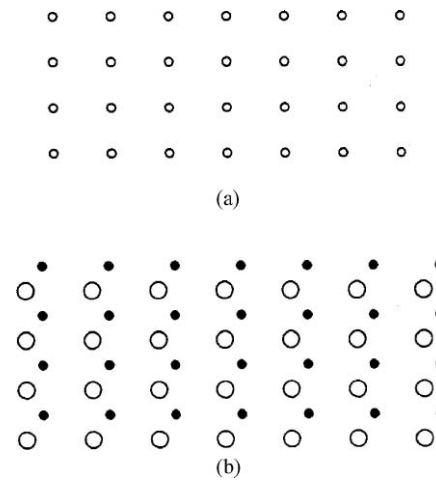


FIGURE 1.4. (a) A two-dimensional (2-D) space lattice, and (b) a 2-D crystal structure with basis of atoms attached to each lattice point.

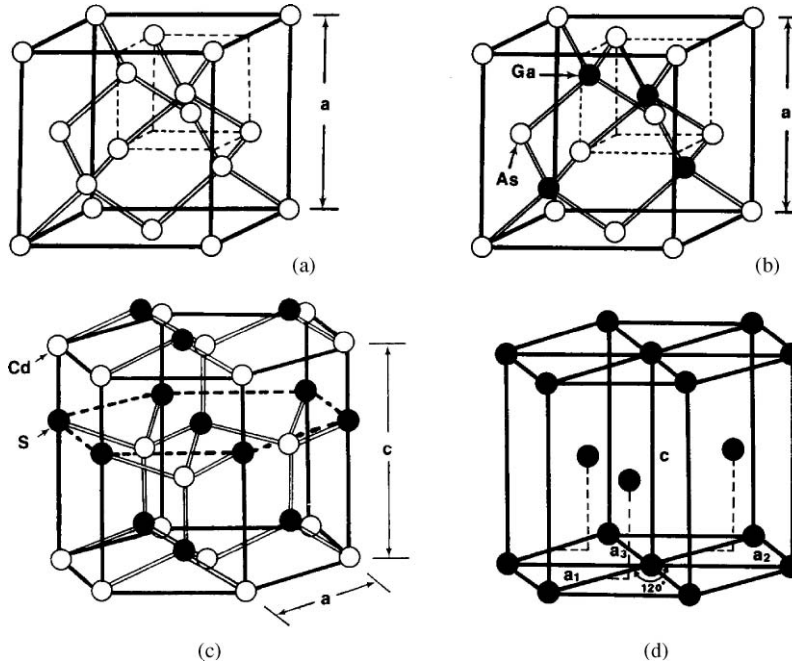


FIGURE 1.5. Four important crystal structures in semiconductors: (a) diamond structure, (b) zinc-blende structure, (c) wurtzite structure, and (d) hexagonal close-packed structure.

the four most commonly observed crystal structures in a semiconductor. The diamond structure shown in Figure 1.5a is actually formed by two interpenetrating face-centered cubic lattices with the vertex atom of one FCC sublattice located at (0, 0, 0) and the vertex atom of another FCC sublattice located at $(a/4\ a/4, a/4)$, where a is the lattice constant. In the diamond lattice structure, the primitive basis of two identical atoms located at (0, 0, 0) and $(a/4, a/4, a/4)$ is associated with each lattice point of the FCC lattice. Elemental semiconductors such as silicon and germanium belong to this crystal structure. The zinc-blende structure shown in Figure 1.5b is similar to the diamond structure except that the two FCC sublattices are occupied alternately by two different kinds of atoms (e.g., Ga and As in a GaAs crystal). III-V compound semiconductors such as GaAs, InP, and InSb have the zinc-blende structure. The wurtzite structure shown in Figure 1.5c is formed by two interpenetrating hexagonal close-packed structures occupied alternately by two different kinds of atoms. II-VI compound semiconductors such as CdS, CdTe, ZnS, and ZnSe have this type of crystal structure. Both the diamond and zinc-blende structures belong to the tetrahedral phase, with each atom surrounded by four equidistant nearest-neighbor atoms at the vertices of a tetrahedron. Figure 1.5d shows a hexagonal close-packed structure. It should be noted that some of the III-V and II-VI compound semiconductors including GaP, ZnS, and

TABLE 1.3. The crystal structures and lattice constants for elemental and compound semiconductors.

Semiconductors	Elements	Lattice structure	Lattice constant (Å)
Elemental semiconductors	Ge	Diamond	5.66
	Si	Diamond	5.43
IV-IV semiconductor	SiC	Zinc blende	4.36
III-V compound semiconductors	GaN	Zinc blende	4.50
		Wurtzite	$a = 3.189, c = 5.185$
	AIN	Wurtzite	$a = 3.11, c = 4.98$
	InN	Wurtzite	$a = 3.54, c = 5.70$
	GaP	Zinc blende	5.45
	GaAs	Zinc blende	5.65
		Wurtzite	$a = 5.18, c = 5.17$
	InP	Zinc blende	5.87
	InAs	Zinc blende	6.06
	InSb	Zinc blende	6.48
II-VI compound semiconductors	CdS	Zinc blende	5.83
		Wurtzite	$a = 4.16, c = 6.75$
	CdSe	Zinc blende	6.05
		Wurtzite	$a = 4.30, c = 7.01$
	CdTe	Zinc blende	6.48
	ZnSe	Zinc blende	5.88
	ZnS	Zinc blende	5.42
		Wurtzite	$a = 3.82, c = 6.26$
IV-VI compound semiconductors	PbS	Cubic	5.93
	PbTe	Cubic	6.46

CdSe may be crystallized either in a zinc-blende or a wurtzite structure. Table 1.3 lists the crystal structures and the lattice constants for elemental and compound semiconductors.

1.4. Miller Indices and Crystal Planes

The orientation of a crystal plane can be determined by three integers, h, k, and l, known as the Miller indices. They are related to the orientations of a crystal plane in the following manner: If h', k', and l' represent the intercepts of a particular crystal plane on the three crystal axes (i.e., x, y, z) in units of the lattice constant a, then the three smallest integers h, k, and l that satisfy the relation

$$hh' = kk' = ll' \tag{1.3}$$

are the Miller indices. As an example, Figure 1.6 illustrates an arbitrary plane that intercepts the three crystal axes at $h' = 2a$, $k' = a$, and $l' = a$, where a is the lattice

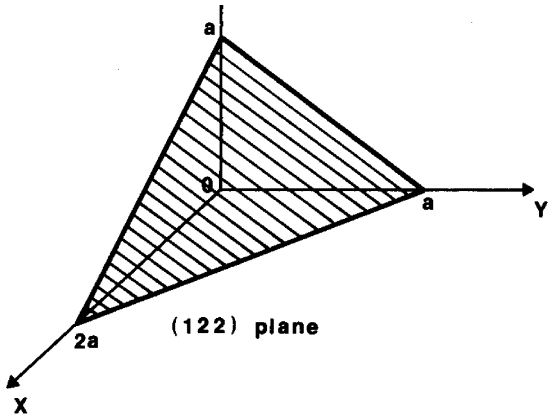


FIGURE 1.6. The Miller indices and the lattice plane.

constant. In this case, the smallest integers that satisfy (1.3) are $h = 1$, $k = 2$, and $l = 2$. These three integers are therefore the Miller indices, and the plane defined by them is called the (122) plane. If a plane is parallel to one of the crystal axes with no interception, then the corresponding Miller index for that axis is zero (i.e., $k' \to \infty$ and $k = 0$). For example, a plane set parallel to the y-z plane and intercepted at the x-axis is called the (100) plane. Furthermore, a set of equivalent planes can be represented collectively by enclosing the Miller indices with in curly braces. For example, the $\{100\}$ planes represent a family of planes consisting of the (100), (010), (001), $(\bar{1}00)$, $(0\bar{1}0)$, and $(00\bar{1})$ planes. The bar on the top of a particular Miller index represents a plane that is intercepted at a negative crystal axis. Figure 1.7 shows the (010), (110), (111), and $(\bar{1}, \bar{1}, \bar{1})$ crystal planes for a simple cubic crystal.

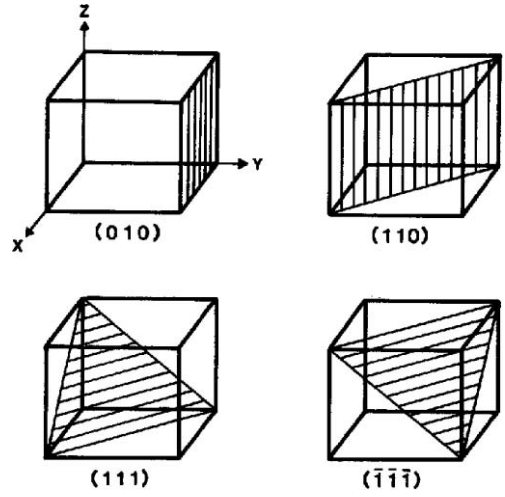


FIGURE 1.7. Some lattice planes in a cubic crystal.

1.5. The Reciprocal Lattice and Brillouin Zone

The Bravais lattice described in Section 1.2 is a space lattice, which has translational symmetry in real space. Since the motion of electrons in a crystal is usually described in both real space and momentum space (or k-space), it is important to introduce here the concepts of reciprocal space and reciprocal lattice. In analogy to a periodic time-varying function, which can be described in terms of the sum of Fourier components in the frequency domain, the spatial properties of a periodic crystal can be described by the sum of the components in Fourier space, or the reciprocal space. For a perfect crystal, the reciprocal lattice in the reciprocal space consists of an infinite periodic three-dimensional (3-D) array of points whose spacing is inversely proportional to the distance between the lattice planes of a Bravais lattice.

The reciprocal lattice is a geometrical construction that allows one to relate the crystal geometry directly to the electronic states and the symmetry properties of a crystal in the reciprocal space. Many important physical, electrical, and optical properties of semiconductors and metals can be understood using the concept of reciprocal lattice. The unit cell of a reciprocal lattice is also known as the Brillouin zone or the Wigner–Seitz cell. The importance of the Brillouin zone in a crystalline solid will become clear when we discuss the lattice dynamics and the energy band theories in Chapters 2 and 4, respectively.

While the basis vector of a direct lattice has the dimension of length, the basis vector of a reciprocal lattice has the dimension of reciprocal length. The translational basis vector of a direct lattice is defined by (1.2). In a reciprocal lattice a set of reciprocal basis vectors $\boldsymbol{b}_1{}^*, \boldsymbol{b}_2{}^*, \boldsymbol{b}_3{}^*$ can be defined in terms of the basis vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ of a direct lattice. This is given by

$$\boldsymbol{b}_1{}^* = \frac{2\pi(\boldsymbol{b}_2 x \boldsymbol{b}_3)}{|\boldsymbol{b}_1 \cdot \boldsymbol{b}_2 x \boldsymbol{b}_3|}, \quad \boldsymbol{b}_2{}^* = \frac{2\pi(\boldsymbol{b}_3 x \boldsymbol{b}_1)}{|\boldsymbol{b}_1 \cdot \boldsymbol{b}_2 x \boldsymbol{b}_3|}, \quad \boldsymbol{b}_3{}^* = \frac{2\pi(\boldsymbol{b}_1 x \boldsymbol{b}_2)}{|\boldsymbol{b}_1 \cdot \boldsymbol{b}_2 x \boldsymbol{b}_3|}. \tag{1.4}$$

The reciprocal lattice vector can be defined in terms of the reciprocal basis vectors and Miller indices by

$$\boldsymbol{K} = h\boldsymbol{b}_1{}^* + k\boldsymbol{b}_2{}^* + l\boldsymbol{b}_3{}^*, \tag{1.5}$$

where $\boldsymbol{b}_1{}^*, \boldsymbol{b}_2{}^*$, and $\boldsymbol{b}_3{}^*$ are given by (1.4), and h, k, l are the Miller indices. The reciprocal lattice vector defined by (1.5) may be used to generate all the reciprocal lattice points in the entire reciprocal space with its unit cell spanned by the reciprocal basis vectors defined by (1.4). Some important properties of the reciprocal lattice are summarized here:

1. Each reciprocal lattice vector in the reciprocal lattice is perpendicular to a set of lattice planes in the direct lattice, as illustrated in Figure 1.8. Using (1.2) and (1.5) one obtains

$$\boldsymbol{R} \cdot \boldsymbol{K} = 2\pi(n_1 h + n_2 k + n_3 l) = 2\pi N, \tag{1.6}$$

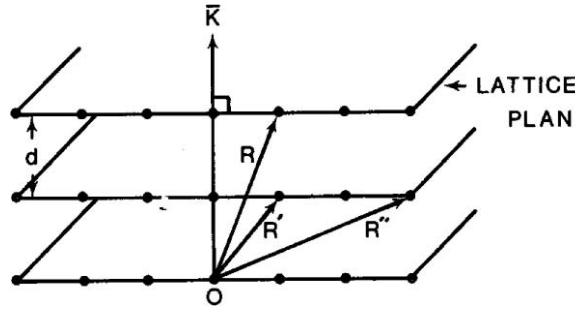


FIGURE 1.8. The reciprocal lattice vector and the lattice planes of a direct lattice, where $\boldsymbol{K}$ denotes the reciprocal lattice vector and $\boldsymbol{R}$ is the translational lattice vector; d is the distance between the lattice planes.

or

$$\exp(\mathrm{i}\boldsymbol{K} \cdot \boldsymbol{R}) = 1, \tag{1.7}$$

where N, n_1, n_2, and n_3 are integers. Equation (1.6) shows that the projection of the translational vector $\boldsymbol{R}$ in the direction of $\boldsymbol{K}$ has length given by

$$d_{hkl} = 2\pi N/|\boldsymbol{K}|, \tag{1.8}$$

where d_{hkl} is the spacing between the two nearby planes of a direct lattice, as shown in Figure 1.8. Equation (1.7) defines the reciprocal lattice in the reciprocal space.

2. The volume of a unit cell in the reciprocal lattice is inversely proportional to the volume of a unit cell in the direct lattice. The denominator of (1.4) represents the volume of the unit cell of a direct lattice, which is given by

$$V_{\mathrm{d}} = |\boldsymbol{b}_1 \cdot \boldsymbol{b}_2 x \boldsymbol{b}_3| \,. \tag{1.9}$$

The volume of the unit cell of a reciprocal lattice is defined by the three reciprocal basis vectors and is given by

$$V_r = \left|\boldsymbol{b}_1{}^* \cdot \boldsymbol{b}_2{}^* x \boldsymbol{b}_3{}^*\right| = \frac{8\pi^3}{V_{\mathrm{d}}} \,. \tag{1.10}$$

The factor $8\pi^3$ given in (1.10) is included so that the reciprocal lattice is defined in such a way that the dimension of the reciprocal lattice vector is the same as the wave vector of phonons or electrons in the momentum (k-) space, as will be discussed further in Chapters 2 and 4.

3. A direct lattice is the reciprocal of its own reciprocal lattice; this can be shown using (1.10).
4. The unit cell of a reciprocal lattice need not be a parallelepiped. In fact, one always deals with the Wigner–Seitz cell of the reciprocal lattice, which is also known as the first Brillouin zone in the reciprocal space, as shown in Figure 1.9.

Construction of the first Brillouin zone in the reciprocal lattice will be discussed next. The first Brillouin zone is the unit cell of the reciprocal lattice. It is the basic

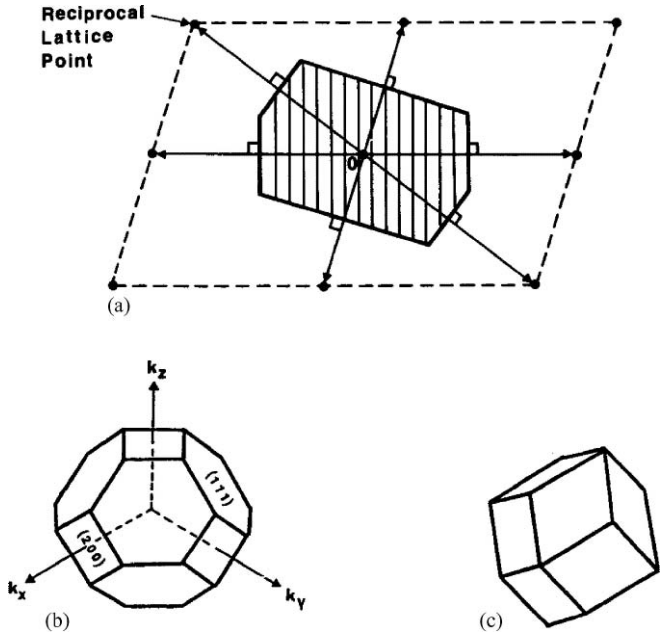


FIGURE 1.9. Construction of the first Brillouin zone for (a) a two-dimensional reciprocal lattice, (b) a face-centered cubic lattice, and (c) a body-centered cubic lattice.

building block, having the smallest volume in the reciprocal space, centered at one reciprocal lattice point, and bounded by a set of planes that bisect the reciprocal lattice vectors connecting this reciprocal lattice point to all its neighboring reciprocal lattice points. As an example, Figure 1.9a shows the construction of the first Brillouin zone for a two-dimensional reciprocal lattice. It is obtained by first drawing a number of reciprocal lattice vectors from the center reciprocal lattice point, say (0, 0), to all its nearest-neighboring reciprocal lattice points, and then drawing the bisecting lines perpendicular to each of these reciprocal lattice vectors. The smallest area enclosed by these bisecting lines is called the first Brillouin zone, or the unit cell of this 2-D reciprocal lattice. The first Brillouin zone for a three-dimensional crystal lattice can be constructed in a similar way to that of a 2-D reciprocal lattice described above. This is done by first drawing the reciprocal lattice vectors from a chosen reciprocal lattice point to all its nearest-neighboring reciprocal lattice points, and then drawing the bisecting planes perpendicular to each of these reciprocal lattice vectors. The smallest volume enclosed by these bisecting planes will normally form a polyhedron about the central reciprocal lattice point, and this polyhedron is called the first Brillouin zone or the Wigner–Seitz cell of the reciprocal lattice. Figure 1.9b, c shows the first Brillouin zones for a face-centered cubic lattice and a body-centered cubic lattice, respectively. It is noted that the first Brillouin zone for a diamond lattice and a zinc-blende

lattice structure is identical to that of the face-centered cubic lattice shown in Figure 1.9b.

The importance of the first Brillouin zone can be best illustrated by considering the wave function of an electron wave packet in a crystalline solid, which is described by the wave vector k in momentum (or the reciprocal lattice) space. In a periodic crystal, it can be shown using translational operation that for any given wave vector k' of the electron wave packet in the reciprocal space, there is a corresponding wave vector k inside the first Brillouin zone, which is related to k' by

$$\boldsymbol{k}' = \boldsymbol{k} + \boldsymbol{K}, \tag{1.11}$$

where $\boldsymbol{K}$ is the reciprocal lattice vector defined by (1.5). Therefore, for a given reciprocal lattice point in the reciprocal space, there is a corresponding reciprocal lattice point in the first Brillouin zone, which can be obtained through the translational operation by substituting (1.7) into (1.11). In fact, one can show that except for a phase factor difference, the wave function of an electron at any given reciprocal lattice point in the reciprocal space is identical to the wave function of a corresponding reciprocal lattice point in the first Brillouin zone obtained via the translational operation of (1.7). This is important since it allows one to describe the entire physical properties of electrons or phonons in the first Brillouin zone of the reciprocal space using the reduced zone scheme. In fact, the phonon dispersion relation and the electronic states (or the energy bands) in a solid can be described using the concept of reciprocal lattice and the first Brillouin zone described in this section. This will be discussed further in Chapters 2 and 4.

1.6. Types of Crystal Bindings

In Section 1.2, we described the classification of solids based on the geometrical aspects of the crystal lattice. In this section, we present the classification of solids according to their binding energy (i.e., the energy responsible for holding the atoms of a solid together). Based on the types of chemical binding energy, we can divide the crystalline solids into four categories. These are discussed next.

(i) Ionic crystals. In an ionic crystal, the electrostatic bonding normally comes from the transfer of electrons from alkali atoms to halogen atoms, resulting in the bonding of positively and negatively charged ions by the Coulomb attractive force. Typical examples are alkali metals such as sodium and potassium, in which each of these atoms has one extra valence electron to transfer to the atoms of halogens such as chlorine and bromine to form an alkali halide salt (e.g., NaCl, KCl, NaBr). The II-VI (e.g., CdS, ZnSe, and CdTe) and III-V (e.g., GaAs, InP, and InSb) compound semiconductors also show certain ionic crystal properties. The

ionic crystal usually has high binding energy due to the strong Coulombic force between the positive and negative ions. Ionic crystals formed by the group I and group VII elements (e.g., NaCl, KCl) in the periodic table belong to this category. Although they are good electrical insulators at room temperature due to their large binding energy, these ions may become mobile at very high temperatures and diffuse through the crystal, which results in an increase of electrical conductivity. The electrical conductivity of an ionic crystal is usually many orders of magnitude smaller than the electrical conductivity of a metal, since the mass of the ion is about 10^4 times larger than the electron mass in a metal. The conductivity of an ionic crystal at elevated temperatures is related to the diffusion constant D of the mobile ion by

$$\sigma_i = \frac{Nq^2D}{k_B T}, \tag{1.12}$$

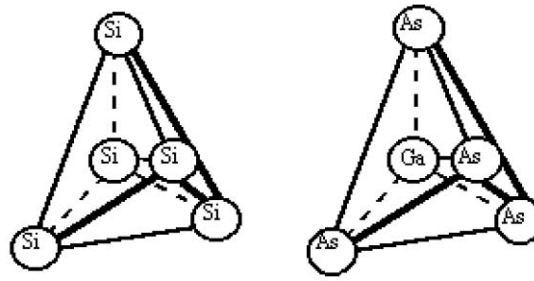
where σ_i is the electrical conductivity of the ionic crystal, N is the density of mobile ions, q is the electronic charge, and k_B is the Boltzmann constant.

One important feature of alkali-halide crystals is that they are transparent to visible and infrared (IR) optical radiation, and hence are widely used as optical window materials in the visible to IR spectral range. For example, crystalline Nacl which is transparent to optical radiation from 0.4 to 16 μm, is widely used as the prism material for grating monochromators in this spectral range.

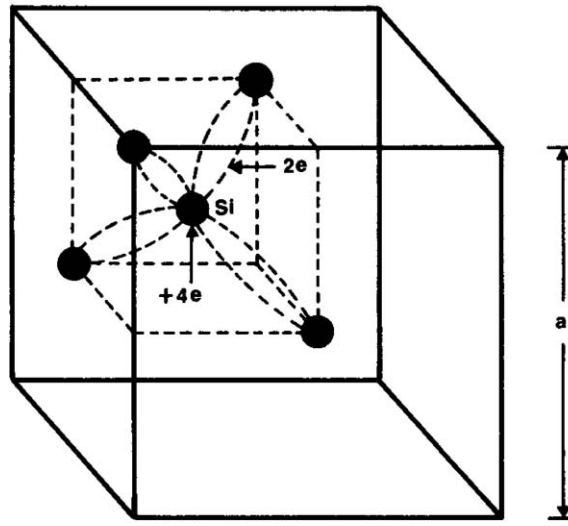
(ii) Covalent crystals. In a covalent crystal, the binding energy comes from the reciprocal sharing of valence electrons of the nearest-neighboring atoms rather than from the transfer of valence electrons as in the case of ionic crystals. Elemental semiconductors such as silicon and germanium are typical covalent crystals.

The structure of a covalent crystal depends strongly on the nature of bonding itself. Covalent crystals such as germanium, silicon, and carbon have four valence electrons per atom, which are shared reciprocally with the nearest-neighboring atoms, contributing to the bonding of the crystal. Figure 1.10a illustrates the tetrahedral bonding for silicon and GaAs crystals, and Figure1.10b the charge distribution of a silicon crystal. Each silicon atom has four valence electrons, which are shared reciprocally by its neighboring atoms and form a tetrahedral bonding. The diamond structure of a silicon crystal is a structure in which each atom is at the center of a tetrahedron, symmetrically surrounded by the four nearest-neighbor atoms located at the vertices. The tetrahedral bonding shown in Figure 1.10a for silicon crystalline can be explained by a linear combination of the s- and p-like atomic orbitals, called the sp^3 hybrids.

High-purity covalent crystals can have very high electrical resistivity and behave like insulators at room temperature. However, the binding force holding the valence electrons in orbit is not as strong as that of an ionic crystal. For example, while the energy required to break an ionic bond in most ionic crystals may be as high as



(a)



(b)

FIGURE 1.10. (a) The tetrahedral bonding configuration of silicon and GaAs crystals, and (b) charge distribution in a diamond lattice structure, showing the tetrahedral covalent bond of a silicon lattice.

10 eV, the energy necessary to break a covalent bond is much smaller, having values ranging from 0.1 eV to around 6.2 eV. Therefore, at room temperature or higher, the thermal energy may be sufficient to break the covalent bonds, thus freeing the valence electrons for electrical conduction in a covalent crystal. Furthermore, the broken bonds left behind by the valence electrons may be treated as free holes in the covalent crystal, which in turn can also contribute to the electrical conductivity in the valence band. In fact, both electrons and holes can contribute to the electrical conduction in an intrinsic semiconductor, as will be discussed further in Chapter 5.

The fact that III-V compound semiconductors such as GaAs, InP, and InAs crystallize in a zinc-blende structure implies that covalent bonding occurs in these crystals. However, in order to form covalent (homopolar) bonding in III-V semiconductors, the sp^3 orbital surrounding each group-III and group-V atom requires four valence electrons per atom. This means that a transfer of one electronic charge from the group-V atom to the group-III atom will occur in III-V compound semiconductors. This will result in group-III atoms becoming negatively charged (III^{1-}) and group-V atoms becoming positively charged (V^{1+}). A negatively charged (III^{1-}) atom together with positively charged (V^{1+}) atom constitutes a nonneutral situation involving Coulombic interaction and hence ionic bonding. This partial ionic bonding characteristic is responsible for some striking differences between the III-V compound semiconductors and elemental semiconductors such as silicon and germanium.

(iii) Metallic crystals. One of the most striking features of a metal is its high electrical conductivity. The binding energy of a metal comes mainly from the average kinetic energy of its valence electrons, and there is no tendency for these electrons to be localized within any given portion of the metal. For example, in a monatomic metal such as sodium or potassium, there are some 10^{23} cm^{-3} valence electrons that can participate in the electrical conduction in these metals. In the classical theory of metals, valence electrons are treated as free electrons, which can move freely inside the metal. The valence electrons form an electron sea in which the positive ions are embedded. Typical examples are the 2s electrons in a lithium crystal and the 3s electrons in a sodium crystal, which are responsible for the binding force of these metallic crystals.

In general, the binding energy of a monatomic metal is mainly due to the average kinetic energy of the valence electrons, which is usually much smaller than that of the ionic and covalent crystals. However, for transition metals, the binding energy, which is due to the covalent bonds of the d-shell electrons, can be much higher than that of monatomic metals.

(iv) Molecular crystals. Argon, neon, and helium are solids that exhibit properties of molecular binding. These substances generally have a very small binding energy, and consequently have low melting and boiling temperatures. The binding force that holds the saturated molecules together in solid phase comes primarily from the van der Waals force. This force is found to vary as r^{-6}, where r is the distance between the two molecules. To explain the origin of this force, it is noted that molecules in such a substance carry neither net electric charge nor permanent electric dipole moment. The instantaneous dipole moment on one molecule will give rise to an electric field, which induces dipole moments on the neighboring molecules. It is the interactions of these instantaneous dipole moments that produce the cohesive energy of a molecular crystal. Since the individual molecules of a molecular crystal are electrically neutral and interact only weakly with one another, they are good electrical insulators, showing neither electronic nor ionic conductivity.

1.7. Defects in a Crystalline Solid

It is generally known that a perfect crystal lattice is possible only mathematically, and in fact does not exist in real crystals. Defects or imperfections are found in all crystalline solids. The existence of defects usually has a profound effect on the physical properties of a crystal, which is particularly true for semiconductor materials. Therefore, it is important to discuss various types of defects that are commonly observed in a crystalline solid.

In general, defects may be divided into two broad categories: One class of defects, which is called dynamic defects, refers to phonons, electrons, and holes. Another class of defects, which is known as stationary defects, is composed of point defects (e.g., vacancies, interstitials, antisite defects, and foreign impurities), line defects (e.g., dislocations), and surface defects (e.g., grain boundaries). Stationary defects play a key role in affecting the electronic, optical, and physical properties of semiconductors. The physical properties and formation of these stationary defects are discussed next.

1.7.1. Vacancies and Interstitials

Both vacancies and interstitials are defects of atomic dimensions; they can be observed only through the use of modern field-ion microscopy or infrared microscopy techniques. Vacancies are always present in crystals, and the density of vacancies depends strongly on temperature. In fact, temperature fluctuation can cause a constant creation and annihilation of vacancies in a crystal. Figure 1.11 shows the formation of vacancy, interstitial, and foreign impurity defects in a crystal lattice.

A vacancy is created when an atom migrates out of its regular lattice site to an interstitial position or to the surface of the crystal. The energy required to remove an atom from its regular lattice site is defined as the activation energy of the vacancy. Two types of defects are usually associated with the creation of vacancies, namely, the Frenkel and Schottky defects. A Frenkel defect is created when an atom is moved from its regular lattice site to an interstitial site, while a Schottky defect is formed when the atom is moved from its regular lattice site to the surface of the crystal. Figure 1.12 shows both the Frenkel and Schottky defects. Another type of point defect, which is commonly found in a semiconductor, is created by

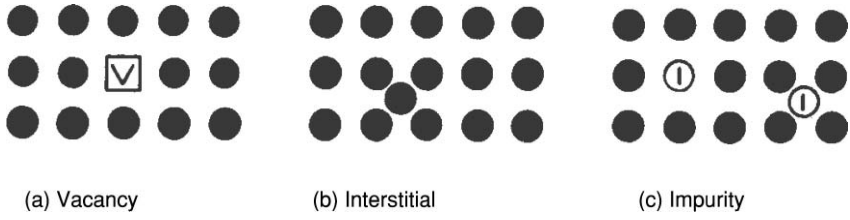


FIGURE 1.11. Vacancy, interstitial, and impurity point defects in a crystal lattice.

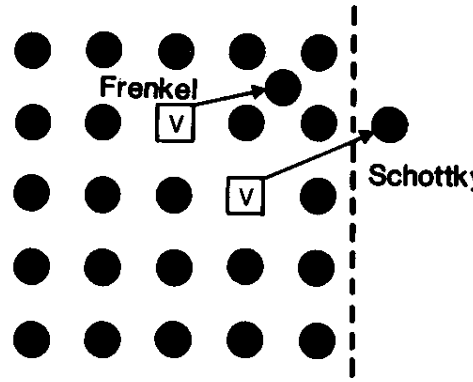


FIGURE 1.12. Formation of Frenkel and Schottky defects in a crystal lattice.

the introduction of foreign impurities into the crystal, either intentionally (e.g., by thermal diffusion or ion implantation), or unintentionally (due to metallic or chemical contaminations), as shown in Figure 1.11c. The Frenkel defects can be created by radiation damages such as high-energy (1 MeV) electron radiation and low-energy proton radiation in semiconductor devices.

The density of vacancies in a crystal can be calculated using classical statistics and thermodynamic principles. In thermal equilibrium, the entropy of a crystal is increased by the presence of disorders, and thus a certain number of vacancies are always present in the crystal. According to the principles of thermodynamics, the equilibrium condition of a system at a finite temperature is established when the free energy of the system is at a minimum. If there are n vacancies distributed randomly among N lattice sites, then the increase of entropy and free energy in the crystal can be calculated as follows: If E_v is the activation energy of a vacancy, then the total incremental internal energy U of the crystal due to the creation of n vacancies is equal to nE_v, where n is the number of vacancies at temperature T. The total number of ways of arranging n vacancies among N lattice sites is given by

$$P = \frac{N!}{(N-n)!n!}. \tag{1.13}$$

The increase of entropy due to the creation of n vacancies in a crystal can be expressed by

$$S = k_B \ln(P) = k_B \ln[N!(N-n)!n!], \tag{1.14}$$

where S is the entropy, and k_B is the Boltzmann constant. Thus, the total change in the free energy of the system is given by

$$F = U - TS = nE_v - k_B T \ln[N!/(N-n)!n!]. \tag{1.15}$$

In thermal equilibrium, the incremental free energy F must be at its minimum with respect to n. The factorials given in (1.15) can be simplified using Stirling's

approximation when n is very large (i.e., $\ln n! \approx n \ln n - n$, for $n \gg 1$). Thus, for $n \gg 1$, (1.14) can be simplified to

$$S \approx k_B[N \ln N - (N - n) \ln(N - n) - n \ln n]. \tag{1.16}$$

Using (1.15) and (1.16), the minimum free energy can be obtained by differentiating F with respect to n in (1.15) and setting the result equal to zero (i.e., $\partial F/\partial n = 0$), which yields

$$n = (N - n) \exp(-E_v/k_B T), \tag{1.17}$$

or

$$n \approx N \exp(-E_v/k_B T) \qquad \text{for } N \gg n. \tag{1.18}$$

Equation (1.18) shows that the density of vacancies increases exponentially with temperature. For example, assuming $E_v = 1$ eV and $N = 10^{23}$ cm^{-3}, the density of vacancies n at $T = 1200$ K is equal to 4.5×10^{18} cm^{-3}.

A similar procedure may be employed to derive the expressions for the density of Frenkel and Schottky defects in a crystal. For Schottky defects, this is given by

$$n \approx N \exp(-E_s/k_B T), \tag{1.19}$$

where E_s is the activation energy for creating a Schottky defect. For Frenkel defects, we obtain

$$n \approx (NN')^{1/2} \exp(-E_f/2k_B T), \tag{1.20}$$

where E_f is the activation energy of a Frenkel defect. Note that N' is the density of interstitial sites. In general, it is found that $E_s > E_f > E_v$. For example, for aluminum, E_v has been found equal to 0.75 eV and $E_s \approx 3$ eV.

It is interesting to note that Frenkel defects may be created by nuclear bombardment, high-energy electron and proton irradiation, or ion implantation damage. In fact, the radiaton damage created by high-energy particle bombardment in a solid is concerned almost entirely with the creation and annihilation of Frenkel defects. The Frenkel defects may be eliminated or reduced via thermal or laser annealing processes. Both thermal and laser annealing procedures are widely used in semiconductor material processing and device fabrication. Recently, rapid thermal annealing (RTA) and laser annealing techniques have also been used extensively in the semiconductor industry for removing damages created by radiation, ion implantation, and device processing.

Foreign impurities constitute another type of point defect, one that deserves special mention. Both substitutional and interstitial impurity defects may be introduced by doping the host crystal with foreign impurities using thermal diffusion or ion implantation. It is a common practice to use foreign impurities to modify the electrical conductivity and the conductivity types (i.e., n- or p-type) of a semiconductor. Foreign impurities may either occupy a regular lattice site or reside in an interstitial site of the host crystal. As will be discussed in Chapter 5, both

shallow- and deep-level impurities may play a very important role in controlling the physical and electrical properties of a semiconductor. Finally, point defects may also be created by quenching the crystal at high temperatures or by severe deformation of the crystal through hammering or rollling.

1.7.2. Line and Surface Defects

Another type of crystal defect, known as line defects, may be created in both single- and polycrystalline solids. The most common type of line defect created in a crystalline solid is called dislocation. Dislocations are lattice defects created in a crystal that can best be described in terms of partial internal slip. There are two types of dislocations that are commonly observed in a crystalline solid. They are the edge and screw dislocations. The creation of these dislocations and their physical properties are discussed next.

(i) Edge dislocation. An edge dislocation can best be described by imagining a perfect crystal that is cut open along line AO, shown in Figure 1.13a; the plane of the cut is perpendicular to that of the page. An extra monolayer crystal plane of depth AO is then inserted in the cut and the crystal lattice is repaired as well as possible, leaving a line perpendicular to the plane of the paper and passing through the point O around which the crystal structure is severely distorted. The distortion of a crystal lattice can be created by the partial insertion of an extra plane of atoms into the crystal. This distortion is characterized by a line defect. The local expansion (known as the dilatation) around the edge dislocation can be described by a simple expression, which reads

$$d = \left(\frac{b}{r}\right) \sin\theta, \tag{1.21}$$

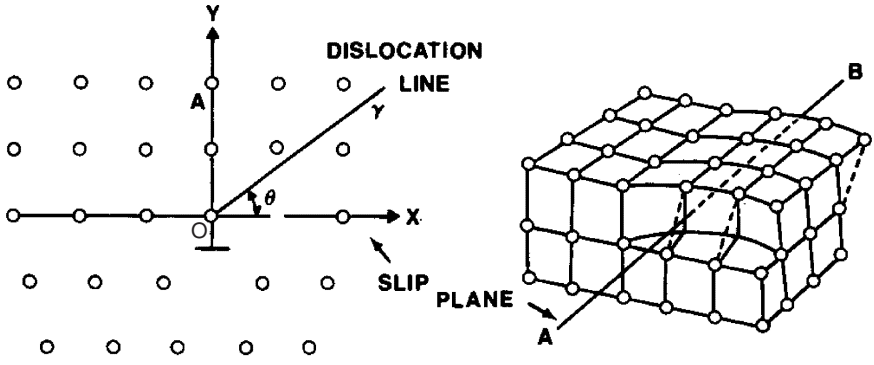


FIGURE 1.13. (a) Edge dislocation and (b) screw dislocation in a crystal lattice.

where b is the Burgers vector (a measure of the strength of distortion caused by dislocation), r is the radial distance from a point in the crystal to the dislocation line, and θ is the angle between r and the slip plane. The sign of dilatation is positive for expansion and negative for compression. The Burgers vector for an edge dislocation is perpendicular to the dislocation line and lies in the slip plane.

(ii) Screw dislocation. The second type of dislocation, known as screw dislocation, is shown in Figure 1.13b. As shown in the horizontal plane of the figure, the screw dislocation is produced by cutting the crystal partially through and pushing the upper part of the crystal one lattice spacing over. A line of distortion is clearly shown along the edge of the cut. This line is usually called the screw dislocation. In contrast to the type of distortion surrounding an edge dislocation, the atoms near the center of a screw dislocation are not in dilatation, but are on a twisted or sheared lattice. It is noted that in a screw dislocation the relative displacements of the two halves of the crystal lie in the direction of the dislocation line rather than normal to it. Again, the Burgers vector can be used to specify the amount of displacement that has occurred.

Dislocations may be created in a number of ways. For example, a plastic deformation creates dislocation and consequently creates damage in the lattice. The dislocations themselves introduce defect levels in the forbidden gap of a semiconductor. For semiconductors with a diamond lattice structure the dislocation velocity depends exponentially on the temperature, and hence the dislocation generation requires that the plastic deformation take place in a semiconductor at very high temperatures. The density of dislocations, which is defined by the number of dislocation lines intersected by a plane of unit area, can be counted using either the etch-pit or the X-ray diffraction technique. In the etch-pit technique, the sample is first polished and then chemically etched. Conical pits are formed at places where dislocation lines intersect the crystal surface, and the number of etch pits is counted. In the X-ray diffraction technique, the observed spread of angle θ in the Bragg diffraction pattern is a measure of the dislocation density.

If the dislocation density is sufficiently high (e.g., $N_D > 10^7$ cm^{-2}), then the electrical and mechanical properties of a crystalline solid may be affected by the presence of these dislocations. For example, the electrical conductivity of a semiconductor measured parallel and perpendicular to the dislocation line can vary considerably when the density of these dislocation lines is very high. It is noted that a dislocation line may be considered as a line charge, which can trap the minority carriers and affect the minority carrier lifetime in a semiconductor. In a pure silicon or germanium crystal, the dislocation density may range from a few hundred to several tens of thousands per cm^2, depending on the conditions of crystal growth and heat treatment. In general, if the dislocation density is less than 10^6 cm^{-2}, then its effect on the electrical properties of a semiconductor becomes negligible. Semiconductors can be produced with zero or few dislocations per

unit area. In fact, dislocation-free germanium and silicon single crystals have been routinely grown using current crystal pull technology. However, for polycrystalline materials, the dislocation density is usually very high, and thus dislocations play a much more important role in a polycrystalline semiconductor than in a single-crystalline semiconductor. Its effects on the minority carrier lifetimes and the majority carrier mobility of a polycrystalline material are also more pronounced than for a single-crystalline semiconductor.

The surface defect is another type of defect that can affect the performance of a semiconductor device. A typical example of a surface defect is the grain boundaries in a polycrystalline semiconductor. In general, an array of edge dislocations can be formed near the grain boundaries of any two subregions of a polycrystalline material. Grain boundaries often play an important role in influencing the electrical and transport properties of a polycrystalline semiconductor. For example, depending on the heat treatment used during and after the film growth, the grain size of polycrystalline silicon thin films grown by the low-pressure chemical vapor deposition (LPCVD) technique may vary from a few hundred angstroms to a few tens of micrometers. On the other hand, for a bulk polycrystalline silicon material, the grain size may vary from a few millimeters to a few centimeters. Polycrystalline silicon thin films prepared by the LPCVD technique are widely used for interconnects and thin-film resistors in silicon integrated circuits. Bulk multicrystalline silicon and polycrystalline thin-film materials from II-VI and I-III-VI compound semiconductors such as CdTe and $CuInSe_2$ (CIS) are widely used in fabricating low-cost solar cells for terrestrial power generation.

Problems

1.1. Show that the maximum proportion of the available volume that can be filled by hard spheres for the following lattice structures is given by
   (a) Simple cubic: $\pi/6$.
   (b) Body-centered cubic: $\sqrt{3}\pi/8$.
   (c) Face-centered cubic: $\sqrt{2}\pi/6$.
   (d) Hexagonal-closed-packed: $\sqrt{2}\pi/6$.
   (e) Diamond: $\sqrt{3}\pi/16$.

1.2. Explain why the following listed lattices are not Bravais lattices.
   (a) Base-centered tetragonal.
   (b) Face-centered tetragonal.
   (c) Face-centered rhombohedral.

1.3. Show that a crystal lattice cannot have an axis with fivefold and sevenfold rotational symmetry.

1.4. Construct a primitive cell for a body-centered cubic (BCC) and a face-centered cubic (FCC) lattice, and write down the primitive basis vectors and the volume of these two primitive cells.

1.5. (a) Show that a diamond lattice structure is made up of two interpenetrating face-centered cubic lattices.

(b) If the cubic edge (or the lattice constant) of a diamond lattice is equal to 3.56 Å, calculate the distance between the nearest neighbors and the total number of atoms per unit cell.
(c) Repeat for silicon (cubic edge $a = 5.43$ Å) and gemanium (cubic edge $a = 5.62$ Å).

1.6. Show that a body-centered tetragonal lattice with $a = \sqrt{2}b$ has the symmetry of a face-centered cubic lattice.
1.7. Find the number of nearest neighbors and the primitive lattice vectors for a diamond lattice structure.
1.8. Show that the reciprocal lattice of a body-centered cubic lattice is a face-centered cubic lattice.
1.9. Draw the crystal planes for the following lattice structures:
(a) (200), (222), (311) planes for a cubic crystal.
(b) $(10\bar{1}0)$ plane of a hexagonal crystal. (*Hint*: the Miller indices for a hexagonal lattice are represented by (a_1, a_2, a_3, c).)
1.10. Show that the first Brillouin zone of a diamond lattice structure is enclosed by eight $\{111\}$ and six $\{200\}$ planes.
1.11. (a) Find the total number of planes for $\{100\}$, $\{110\}$, $\{111\}$, and $\{200\}$ of a cubic lattice.
(b) Find the normal distance from the origin of the unit cell to the planes listed in (a).
1.12. Show that in a hexagonal close-packed lattice structure, the length of the c-axis is equal to $\sqrt{8/3}a$, where a is the length of one side of the hexagonal base plane.
1.13. Draw the first four Brillouin zones of a two-dimensional square lattice, and show that the areas of all the zones are identical.
1.14. Show that the density of the Schottky and Frenkel defects in a crystal are given, respectively, by (1.19) and (1.20).

Bibliography

F. J. Blatt, *Physics Propagation in Periodic Structures*, 2nd ed., Dover, New York (1953).
L. Brillouin, *Wave Propagation in Periodic Structures*, 2nd ed., Dover, New York (1953).
M. J. Buerger, *Elementary Crystallography*, Wiley, New York (1963).
A. J. Dekker, *Solid State Physics*, 6th ed., Prentice-Hall, Englewood Cliffs (1962).
B. Henderson, *Defects in Crystalline Solids*, Crane Russak & Co., New York (1972).
C. Kittel, *Introduction to Solid State Physics*, 5th ed., Wiley, New York (1976).
T. L. Martin, Jr. and W. F. Leonard, *Electrons and Crystals*, Brooks/Cole, California (1970).
J. P. McKelvey, *Solid State and Semiconductor Physics*, Harper & Row, New York (1966).
A. G. Milnes, *Deep Impurities in Semiconductors*, Wiley Interscience, New York (1973).
L. Pauling, *The Nature of the Chemical Bond*, Cornell University Press, Ithaca, New York (1960).
J. C. Phillips, *Bonds and Bands in Semiconductors*, Academic Press, New York (1973).
F. C. Phillips, *An Introduction to Crystallography*, Longmans, Green & Co., London (1946).

M. Shur, M. Levinshtein, S. Rumyantsev (eds.), *Properties of Advanced Semiconductor Materials: GaN, AIN, InN, BN, SiC, and SiGe*. Wiley Interscience, New York (2001).

J. C. Slater, *Quantum Theory of Molecules and Solids*, vol. 2, McGraw-Hill, New York (1965).

R. W. G. Wyckoff, *Crystal Structures*, 2nd ed., Interscience Publishers, New York (1963).

2
Lattice Dynamics

2.1. Introduction

This chapter presents the thermal properties and lattice dynamics of solids. In thermal equilibrium, the mass centers or the nuclei of the atoms in a solid are not at rest, but instead they vibrate with respect to their equilibrium positions. In fact, many thermal properties of solids are determined by the amplitude and phase factor of the atomic vibrations. For example, the specific heat of an insulator is due entirely to its lattice vibrations. Solid argon, which is perhaps the simplest solid of all, consists of a regular array of neutral atoms with tightly bound closed-shell electrons. These electrons are held together primarily by the van der Waals force, and hence interact only with their nearest-neighbor atoms. The physical properties of such a solid are due entirely to the thermal vibrations of its atoms with respect to their equilibrium positions. Therefore, the specific heat for such a solid results entirely from its lattice vibrations. On the other hand, the specific heat for metals is dominated by the lattice-specific heat at high temperatures, and by the electronic specific heat at very low temperatures. The most important effect of the lattice vibration on metals or intrinsic semiconductors is that it is the main scattering source that limits the carrier mobility in these materials. In fact, the interaction between the electrons and lattice vibrations is usually responsible for the temperature dependence of the resistivity and carrier mobility in metals or lightly doped semiconductors. Furthermore, such interactions may also play an important role in the thermoelectric effects of metals and semiconductors.

According to the classical Dulong and Petit law, the lattice-specific heat for a solid is constant and equal to $3R$ ($= 5.96$ cal/(mol °C). The Dulong and Petit law gives a correct prediction of lattice-specific heat for most solids at high temperatures but fails at very low temperatures. The lattice specific heat can be derived from classical statistics as follows.

Consider a solid with N identical atoms that are bound together by an elastic force. If each atom has three degrees of freedom, then there will be $3N$ degrees of freedom for the N atoms, to produce $3N$ independent vibration modes, each with the same vibration frequency. According to classical statistics, the mean energy for each lattice vibration mode is $k_B T$, and hence the total energy U for $3N$ vibration

modes in a solid is equal to $3N\mathrm{k_B}T$. Thus, the lattice specific heat under a constant-volume condition is given by

$$C_{\mathrm{v}} = \frac{\mathrm{d}U}{\mathrm{d}T} = 3Nk_{\mathrm{B}} = 3R, \tag{2.1}$$

where $R(= Nk_{\mathrm{B}})$ is the ideal gas constant, and k_{B} is the Boltzmann constant ($= 1.38 \times 10^{-23}$ joule/K). For an ideal gas system, by substituting $N = 6.025 \times 10^{23}$ atoms/(g mol) (Avogadro's number) into (2.1) one finds that C_{v} is equal to 5.96 cal/(mol °C). This value is in good agreement with the experimental data for solids at high temperatures. However, (2.1) fails to predict correctly the lattice specific heat for most solids at very low temperatures. This is due to the fact that at very low temperatures, atoms in a solid are no longer vibrating independently of one another. Instead, the lattice vibration modes can be considered as a quasi-continuum, with a broad spectrum of vibration frequencies from very low frequencies up to a maximum frequency determined by the number of vibration modes available in the lattice.

In Section 2.2, expressions for the dispersion relations of a one-dimensional (1-D) monatomic linear chain and a diatomic linear chain are derived and described. The dispersion relation for a three-dimensional (3-D) lattice is derived and discussed in Section 2.3. The concept of phonons (i.e., quantized lattice vibration modes) in crystalline solids is discussed in Section 2.4. In Section 2.5, the phonon density of states function is derived, and the lattice spectra for some metals and semiconductors are presented. The Debye model for predicting the lattice specific heat of a solid over the entire range of temperature is discussed in Section 2.6.

2.2. The One-Dimensional Linear Chain

To understand the thermal and physical properties associated with atomic (lattice) vibrations in a solid, it is useful to consider two simple cases, namely, the one-dimensional (1-D) monatomic linear chain and the 1-D diatomic linear chain.

2.2.1. The Monatomic Linear Chain

In a 1-D monatomic linear chain, there is one atom per unit cell. If only the nearest- neighbor interaction is considered, then the linear chain can be represented by a string of identical masses connected to one another by a massless spring, as illustrated in Figure 2.1. In this case, the equation of motion for the atomic displacement can be easily derived using Hooke's law. According to this classical law, the force acting on the nth atom with mass m can be expressed as

$$\begin{aligned} F_n &= m\frac{\partial^2 u_n}{\partial t^2} = -\beta(u_n - u_{n+1}) - \beta(u_n - u_{n-1}) \\ &= -\beta(2u_n - u_{n+1} - u_{n-1}), \end{aligned} \tag{2.2}$$

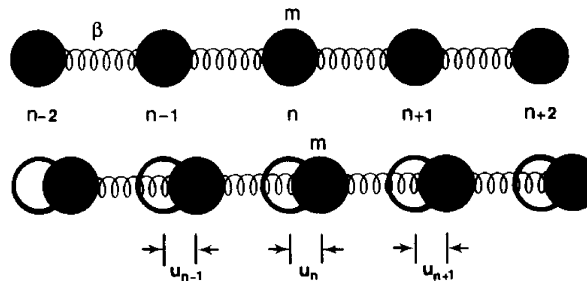


FIGURE 2.1. Lattice vibration of a one-dimensional monatomic linear chain. Here a is the lattice constant, β is the force constant, and μ_n is the displacement of the nth atom from its equilibrium position.

where β is the force constant between two adjacent atoms, and u_n, u_{n-1}, u_{n+1} denote the displacements of the nth, $(n-1)$th, and $(n+1)$th atoms, respectively. The solution of (2.2) has the form of a traveling wave, which is given by

$$u_n = u_q e^{i(naq-\omega t)}, \tag{2.3}$$

where $\boldsymbol{q}$ is the wave vector ($q = 2\pi/\lambda$) of the lattice wave, a is the lattice constant, and n is an integer. Note that u_q denotes the amplitude function of the lattice wave, which is also a function of wave vector $\boldsymbol{q}$. Substituting (2.3) into (2.2), one obtains

$$m\omega^2 = -2\beta(\cos qa - 1). \tag{2.4}$$

Equation (2.4) is the solution for a simple harmonic oscillator, which has a dispersion relation (ω vs. q) given by

$$\omega = 2\sqrt{\frac{\beta}{m}} \sin\left(\frac{qa}{2}\right) = \omega_m \sin\left(\frac{qa}{2}\right), \tag{2.5}$$

where $\omega_m = 2(\beta/m)^{1/2}$ is the maximum frequency of the lattice vibration modes. Figure 2.2 shows a plot of the dispersion relation for a 1-D monatomic linear chain obtained from (2.5). As shown in this figure, the dispersion curve has a period of $2\pi/a$.

The dispersion relation given by (2.5) for a 1-D monatomic linear chain exemplifies several fundamental physical properties of lattice dynamics in a solid. First, all the possible lattice vibration modes are limited by the allowed values of wave vector q, which fall in the range $-\pi/a \le q \le \pi/a$. This range is known as the first Brillouin zone for the dispersion curve of a 1-D monatomic linear chain. There are n independent wave vectors within the first Brillouin zone, representing n (i.e., $n = N$, where N is the total number of atoms) independent vibration modes. Each atomic displacement contributes to one lattice vibration mode. The maximum wave number $q_{\max}$, which occurs at the zone boundary, is given by

$$q_{\max} = \pi/a \approx 10^8 \text{ cm}^{-1}, \tag{2.6}$$

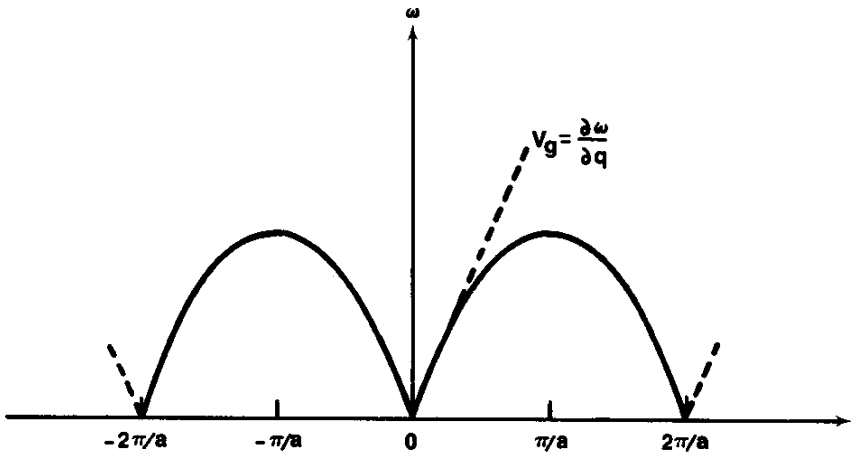


FIGURE 2.2. Dispersion curve for a one-dimensional monatomic linear chain.

where a is the lattice constant. Since the frequency ω is a periodic function of wave vector q in q-space, for any given wave vector q' outside the first Brillouin zone there is a corresponding wave vector q in the first Brillouin zone, which can be obtained by translational operation (i.e., $\boldsymbol{q}' = \boldsymbol{q} \pm \boldsymbol{K}$, where $\boldsymbol{K}$ is the reciprocal lattice vector). The translational symmetry operation in a crystal lattice has been discussed in detail in Chapter 1. At the zone boundaries, the solution of (2.3) does not represent a traveling wave but a standing wave. Thus, at the zone boundaries, $q_{\max} = \pm(n\pi/a)$, and u_n is given by

$$u_n = u_{q\,\max}\, \mathrm{e}^{i(n\pi - \omega t)} = u_{q\,\max}\, \mathrm{e}^{-i\omega t} \cos(n\pi). \tag{2.7}$$

Equation (2.7) shows that at the zone boundaries, $\cos n\pi = \pm 1$, depending on whether n is an even or an odd integer. This implies that the vibration modes for the alternate atoms are out of phase at the zone boundaries. The group velocity of the lattice wave packet is defined by

$$v_{\mathrm{g}} = \frac{\mathrm{d}\omega}{\mathrm{d}q}. \tag{2.8}$$

Solving (2.5) and (2.8) yields an expression for the group velocity, which is

$$v_{\mathrm{g}} = (\beta/m)^{1/2} a \cos(qa/2). \tag{2.9}$$

From (2.9), it is noted that at the zone boundaries where $q_{\max} = \pm\pi/a$, the group velocity v_{g} is equal to zero. Thus, the lattice wave is a standing wave packet at the zone boundaries, and the incident and reflected lattice waves have the same amplitude but travel in opposite directions.

In the long-wavelength limit (i.e., for $qa \to 0$), (2.5) reduces to

$$\omega = (\beta/m)^{1/2} aq, \tag{2.10}$$

which shows that for $qa \to 0$, the vibration frequency ω of the lattice waves is directly proportional to the wave vector q. This corresponds to the common property of ordinary elastic waves in a continuum medium. In this case, the group velocity ($v_g = \mathrm{d}\omega/\mathrm{d}q$) and the phase velocity ($v_\mathrm{p} = \omega/q$) are equal, and their values can be determined from the slope of the dispersion curve at small q value, as is shown in Figure 2.2. Using $a = 3$ Å and $\mathrm{v_s} = 10^5$ cm/s, one obtains a value of $(\beta/m)^{1/2} \approx 3 \times 10^{12}$ s^{-1}, and the maximum vibration frequency that a lattice can support is $\omega_\mathrm{max} = 2(\beta/m)^{1/2} = 6 \times 10^{12}$ s^{-1}; this value falls in the infrared spectral regime of the electromagnetic radiation spectrum.

2.2.2. *The Diatomic Linear Chain*

The dispersion relation for a 1-D diatomic linear chain will be derived next. Figure 2.3 shows a 1-D diatomic linear chain, which contains two types of atoms with different masses per unit cell. The atoms are equally spaced, but with different masses placed in alternate positions along the linear chain. If one assumes that only the nearest-neighbor interactions are important, then the force constant β between the two different mass atoms is the same throughout the entire linear chain. Therefore, there are two atoms per unit cell, with masses of m_1 and m_2. Using Hooke's law, the equations of motion for the $2n$th and $(2n + 1)$th atoms of the 1-D diatomic linear chain can be written as

$$m_1 \frac{\partial^2 u_{2n}}{\partial t^2} = \beta(u_{2n+1} + u_{2n-1} - 2u_{2n}), \tag{2.11}$$

$$m_2 \frac{\partial^2 u_{2n+1}}{\partial t^2} = \beta(u_{2n+2} + u_{2n} - 2u_{2n+1}). \tag{2.12}$$

Solutions of (2.11) and (2.12) can be expressed, respectively, as

$$u_{2n} = u_a \mathrm{e}^{i(2naq - \omega t)}, \tag{2.13}$$

$$u_{2n+1} = u_o \mathrm{e}^{i[(2n+1)aq - \omega t]}, \tag{2.14}$$

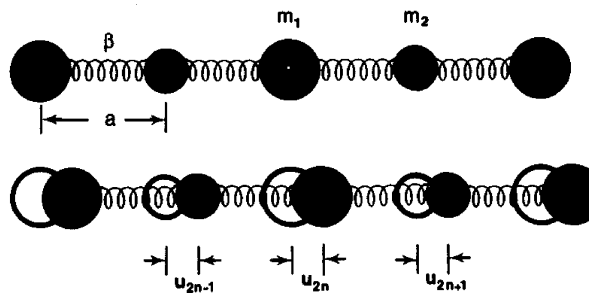


FIGURE 2.3. A diatomic linear chain in equilibrium position and in the displaced position, assuming $m_1 > m_2$.

where u_{2n} and u_{2n+1} are the displacements for the $2n$th and $(2n+1)$th atoms, respectively. Now, substituting (2.13) and (2.14) into (2.11) and (2.12), one obtains

$$2\beta u_a - m_1\omega^2 u_a - 2\beta u_o \cos aq = 0, \tag{2.15}$$

$$2\beta u_o - m_2\omega^2 u_o - 2\beta u_a \cos aq = 0. \tag{2.16}$$

Equations (2.15) and (2.16) will have a nontrivial solution if and only if the determinant for the coefficients of u_a and u_o in both equations is set equal to zero. Thus, the frequency ω must satisfy the secular equation given by

$$\begin{vmatrix} (2\beta - m_1\omega^2) & -2\beta\cos(qa) \\ -2\beta\cos(qa) & (2\beta - m_2\omega^2) \end{vmatrix} = 0. \tag{2.17}$$

Solving (2.17) for ω yields

$$\omega^2 = \beta\left\{\left(\frac{1}{m_1}+\frac{1}{m_2}\right) \pm \left[\left(\frac{1}{m_1}+\frac{1}{m_2}\right)^2 - \frac{4\sin^2(aq)}{m_1 m_2}\right]^{1/2}\right\}. \tag{2.18}$$

Using the same argument as in the case of a monatomic linear chain, one finds that the allowed values of $|q|$ for the diatomic linear chain are given by

$$|q| = \frac{n\pi}{Na}, \tag{2.19}$$

where N is the total number of unit cells in the linear chain and n is an integer. Since the period of a diatomic linear chain is equal to $2a$, the first Brillouin zone is defined by

$$\frac{-\pi}{2a} \le q \le \frac{\pi}{2a}, \tag{2.20}$$

which is a factor of 2 smaller than the first Brillouin zone of the 1-D monatomic linear chain. Figure 2.4 shows the dispersion curves for the 1-D diatomic linear chain with $m_1 > m_2$. The upper curve shown in Figure 2.4 corresponds to the plus sign given by (2.18), and is called the optical branch. The lower curve, which corresponds to the minus sign in (2.18), is known as the acoustical branch. The lattice vibration modes in the optical branch can usually be excited by the infrared optical radiation, which has frequencies in the range from 10^{12} to 10^{14} Hz. For the acoustical branch, the lattice vibration modes can be excited if the crystal is connected to an acoustical wave transducer that produces pressure waves throughout the crystal. In general, the dispersion curves for a solid with two atoms per unit cell contain both the acoustical and optical branches. For example, the dispersion curves for an alkalihalide crystal such as NaCl consist of both acoustical and optical branches, contributed by the positively and negatively charged ions (i.e., Na^+, Cl^-) in the crystal.

The physical insights for the dispersion curves of a diatomic linear chain can be best explained by considering two limiting cases, namely, (i) the long-wavelength limit (i.e., $qa \to 0$) and (ii) near the zone boundaries (i.e., $q \to \pm\pi/2a$). These are discussed as follows:

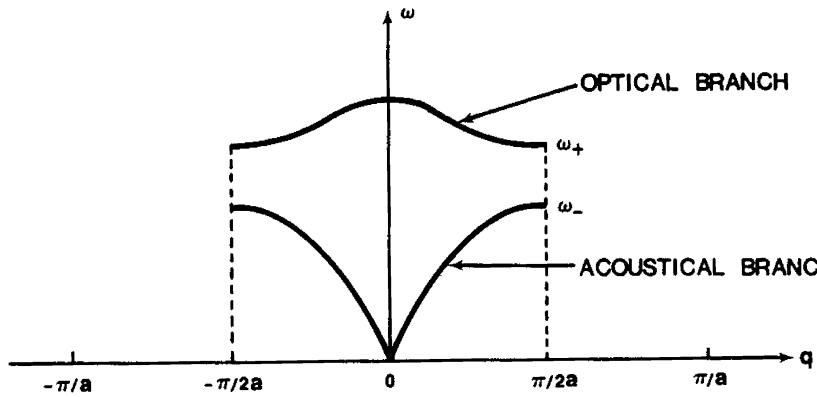


FIGURE 2.4. Dispersion curves of a one-dimensional diatomic linear chain, assuming $m_1 > m_2$.

(i) The acoustical branch. For $qa \to 0$, using the minus sign in (2.18) for the acoustical branch, one obtains

$$\omega = \left[\frac{2\beta}{(m_1 + m_2)}\right]^{1/2} aq. \tag{2.21}$$

Equation (2.21) shows that in the long-wavelength limit, ω is directly proportional to the wave vector q. This result is identical to the monatomic linear chain discussed in the previous section. Furthermore, from (2.16), and for $qa \to 0$, the ratio of the amplitude of two different mass atoms is given by

$$\frac{u_a}{u_o} = 1, \tag{2.22}$$

which implies that in the long-wavelength limit, atoms at odd and even lattice sites are moving in phase with equal amplitude. From the above analysis, it is obvious that in the long-wavelength limit the dispersion curve in the acoustical branch for a diatomic linear chain will reduce to that of a monatomic linear chain if the masses of two different mass atoms are equal (i.e., $m_1 = m_2 = m$).

(ii) The optical branch. The optical branch for the 1-D diatomic linear chain is shown by the upper curve of Figure 2.4. In the long-wavelength limit, when qa approaches zero, one obtains

$$\omega = \left[\frac{2\beta(m_1 + m_2)}{m_1 m_2}\right]^{1/2}. \tag{2.23}$$

Solving (2.15) and (2.16) for $qa \to 0$, the ratio of the amplitude of two different mass atoms is given by

$$\frac{u_a}{u_o} = -\frac{m_2}{m_1}. \tag{2.24}$$

Equation (2.23) shows that in the optical branch as $qa \to 0$, the frequency ω becomes a constant and independent of the wave vector q. Thus, the lattice vibration mode at $qa = 0$ represents a standing wave. The ratio of the amplitude factor given by (2.24) reveals that the lattice vibration modes of alternate masses are out of phase, and the amplitude of the vibration modes is inversely proportional to the mass ratio of the alternate atoms. This can be best explained by considering the alkali-halide crystal. The two types of ions in an alkali-halide crystal (e.g., NaCl) are oppositely charged, and hence will experience opposing forces when an electric field is applied to the crystal. As a result, the motions of atoms in alternate lattice sites are out of phase with each other with an amplitude ratio inversely proportional to their mass ratio. If electromagnetic waves with frequencies corresponding to the frequencies of optical lattice vibration modes are applied to the crystal, resonant absorption takes place. Since the frequencies in which the lattice vibration modes are excited in this branch usually fall in the infrared spectral range, it is referred to as the optical branch.

Another feature of the dispersion curves shown in Figure 2.4 is the existence of a forbidden gap between $\omega_- = (2\beta/m_1)^{1/2}$ and $\omega_+ = (2\beta/m_2)^{1/2}$ at the zone boundaries (i.e., $q_{\max} = \pm\pi/2a$). The forbidden region corresponds to frequencies in which lattice waves cannot propagate through the linear chain without attenuation. This can be easily verified by substituting a value of ω, which falls in the forbidden region of the dispersion curve shown in Figure 2.4, into (2.18). In this case, the wave vector q becomes a complex number, and the lattice waves with frequencies falling in the forbidden zone will attenuate when they propagate through the linear chain. It is interesting to note that a similar situation also exists in the energy band scheme of a semiconductor in which a forbidden band gap exists between the valence band and the conduction band at the zone boundaries of the first Brillouin zone. This will be discussed in detail in Chapter 4.

2.3. Dispersion Relation for a Three-Dimensional Lattice

The dispersion relation for the 1-D linear chain derived in Section 2.2 can be easily extended to the 2-D and 3-D lattices by considering the lattice vibration modes as simple harmonic oscillators. As frequently encountered in quantum mechanics, the displacement of atoms can be expressed in terms of the normal coordinates and normal modes of quantum oscillators. According to quantum mechanics, the lattice vibration modes generated by the atomic vibrations in their equilibrium positions can be represented by the harmonic oscillators, with each vibration mode having its own characteristic frequency ω and wave vector q. According to quantum theory, the equation of motion for a 3-D harmonic oscillator is given by

$$\ddot{Q}_{q,s} + \omega^2(q, s)Q_{q,s} = 0, \tag{2.25}$$

where $Q_{q,s}$ is the normal coordinate and $\omega(q, s)$ denotes the normal frequency. For a 3-D lattice both $Q_{q,s}$ and $\omega(q, s)$ are functions of wave vector q and polarization index s (i.e., $s = 1, 2, 3$).

In general, if a crystal contains only one atom per unit cell, then there are three possible polarizations for each wave vector q, one longitudinal and two transverse modes of polarization. In the longitudinal mode of lattice vibration, the motion of atoms is along the direction of wave propagation, while for the transverse modes the motion of atoms is in the plane perpendicular to the direction of wave propagation. If a crystal contains N atoms per unit cell, then the index n varies from 1 to 3 N. For example, if $N = 2$ and $s = 3$, then there are three polarizations (i.e., one longitudinal and two transverse modes) in the acoustical branch and three polarizations in the optical branch.

Figure 2.5a–c shows the measured lattice dispersion curves for silicon, GaAs, and aluminum, respectively, obtained from the inelastic slow neutron experiment.[1–3] As can be seen in this figure, the dispersion curves for these materials are strongly dependent on the crystal orientations. This is due to the fact that lattice vibration modes depend strongly on crystal symmetry and atomic spacing along a particular crystal orientation. For example, the atomic spacing for a silicon crystal along the (100) axis is different from those along the (111) and (110) axes. As a result, the dispersion relations for the silicon lattice are different along the (100), (110), and (111) orientations. A similar situation exists in GaAs and aluminum. In general, the dispersion curves for most solids can be determined from the inelastic slow neutron experiment. In this experiment, the energy losses of a slow neutron due to scattering by the lattice vibrations and the change of wave vector during scattering can be determined experimentally by the conservation of energy and momentum. A slow neutron impinging on a crystal sees the crystal lattice mainly by interacting with the nuclei of the atoms. The momentum conservation for the slow neutron scattering by a lattice vibration mode can be described by

$$\boldsymbol{k} = \boldsymbol{k}' \pm \boldsymbol{q} + \boldsymbol{K}, \tag{2.26}$$

where $\boldsymbol{k}$ is the wave vector of the incident neutron, $\boldsymbol{k}'$ is the wave vector of the scattered neutron, $\boldsymbol{q}$ is the phonon wave vector, and $\boldsymbol{K}$ is the reciprocal lattice vector. The plus sign in (2.26) denotes the creation of a phonon, while the minus sign is for annihilation of a phonon. Note that we have introduced here the terminology "phonon" to represent the quantized lattice vibration.

The conservation of energy for scattering of a slow neutron by a lattice atom is given by

$$\frac{\hbar^2 k^2}{2M_n} = \frac{\hbar^2 k'^2}{2M_n} \pm \hbar\omega_q, \tag{2.27}$$

where $\hbar\omega_q$ is the phonon energy. In (2.27), the plus sign is for phonon emission and the minus sign is for phonon absorption. The dispersion relation for the lattice vibration modes of a crystalline solid can be determined by (2.26) and (2.27) using the energy gain and loss of the scattered neutrons as a function of the scattering direction (i.e., $\boldsymbol{k} - \boldsymbol{k}'$) from the slow neutron experiment. This method has been widely used in determining the phonon spectra of metals, insulators, and semiconductors. The concept of phonon will be discussed next.

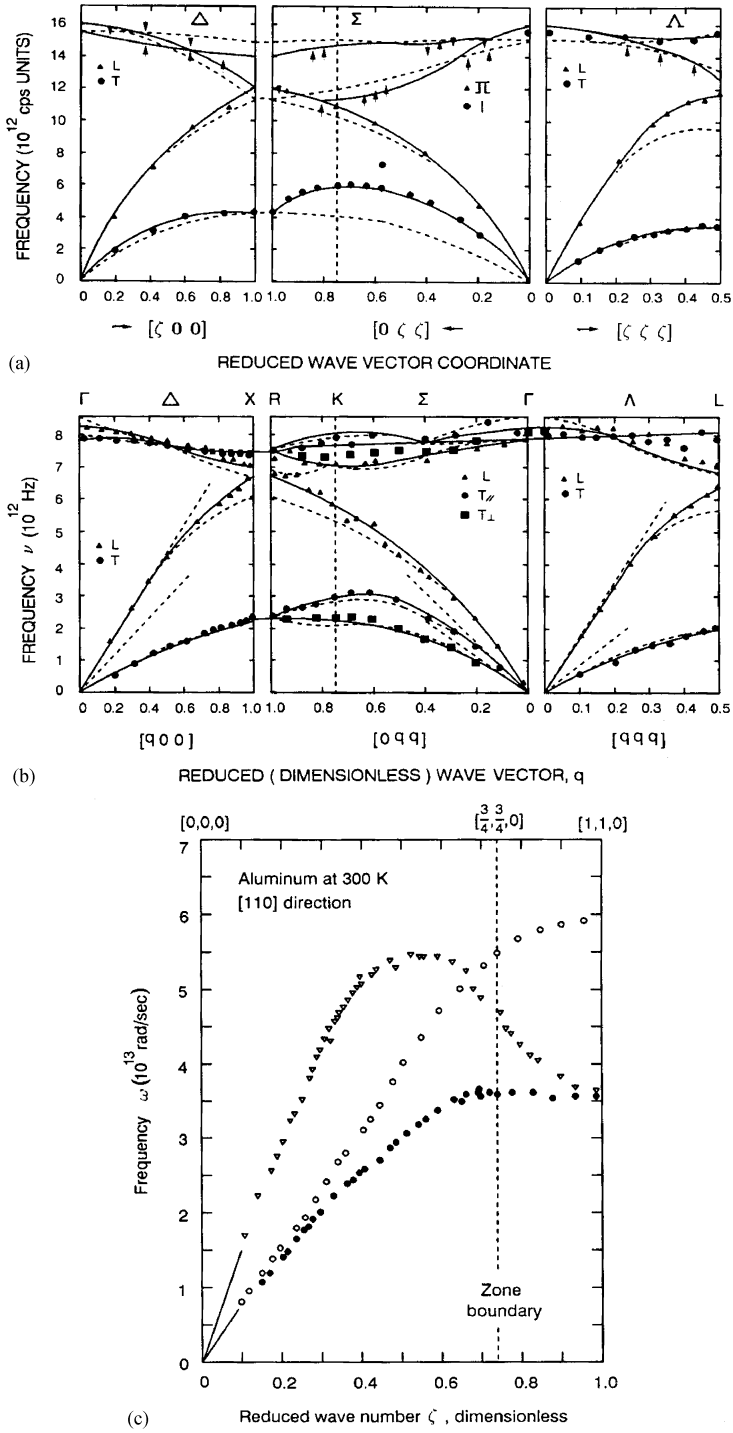


FIGURE 2.5. (a) Dispersion curves for silicon along the (100) and (111) orientations, (b) for GaAs along the (111), (110), and (100) orientations, and (c) for aluminum. After Dolling,[1] Waugh and Dolling,[2] Wallis,[3] by permission.

2.4. The Concept of Phonons

The dispersion relations derived in Section 2.2 for the 1-D monatomic and diatomic linear chains are based on Hooke's law. The results of this classical approach provide a good insight concerning the physical properties of lattice waves, which is important to the understanding of the specific heat of a solid. However, it is not a common practice to use the wave concept of lattice vibrations to deal with the problems of interactions between electrons and lattice waves in a crystalline solid, such as scattering of electrons by the lattice waves in a semiconductor or a metal. Instead, the quantum-mechanical approach is usually used to solve the problems of scattering of electrons by the lattice vibrations in a solid. In the framework of quantum mechanics, each lattice vibration mode is quantized and can be treated as a quantum oscillator with a characteristic frequency ω and a wave vector q. This quantized lattice vibration mode is usually referred to as the "phonon," analogous to the term " photon" as a quantum unit of the electromagnetic radiation. Therefore, it is appropriate to introduce here the concept of "phonons" to represent the quantized lattice vibration modes in a crystalline solid.

In Section 2.3, the normal coordinates and normal modes are introduced to describe the quantum oscillators for the 3-D lattice vibration modes in a crystalline solid. A quantized lattice vibration mode (phonon) can be represented by a harmonic oscillator, which has a characteristic frequency ω, wave vector q, and polarization index s. According to quantum theory, the energy of a harmonic oscillator is given by

$$E_n = (n + 1/2)\hbar\omega, \tag{2.28}$$

where $n = 0, 1, 2, 3, \ldots, \hbar = h/2\pi$; h is Planck's constant, and ω is the characteristic frequency of the quantum oscillator. Using (2.28), the phonon energy can be written as

$$E_n(q, s) = (n_{q,s} + 1/2)\hbar\omega(q, s), \tag{2.29}$$

where $n_{q,s} = 1/[\exp(\hbar\omega/k_B T) - 1]$ is the average phonon occupation number, which can be derived using Bose–Einstein statistics, to be discussed in Chapter 3. The quantity $\hbar\omega/2$ on the right-hand side of (2.29) represents the zero-point phonon energy (i.e., $n_{q,s} = 0$). It should be noted that the zero-point energy does not affect the phonon distribution function in any way, nor does it contribute to the average internal energy and the specific heat of a solid at temperatures above absolute zero. A large value of $n_{q,s}$ in (2.29) corresponds to phonons with large amplitude, and vice versa. In the dispersion curves shown in Section 2.2, the acoustical branch consists of both the longitudinal and transverse acoustical (LA and TA) phonons, while the optical branch is composed of longitudinal and transverse optical (LO and TO) phonons.

It should be noted that phonon scatterings are usually the dominant scattering mechanisms in intrinsic and lightly doped semiconductors, and hence they control the carrier mobilities in these semiconductors. This will be discussed further in Chapters 7 and 8.

2.5. The Density of States and Lattice Spectrum

The density of states function for phonons in a crystalline solid can be derived using the periodic boundary conditions of the crystalline solids. For a 3-D cubic lattice, if the length of each side of the cubic unit cell is equal to L, then the density of states function can be derived using the periodic boundary conditions over the N atoms within the cubic unit cell with a volume of L^3. Values of the phonon wave vector q are determined using the 3-D periodic boundary conditions, given by

$$e^{i(q_x x+q_y y+q_z z)} = e^{i[q_x(x+L)+q_y(y+L)+q_z(z+L)]}, \tag{2.30}$$

which reduces to

$$e^{i(q_x+q_y+q_z)L} = 1. \tag{2.31}$$

From (2.31) one obtains $q_x, q_y, q_z = 0, \pm 2\pi/L, \pm 4\pi/L, \ldots, N\pi/L$. Therefore, there is one allowed value of q per unit volume $(2\pi/L)^3$ in the reciprocal space (i.e, the q-space). To find a general expression for the phonon density of states function $D(\omega)$ for a 3-D crystal lattice, the total number of states per unit volume with frequencies between ω and $\omega + d\omega$ can be expressed by

$$D(\omega)\,\mathrm{d}\omega = \left(\frac{L}{2\pi}\right)^3 \int_{\text{shell}} \mathrm{d}^3 q. \tag{2.32}$$

The integrand of (2.32) represents the total number of states available within a spherical shell in q-space with frequencies varying between ω and $\omega + \mathrm{d}\omega$, and $(2\pi/L)^3$ is the volume of the unit cell in q-space. As shown in Figure 2.6, $\mathrm{d}S_\omega$ is the area element of the constant frequency surface in q-space, and $\mathrm{d}^3 q$ is the

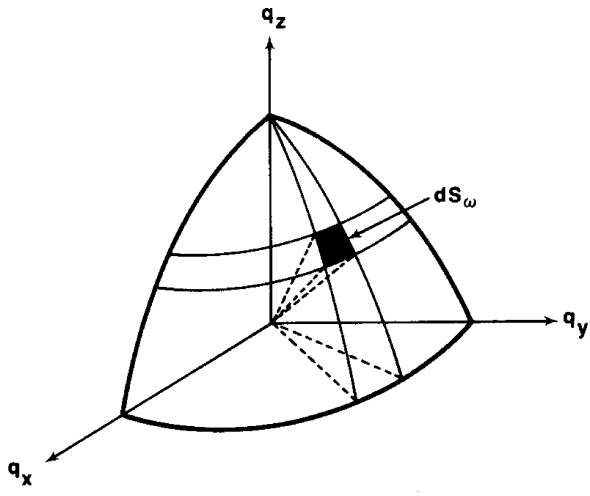


FIGURE 2.6. Constant-frequency surface in q-space; $\mathrm{d}S_\omega \mathrm{d}q$ is the volume element between two surfaces of constant frequency ω and $\omega + \mathrm{d}\omega$.

volume element, which can be expressed as

$$\mathrm{d}^3q = \mathrm{d}S_\omega \mathrm{d}q_\perp. \tag{2.33}$$

Now substituting (2.33) into (2.32) and using the relation $\mathrm{d}q_\perp = \mathrm{d}\omega / \left|\nabla_q \omega\right|$, one obtains a general expression for the phonon density of states function as

$$D(\omega) = \left(\frac{L}{2\pi}\right)^3 \int_{\text{shell}} \frac{\mathrm{d}S_\omega}{\left|\nabla_q \omega\right|}, \tag{2.34}$$

or

$$D(\omega) = \left(\frac{L}{2\pi}\right)^3 \int \frac{\mathrm{d}S_\omega}{v_\mathrm{g}} \tag{2.35}$$

where $v_\mathrm{g} = |\nabla_q \omega|$ is the group velocity of the phonons. Note that integration of (2.34) is carried out over the constant-frequency surface in q-space.

It is clear that an expression for the phonon density of states function can be derived from (2.34), provided that the dispersion relation between ω and q is known. Figure 2.7a shows the plot of phonon density of states as a function of frequency for copper, and Figure 2.7b illustrates the corresponding density of states function derived using a dispersionless relation, $\omega = u_\mathrm{s} q$, in (2.34), where u_s is the velocity of sound.

The phonon density of states plot shown in Figure 2.7a for the copper crystal was obtained from the numerical analysis of the measured dispersion curve. In general, if the constant-frequency surface in q-space is spherical, then in the long-wavelength limit, the density of states function $D(\omega)$ is proportional to the square of the frequency. It should be noted that (2.34) could be applied to the derivation of the density of states function for electrons in the conduction band or for holes in the valence band of a semiconductor. This can be achieved by including the spin

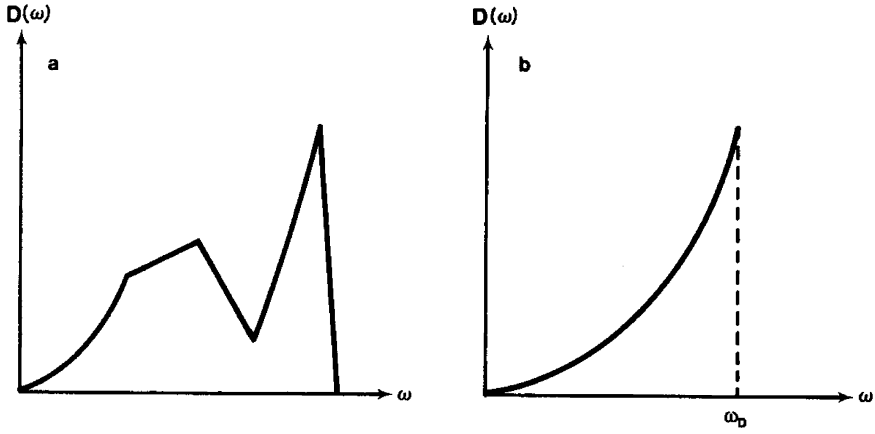


FIGURE 2.7. (a) Density of phonon states versus frequency for copper. (b) Debye lattice spectrum, where ω_D is the Debye cutoff frequency.

degeneracy factor (= 2) due to the Pauli's exclusion principle and by replacing the frequency of phonons by the energy of electrons or holes in (2.34), as will be discussed further in Chapter 4.

2.6. Lattice Specific Heat

The classical Dulong and Petit law described in Section 2.1 fails to predict correctly the temperature dependence of the lattice specific heat of solids at very low temperatures. The reason for its failure arises from the fact that Dulong and Petit's law does not consider all lattice vibration modes with different vibration frequencies, particularly the long-wavelength phonons, which are the dominant vibration modes at low temperatures. In deriving the lattice specific heat for solids, Debye uses a continuum model to account for all the possible lattice vibration modes. This assumption is valid as long as the wavelength of phonons is large compared to the interatomic spacing. In this respect, a solid is considered as a continuous medium to the lattice phonons. Furthermore, the number of vibration modes is limited by the total number of constituent atoms in the crystal, which is equal to N. Therefore, for N atoms each with three degrees of freedom, the total number of vibration modes is equal to $3N$. In other words, the frequency spectrum corresponding to a perfect continuum is cut off so as to comply with a total of $3N$ vibration modes. The Debye cutoff frequency, ω_{D}, corresponds to the maximum frequency that the transverse and longitudinal vibration modes can support. The Debye model for the lattice specific heat of a solid is discussed next.

To derive the lattice specific heat of a crystalline solid, it is necessary to find the total internal energy due to the thermal vibrations of lattice atoms. Using (2.29) for the average phonon energy ($= n_{q,s}\,\hbar\omega$) and ignoring the zero-point energy, the total energy of the lattice phonons with frequencies varying from zero to a cutoff frequency ω_{D} is given by

$$U = \int_0^{\omega_{\mathrm{D}}} \frac{D(\omega)\,\hbar\omega \mathrm{d}\omega}{\left(\mathrm{e}^{\hbar\omega/k_{\mathrm{B}}T} - 1\right)}, \tag{2.36}$$

where $D(\omega)$ is the density of states function per unit frequency given by (2.34). To find the solution of (2.36), the expression for $D(\omega)$ and the dispersion relation between ω and q must be first derived. In the Debye model, it is assumed that the solid under consideration is an isotropic dispersionless continuum medium, and hence the relation between ω and q is given by

$$\omega = v_{\mathrm{g}} q = v_{\mathrm{p}} q = u_{\mathrm{s}} q, \tag{2.37}$$

where v_{g}, v_{p}, and u_{s} denote the group velocity, phase velocity, and the velocity of sound in a solid, respectively. From (2.37), it is noted that the group and phase velocities are equal to the velocity of sound in a dispersionless continuum medium. Furthermore, in an isotropic medium, the phase velocity of phonons is independent of the direction of the wave vector q. To derive the phonon density of states function, one can consider the spherical shell in q-space as shown in Figure 2.6.

From (2.34) through (2.37) one can determine the number of vibration modes within the spherical shell for frequencies between ω and $\omega + \mathrm{d}\omega$, and the result yields

$$D(\omega)\mathrm{d}\omega = \left(\frac{L}{2\pi}\right)^3 \int \frac{\mathrm{d}S_\omega}{v_\mathrm{g}}\mathrm{d}\omega = \left(\frac{3V\omega^2}{2\pi^2 u_\mathrm{s}^3}\right)\mathrm{d}\omega, \tag{2.38}$$

where $V = L^3$ is the volume of the cubic unit cell, and the surface integral ($\int \mathrm{d}S_\omega$) is equal to $4\pi q^2$, or $4\pi\omega^2/{u_\mathrm{s}}^2$. A typical Debye spectrum calculated from (2.38) is shown in Figure 2.7b. A factor of 3 is included in (2.38) to account for the three components of polarizations (i.e., two transverse and one longitudinal) per wave vector. In general, the propagation velocities for the transverse-mode phonons and the longitudinal-mode phonons are not equal (i.e., $v_\mathrm{t} \neq v_\mathrm{l}$), and hence (2.38) must be replaced by

$$D(\omega)\mathrm{d}\omega = \left(\frac{V}{2\pi^2}\right)\left(\frac{2}{v_\mathrm{t}^3} + \frac{1}{v_\mathrm{l}^3}\right)\omega^2\mathrm{d}\omega. \tag{2.39}$$

The Debye cutoff frequency ω_D can be obtained by integrating (2.39) for frequencies from 0 to ω_D, and using the fact that there are $3N$ total vibration modes in the crystal. Thus, the total number of vibration modes is given by

$$\int_0^{\omega_\mathrm{D}} D(\omega)\,\mathrm{d}\omega = 3N. \tag{2.40}$$

Substituting (2.38) for $D(\omega)$ into (2.40) yields

$$\omega_\mathrm{D} = (6\pi^2 n)^{1/3} u_\mathrm{s}, \tag{2.41}$$

where $n = N/V$ is the number of atoms per unit volume and u_s is given by

$$u_\mathrm{s} = \left[\frac{1}{3}\left(\frac{2}{v_\mathrm{t}^3} + \frac{1}{v_\mathrm{l}^3}\right)\right]^{-1/3}, \tag{2.42}$$

where u_s is the velocity of sound in the solid. The total energy of the phonons can be obtained by substituting (2.38) into (2.36) and integrating (2.36) from $\omega = 0$ to $\omega = \omega_\mathrm{D}$, which yields

$$U = \left(\frac{3V}{2\pi^2 u_\mathrm{s}^3}\right)\int_0^{\omega_\mathrm{D}} \frac{\hbar\omega^2\mathrm{d}\omega}{(\mathrm{e}^{\hbar\omega/k_\mathrm{B}T} - 1)} = \left(\frac{3Vk_\mathrm{B}^4T^4}{2\pi^2\hbar^3 u_\mathrm{s}^3}\right)\int_0^{x_m} \frac{x^3\mathrm{d}x}{(\mathrm{e}^x - 1)}, \tag{2.43}$$

where $x = \hbar\omega/k_\mathrm{B}T$, $x_m = \hbar\omega_\mathrm{D}/k_\mathrm{B}T = T_\mathrm{D}/T$, and $T_\mathrm{D} = \hbar\omega_\mathrm{D}/k_\mathrm{B}$ is called the Debye temperature. The lattice specific heat under constant volume can be obtained from (2.43) by differentiating the total energy U with respect to temperature, which yields

$$C_\mathrm{v} = \frac{\mathrm{d}U}{\mathrm{d}T} = 9Nk_\mathrm{B}\left(\frac{T}{T_\mathrm{D}}\right)^3 \int_0^{T_\mathrm{D}/T} \frac{\mathrm{e}^x x^4\,\mathrm{d}x}{(\mathrm{e}^x - 1)^2}, \tag{2.44}$$

where $T_\mathrm{D} = \hbar\omega_\mathrm{D}/k_\mathrm{B} = (\hbar u_\mathrm{s}/k_\mathrm{B}L)(6\pi^2 N)^{1/3}$ is used in (2.44). It is noted that an

analytical expression valid over the entire temperature range could not be obtained from (2.44). However, an analytical expression may be derived for two limiting cases, namely, for $T \gg T_D$ and $T \ll T_D$. They are described as follows:

(i) The high-temperature regime ($T \ll T_D$ *or* $x \ll 1$). In the high-temperature regime, (2.44) can be simplified to

$$C_v \approx 9Nk_B \left(\frac{T}{T_D}\right)^3 \int_0^{T_D/T} \frac{x^4\,dx}{x^2} = 3Nk_B = 3R, \tag{2.45}$$

which is identical to the result predicted by the classical Dulong and Petit's law for the lattice specific heat of solids.

(ii) The low-temperature regime ($T \ll T_D$ *or* $x \gg 1$). In this case, the upper limit of the integral in (2.43) for the total energy of phonons may be replaced by infinity, and the definite integral is given by

$$\int_0^\infty \frac{x^3\,dx}{(e^x - 1)} = \frac{\pi^4}{15}. \tag{2.46}$$

Now, substituting (2.46) into (2.43) and differentiating the total energy U with respect to T, one obtains the lattice specific heat as

$$C_v = \left(\frac{12\pi^4}{5}\right)(Nk_B)\left(\frac{T}{T_D}\right)^3. \tag{2.47}$$

Equation (2.47) shows that the lattice specific heat of a crystalline solid is proportional to T^3 at low temperatures. The result given by (2.47) provides a correct prediction of the temperature dependence of the lattice specific heat for both semiconductors and insulators at low temperatures. The reason for the good agreement is attributed to the fact that the Debye model takes into account the contribution of the long-wavelength acoustical phonons to the lattice specific heat, which is dominant at low temperatures. Figure 2.8 shows a comparison of the lattice specific heat versus temperature predicted by the Debye model and by Dulong and Petit's law.

Although the Debye model generally gives a correct prediction of the lattice specific heat for both insulators and semiconductors over a wide range of temperatures, the Debye temperature used in theoretical fitting of the experimental data varies from material to material. For example, the Debye temperature is $T_D = 640$ K for silicon and 370 K for germanium.

In spite of the success of the Debye model in predicting the correct temperature behavior of lattice specific heat of semiconductors and insulators over a wide range of temperatures, it fails to predict the correct temperature dependence of the specific heat of metals at very low temperatures. The reason for its failure stems from the fact that the electronic specific heat, which is influenced by the total kinetic energy of electrons in a metal, becomes dominant at very low temperatures. In fact, the specific heat of a metal is dominated by the electronic specific heat rather than by the lattice specific heat at very low temperatures. Using Fermi–Dirac

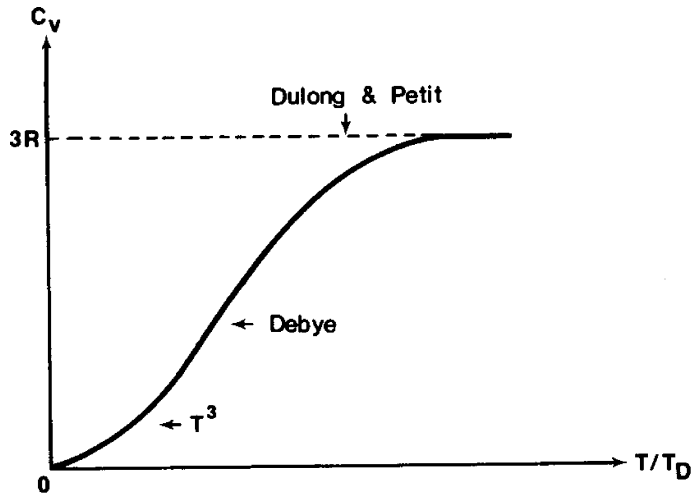


FIGURE 2.8. Lattice specific heat C_v versus normalized temperature T/T_D, as predicted by the Debye model and the Dulong and Petit law.

statistics it can be shown that the electronic specific heat for a metal varies linearly with temperature at very low temperatures, which is in good agreement with the experimental observation of the specific heat of metals at very low temperatures. Thus, the total specific heat for a metal consists of the lattice specific heat and the electronic specific heat, which can be expressed by

$$C_v = C_l + C_e = \alpha T^3 + \beta T, \tag{2.48}$$

where C_l and C_e denote the lattice and electronic specific heats, respectively. Both α and β are constants, which can be determined from the C_v/T versus T^2 plot. From the slope of this plot one can determine the constant α, and the intercept, when extrapolated to $T = 0$ K, yields the constant β. It is noted that at very low temperatures the electronic specific heat prevails in metals and the second term in (2.48) becomes dominant. Therefore, the specific heat of metals varies linearly with temperature at very low temperatures. Derivation of the electronic specific heat for a metal can be made using Fermi–Dirac statistics, to be discussed in Chapter 3.

Problems

2.1. (a) Considering only the nearest-neighbor interaction, find the dispersion relation for the diatomic linear chain of a silicon lattice along the (111) crystal axis. Note that the masses of the silicon atoms are identical, and the positions of the nearest-neighbor atoms on the (111) axis are located at (0,0,0), $(a/4, a/4, a/4)$, and (a, a, a) inside the unit cell. Assume that the force constant between the nearest-neighbor atoms is equal to β.

(b) Plot the dispersion curves (ω vs. q) from the result obtained in (a).
(c) Sketch the atomic displacement for the longitudinal and transverse optical lattice vibration modes at $q = \pi/4\sqrt{3}a$, $\pi/2\sqrt{3}a$, and $\pi/\sqrt{3}a$.

2.2. Using the Einstein model, derive the lattice specific heat for a 3-D crystal lattice. (*Hint*: The Einstein model is similar to the Debye model except that it assumes a single vibration frequency ω_E and energy $E = \hbar\omega_E$ for all lattice phonons. The phonon distribution function is $\langle n \rangle = 1/(e^{\hbar\omega_E/k_B T} - 1)$. Based on this assumption you can derive the average phonon energy and the lattice specific heat.)

2.3. (a) For a 1-D monatomic linear chain with fixed-end boundary condition, show that the density of states function can be expressed by

$$D(\omega) = \frac{L}{\pi} \cdot \frac{dq}{d\omega} = \left(\frac{2L}{\pi a}\right) \cdot \frac{1}{(\omega_m^2 - \omega^2)^{1/2}}.$$

(*Hint*: Use the dispersion relation given by (2.5) to derive $D(\omega)$ for the 1-D case.)

(b) Apply the Debye model to this 1-D linear chain, and show that the density of states function can be expressed by

$$D(\omega) = \frac{L}{\pi v_s},$$

where L is the length of the linear chain.

(c) Plot $D(\omega)$ versus ω for (a) and (b), and explain the difference.

2.4. Using the Debye model derive the specific heat for a 1-D monatomic linear chain with only nearest-neighbor interaction, and show that at low temperatures the specific heat varies linearly with temperature. What is the cutoff frequency for this case?

2.5. (a) Write down the equations of motion for a 1-D linear chain of identical masses that are connected to each other by springs of two difference force constants, β_1 and β_2, in alternating positions. Find the dispersion relation for this linear chain.

(b) Plot the dispersion curve for the linear chain given in (a).

2.6. (a) Write down the equation of motion for a 2-D square lattice with spacing a and atomic mass M. The nearest-neighbor force constant is given by β.

(b) Assume that the solution of (a) is given by

$$u_{lm} = u(0) \exp[i(lk_x a + mk_y a - \omega t)],$$

where u_{lm} denotes the displacement normal to the plane of the square lattice for the atom in the lth column and mth row. Find the dispersion relation in (a).

(c) Plot the dispersion curve for a square lattice based on the result obtained in (b).

2.7. Derive the density of states function $D(\omega)$ for a 1-D linear chain of length L carrying $N + 1$ particles with a spacing of a for the following cases:

(a) The particles $s = 0$ and $s = N$ at the ends of the linear chain are held fixed (i.e., fixed boundary condition).
(b) The linear chain is allowed to form a ring, so that the periodic boundary condition can be applied to the problem [i.e., $u(sa) = u(sa + L)$].
(c) What are the allowed values of the wave vector q in cases (a) and (b)?

2.8. One common method used in determining the phonon spectra of a solid is the slow neutron scattering experiment. Give an example to explain this technique for determining the phonon spectra in a crystal.[4,5]

References

1. G. Dolling, *Proceedings of Symposium on Inelastic Scattering in Solids and Liquids, II*, Chalk River, IAEA, Vienna (1963), p. 37.
2. J. L. T. Waugh and G. Dolling, *Phys. Rev.* **132**, 2410 (1963).
3. R. F. Wallis (ed.), *Lattice Dynamics*, Pergamon Press, New York (1965), p. 60.
4. A. D. B. Woods, B. N. Brookhouse, R. A. Cowley, and W. Cochran, *Phys. Rev*. **131**, 1025 (1963).
5. A. D. B. Woods, W. Cochran, and B. N. Brackhouse, *Phys. Rev.* **119**, 980 (1960).

Bibliography

L. Brillouin, *Wave Propagation in Periodic Structures*, 2nd ed., McGraw-Hill, New York (1946).
M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Oxford University Press, London (1954).
W. Cochran, *Dynamics of Atoms in Crystals*, Crane, Russak, New York (1975).
J. De Launay, in: *Solid State Physics*, Vol. 2 (F. Seitz and D. Turnbull, eds.), Academic Press, New York (1956), p. 219.
C. Kittel, *Introduction to Solid State Physics*, 5th ed., Wiley, New York (1976).
A. A. Maradudin, E. W. Montroll, G. H. Weiss, and I. P. Ipatova, Theory of lattice dynamics in the harmonic approximation, in: *Solid State Physics*, 2nd ed., Suppl. 3 (1971).
H. J. McSkimin, *J. Appl. Phys.* **38**, 2362 (1967).
S. S. Mitra, in: *Solid State Physics*, Vol. 13 (F. Seitz and D. Turnbull, eds.), Academic Press, New York (1956) p. 1.
J. M. Ziman, *Electrons and Phonons*, Oxford University Press, London (1960).

3
Semiconductor Statistics

3.1. Introduction

In this chapter we present three basic statistics that are commonly used in the derivation of distribution functions for gas molecules, photons and phonons, electrons in a metal, and electrons and holes in a semiconductor. These basic statistics are needed to deal with the problems of interactions of a large number of particles in a solid. Since a great deal of physical insight can be obtained from statistical analysis of the particle distribution functions in a solid, it is appropriate for us to devote this chapter to finding the distribution functions associated with different statistical mechanics for particles such as gas molecules, photons, phonons, electrons, and holes. The three basic statistics that govern the distribution of particles in a solid are (1) Maxwell–Boltzmann (M-B) statistics, (2) Bose–Einstein (B-E) statistics, and (3) Fermi–Dirac (F-D) statistics. The M-B statistics are also known as the classical statistics, since they apply only to particles with weak interactions among themselves. In the M-B statistics, the number of particles allowed in each quantum state is not restricted by the Pauli exclusion principle. Particles such as gas molecules in an ideal gas system and electrons and holes in a dilute semiconductor are examples that obey the M-B statistics. The B-E and F-D statistics are known as quantum statistics because their distribution functions are derived based on quantum-mechanical principles. Particles that obey the B-E and F-D statistics in general have a much higher density and stronger interaction among themselves than the classical particles. Particles that obey the B-E statistics, such as photons and phonons, are called bosons, while particles that obey the F-D statistics, such as electrons and holes in a degenerate semiconductor or electrons in a metal, are known as fermions. The main difference between the F-D and the B-E statistics is that the occupation number in each quantum state for the fermions is restricted by the Pauli exclusion principle, while bosons are not subjected to the restriction of the exclusion principle. The Pauli exclusion principle states that no more than two particles with opposite spin degeneracy can occupy the same quantum state. Therefore, the total number of particles with the same spin should be equal to or less than the total number of quantum states available for occupancy in a solid.

The M-B statistics and velocity distribution function for ideal gas molecules are depicted in Section 3.2. Section 3.3 presents the F-D statistics, the physical aspect of F-D distribution function, and its derivative for the free-electron case. Section 3.4 depicts the B-E statistics and the distribution function of phonons and photons. The blackbody radiation formula is also presented. Finally, the distribution functions for electrons in the shallow donor states and holes in the shallow acceptor states in the forbidden gap of a semiconductor are presented in Section 3.5.

3.2. Maxwell–Boltzmann Statistics

In this section, the M-B distribution function for classical noninteracting particles such as ideal gas molecules and electrons in an intrinsic or lightly doped semiconductor is derived. Let us first consider an isolated system that contains N distinguishable particles with a total energy of E. Let us then consider the problem of distributing N particles among the q energy levels. If the system consists of $E_1, E_2, \ldots, E_i, \ldots, E_q$ energy levels with $n_1, n_2, \ldots, n_i, \ldots, n_q$ particles in each corresponding energy level, then there are two constraints imposed on these N particles, namely, the conservation of energy and the conservation of particles. These two constraints can be expressed by

$$C_1(n_1, n_2, \ldots, n_i, \ldots, n_q) = N = \sum_{i=1}^{q} n_i, \tag{3.1}$$

$$C_2(n_1, n_2, \ldots, n_i, \ldots, n_q) = E = \sum_{i=1}^{q} n_i E_i. \tag{3.2}$$

Equation (3.1) states that the total number of particles in the system is constant and equal to N, and (3.2) states that the total energy in the system is constant and equal to E. To derive the classical M-B distribution function, the framework of quantum states and energy levels in a noninteracting system is retained and the Pauli exclusion principle is neglected. Therefore, there is no limitation on the number of particles that can be put into a quantum state at a given energy level in the system. In order to derive the distribution function for particles in a noninteracting system, we first analyze the problem of distributing N_1 and N_2 balls in two boxes and then extend this result to the problem of particle distribution in a solid.

If $W(N_1, N_2)$ represents the total number of independent ways of arranging N_1 and N_2 balls in box 1 and box 2, respectively, then $W(0, N_2)$ is the total number of ways of making box 1 empty and box 2 full. There is only one possible arrangement for this case, and hence $W(0, N_2) = 1$. Next, consider the case of arranging 1 ball in box 1 and $(N - 1)$ balls in box 2. In this case, there are N different ways of putting 1 ball in box 1, and hence the total number of ways of arranging 1 ball in box 1 and $(N - 1)$ balls in box 2 is given by $W(1, N_2) = N$. Next, consider the case of arranging 2 balls in box 1 and $(N - 2)$ balls in box 2. The number of ways of arranging the first ball in box 1 is N, and the number of ways of arranging the second ball in box 1 is $(N - 1)$. Thus, the total number of ways of arranging 2 balls

in box 1 and $(N-2)$ balls in box 2 is $W(2, N-2) = N(N-1)/2!$, where 2! is included to account for the permutation between two identical balls. Similarly, one can extend this procedure to the distribution of N_1 balls in box 1 and N_2 balls in box 2. Thus, one can write the total number of ways of arranging N_1 and N_2 balls in boxes 1 and 2 as

$$W(N_1, N_2) = \frac{N(N-1)(N-2)(N-3)\cdots(N-N_1+1)}{N_i!} = \frac{N!}{N_1 N_2!}. \tag{3.3}$$

Extending the above procedure, the total number of ways of arranging $N_1, N_2, N_3, \ldots, N_q$ balls in boxes $1, 2, 3, \ldots, q$ can be expressed by

$$W(N_1, N_2, N_3, \ldots, N_q) = \frac{N!}{N_1!N_2!N_3!\cdots N_q!} = \frac{N!}{\prod_{i=1}^{q} N_i!}, \tag{3.4}$$

where $N = N_1 + N_2 + N_3 + \cdots + N_q$ is the total number of balls available for distribution in q boxes.

The above results can be applied to the derivation of the M-B distribution function for particles in a solid. Next, consider the distribution of n particles among the $E_1, E_2, E_3, \ldots, E_q$ energy levels in a solid. If one assumes that there are $g_1, g_2, g_3, \ldots, g_q$ degenerate quantum states and $n_1, n_2, n_3, \ldots, n_q$ particles in each corresponding energy level $E_1, E_2, E_3, \ldots . E_q$, then the distribution function for the n particles among the q energy levels each with g_i degenerate states is similar to the distribution of N balls in q boxes discussed above. The only difference is that in this case there are g_i additional quantum states in each E_i energy level (where $i = 1, 2, 3, \ldots, q$). If the quantum state in each energy level is nondegenerate, then the total number of ways of arranging $n_1, n_2, \ldots, n_q$ particles in the $E_1, E_2, E_3, \ldots, E_q$ energy levels is given by (3.4). If there are g_i degenerate quantum states in each energy level E_i, then it is necessary to count the total number of ways of arranging n_i particles in each of the g_i quantum states in the E_i energy level. For example, the number of ways of arranging n_1 particles in g_1 quantum states in the energy level E_1 is given by $(g_1)^{n_1}$. Similarly, there are $(g_2)^{n_2}$ ways of arranging n_2 particles among the g_2 quantum states in the E_2 energy level. Therefore, the total number of ways of arranging $n_1, n_2, \ldots, n_q$ particles among the $g_1, g_2, \ldots, g_q$ quantum states in the $E_1, E_2, \ldots, E_q$ energy levels is given by

$$\begin{aligned} W(n_1, n_2, \ldots, n_i, \ldots, n_q) &= \frac{n!(g_1)^{n_1}(g_2)^{n_2}\cdots(g_q)^{n_q}}{(n_1!n_2!\cdots n_q!)} \\ &= n! \prod_{i=1}^{q} \left(\frac{(g_i)^{n_i}}{n_i!} \right). \end{aligned} \tag{3.5}$$

Taking the natural logarithm on both sides of (3.5), one obtains

$$\ln W(n_1, n_2, n_3, \ldots, n_i, \ldots, n_q) = \ln(n!) + \sum_{i=1}^{q} [n_i \ln(g_i) - \ln(n_i!)]. \tag{3.6}$$

Since values of n_i, g_i, and n are much larger than unity, one can employ Stirling's approximation (i.e., $\ln x! \approx x \ln x - x$, for $x \gg 1$) in (3.6), and the result is

$$\ln W \approx (n \ln(n) - n) + \sum_{i=1}^{q} [n_i \ln(g_i) - n_i \ln(n_i) + n_i]. \tag{3.7}$$

From thermodynamics, the entropy of a solid is defined by $S = k_B \ln W$, where k_B is the Boltzmann constant, and $\ln W$ is given by (3.7). Furthermore, the most probable distribution function for particles in a solid can be obtained by maximizing the entropy of the system. Thus, the distribution function for the noninteracting particles described above can be obtained by differentiating (3.7) with respect to n_i for the maximum entropy, which can be carried out using the method of Lagrange multipliers. Using (3.1), (3.2), and (3.7), one obtains

$$\begin{aligned} \frac{\mathrm{d} \ln W}{\mathrm{d}n_i} &= \frac{\mathrm{d}}{\mathrm{d}n_i} \left[n \ln(n) + \sum_{i=1}^{g} \left\{ n_i \ln\left(\frac{g_i}{n_i}\right) \right\} \right] \\ &= \alpha \frac{\mathrm{d}C_1}{\mathrm{d}n_i} + \beta \frac{\mathrm{d}C_2}{\mathrm{d}n_i}, \end{aligned} \tag{3.8}$$

which yields

$$\ln\left(\frac{g_i}{n_i}\right) - 1 = \alpha + \beta E_i, \tag{3.9}$$

where α and β are constants to be determined. From (3.9), the distribution function of the classical particles is defined by

$$f(E_i) = \frac{n_i}{g_i} = \exp[-(1 + \alpha + \beta E_i)]. \tag{3.10}$$

If one drops the index i from (3.10), then the M-B distribution function can be expressed as

$$f(E) = A \exp(-\beta E), \tag{3.11}$$

where A and β are constants, which can be determined from the distribution of gas molecules in an ideal gas system that obeys the M-B statistics. For example, if there are N monatomic gas molecules that interact only through the collision processes, then the energy of such particles is purely kinetic, and may be written as

$$E = \frac{mv^2}{2} = \left(\frac{m}{2}\right)(v_x^2 + v_y^2 + v_z^2). \tag{3.12}$$

Now substituting (3.12) into (3.11), the M-B distribution function for an ideal gas molecule system can also be expressed in terms of the velocity distribution function, which is given by

$$N(v) = 4\pi v^2 A \, \mathrm{e}^{-\beta m v^2/2}. \tag{3.13}$$

The velocity distribution function $N(v)$ given by (3.13) represents the number of particles in the system whose velocities lie in a range $\mathrm{d}v$ about v. Using (3.13), the number of particles with velocities between v and $v + \mathrm{d}v$ in the velocity space is given by

$$\mathrm{d}N = N(v)\,\mathrm{d}v = 4\pi v^2 A\, \mathrm{e}^{-\beta m v^2/2}\mathrm{d}v. \tag{3.14}$$

The total number of particles in the velocity space is obtained by integrating (3.14) from zero to infinity, and the result is

$$N = \int \mathrm{d}N = \int_0^\infty 4\pi v^2 A\, \mathrm{e}^{-\beta m v^2/2}\mathrm{d}v. \tag{3.15}$$

The total energy of the gas molecules in such a system is given by

$$U = \int_0^\infty \left(\frac{mv^2}{2}\right) 4\pi v^2 A\, \mathrm{e}^{-\beta m v^2/2}\mathrm{d}v. \tag{3.16}$$

To find constants A and β from (3.15) and (3.16), the kinetic energy of the gas molecules must be determined first. This energy may also be obtained independently using the equipartition law, which shows that the kinetic energy for each gas molecule in an ideal gas system is equal to $(3/2)k_{\mathrm{B}}T$. Thus, the total energy for an ideal gas system containing N molecules is given by

$$U = \frac{3}{2}Nk_{\mathrm{B}}T. \tag{3.17}$$

Solving (3.15) to (3.17) yields

$$A = N\left(\frac{m}{2\pi k_{\mathrm{B}}T}\right)^{3/2} \quad \text{and} \quad \beta = \frac{1}{k_{\mathrm{B}}T}. \tag{3.18}$$

The velocity distribution function for a classical particle can be obtained by substituting the expressions of A and β given by (3.18) into (3.13), and the result is

$$N(v) = 4\pi N\left(\frac{m}{2\pi k_{\mathrm{B}}T}\right)^{3/2} v^2\, \mathrm{e}^{-mv^2/2k_{\mathrm{B}}T}. \tag{3.19}$$

Figure 3.1 shows the plot of $N(v)$ versus v for three different temperatures. The average velocity for a classical particle that obeys the M-B statistics can be obtained using the expression

$$\langle v\rangle = \frac{\int_0^\infty vN(v)\,\mathrm{d}v}{\int_0^\infty N(v)\,\mathrm{d}v}, \tag{3.20}$$

where $N(v)$ is given by (3.19). A general expression for the average of velocity to the nth power, $\langle v^n\rangle$, in the velocity space is given by

$$\langle v^n\rangle = \frac{\int_0^\infty v^n N(v)\,\mathrm{d}v}{\int_0^\infty N(v)\,\mathrm{d}v}, \tag{3.21}$$

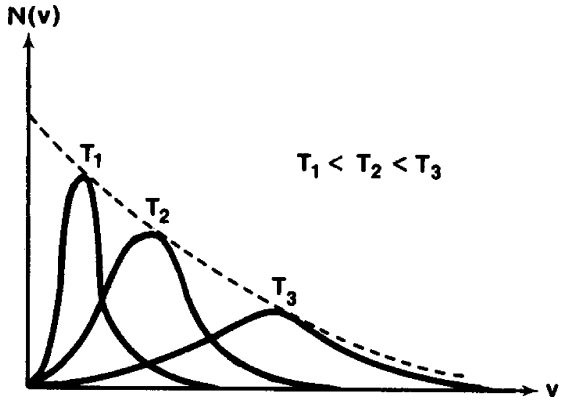


FIGURE 3.1. Maxwell–Boltzmann velocity distribution function for three different temperatures.

where $n = 1, 2, 3, \ldots$. Note that both the average velocity $\langle v \rangle$ and the average kinetic energy ($\langle E \rangle = (1/2)m_0\langle v^2 \rangle$) of a classical particle in the velocity space can be calculated using (3.21). Table 3.1 lists some definite integrals that may be used to calculate the average velocity to the nth power for electrons or holes in a nondegenerate semiconductor or for gas molecules in an ideal gas system.

3.3. Fermi–Dirac Statistics

The M-B statistics described in Section 3.2 are applicable for noninteracting particles, which are assumed to be distinguishable. However, in reality it is usually impossible to distinguish electrons in a metal or in a degenerate semiconductor because of the extremely high density of electrons in these materials ($10^{19} \leq n_0 \leq 10^{23}\,\mathrm{cm}^{-3}$). To apply statistical methods to particles in such a system, an additional constraint imposed by the Pauli exclusion principle from quantum mechanics must be considered. According to the Pauli exclusion principle, no more than one electron with the same spin is allowed per quantum state in a degenerate

TABLE 3.1. Some frequently used integrals for the M-B statistics.

$\int_0^\infty \mathrm{e}^{-ax^2}\,\mathrm{d}x = (\pi/4a)^{1/2} = (\pi k_\mathrm{B}T/2m)^{1/2}$
$\int_0^\infty x\,\mathrm{e}^{-ax^2}\mathrm{d}x = 1/2a = k_\mathrm{B}T/m$
$\int_0^\infty x^2\,\mathrm{e}^{-ax^2}\mathrm{d}x = (\pi/16a^3)^{1/2} = (\sqrt{\pi}/4)(2k_\mathrm{B}T/m)^{3/2}$
$\int_0^\infty x^3\,\mathrm{e}^{-ax^2}\mathrm{d}x = 1/2a^2 = 2(k_\mathrm{B}T/m)^2$
$\int_0^\infty x^4\,\mathrm{e}^{-ax^2}\mathrm{d}x = (3/8a^2)(\pi/a)^{1/2} = (3\sqrt{\pi}/8)(2k_\mathrm{B}T/m)^{5/2}$
$\int_0^\infty x^5\,\mathrm{e}^{-ax^2}\mathrm{d}x = 3/a^3 = 3(2k_\mathrm{B}T/m)^3$
$a = m/2k_\mathrm{B}T$

electron system. The F-D distribution function, which takes into account the Pauli exclusion principle, may be employed to find the electron densities in a metal or in a heavily doped semiconductor. It is interesting to note that several important physical phenomena that cannot be explained properly using the classical M-B statistics at very low temperatures come as a direct result of the F-D statistics.

To derive the F-D distribution function, three basic constraints must be considered. They are the conservation of particles, the conservation of energy, and the Pauli exclusion principle, which are given by

$$C_1(n_1, n_2, \ldots, n_i, \ldots, n_q) = \sum_{i=1}^{q} n_i = n, \tag{3.22}$$

$$C_2(n_1, n_2, \ldots, n_i, \ldots, n_q) = \sum_{i=1}^{q} n_i E_i = E, \tag{3.23}$$

$$n_i \leq g_i, \tag{3.24}$$

where n_i and g_i denote the number of particles and quantum states in the E_i energy level, respectively. The total energy of the system is equal to E, and the total number of particles in the system under consideration is n. Equation (3.24) gives the additional constraint imposed by the Pauli exclusion principle. To derive the F-D distribution function, it is appropriate to first consider the distribution of particles in the E_i energy level. If there are n_i particles and g_i quantum states ($n_i \leq g_i$) in the E_i energy level, then in the g_i quantum states there are g_i ways of arranging the first particle, $(g_i - 1)$ ways of arranging the second particle, $(g_i - 2)$ ways of arranging the third particle, and so on. Thus, the total number of ways of arranging n_i particles in the g_i quantum states in the E_i energy level is given by

$$g_i(g_i - 1)(g_i - 2)\cdots(g_i - n_i + 1) = \frac{g_i!}{(g_i - n_i)!}. \tag{3.25}$$

Since all n_i particles are indistinguishable, permutation among them cannot be counted as independent arrangements. Thus, (3.25) must be modified to account for the permutation of n_i particles, with the result

$$W(n_i) = \frac{g_i!}{n_i!(g_i - n_i)!}. \tag{3.26}$$

Using the above procedure, the total number of independent ways of arranging $n_1, n_2, n_3, n_i, \ldots, n_q$ particles among $g_1, g_2, g_3, g_i, \ldots, g_q$ quantum states in the $E_1, E_2, E_3, E_i, \ldots, E_q$ energy levels, with no more than one particle per quantum state with same spin, is given by

$$\begin{aligned} W(n_1, n_2, \ldots, n_i, \ldots, n_q) &= \frac{g_1!g_2!g_3!\cdots g_q!}{(n_1!n_2!\cdots n_q!)(g_1 - n_1)!(g_2 - n_2)!\cdots(g_4 - n_4)!} \\ &= \prod_{i=1}^{q} \left[\frac{(g_i)!}{n_i!(g_i - n_i)!} \right]. \end{aligned} \tag{3.27}$$

Taking the natural logarithm on both sides of (3.27), one obtains

$$\ln W(n_1, n_2, \ldots, n_q) = \sum_{i=1}^{q} \ln \left[\frac{g_i!}{n_i!(g_i - n_i)!} \right]. \tag{3.28}$$

Now using Stirling's approximation on the right-hand side of (3.28) yields

$$\ln W \approx \sum_{i=1}^{q} [g_i \ln g_i - n_i \ln n_i - (g_i - n_i) \ln(g_i - n_i)]. \tag{3.29}$$

The most probable distribution function of the F-D statistics can be obtained by differentiating (3.29) with respect to n_i, and applying the method of Lagrange multipliers on the two constraints given by (3.22) and (3.23). The result is

$$\frac{\mathrm{d} \ln W}{\mathrm{d}n_i} = \ln \left[\frac{(g_i - n_i)}{n_i} \right] = \eta \frac{\mathrm{d}C_1}{\mathrm{d}n_i} + \beta \frac{\mathrm{d}C_2}{\mathrm{d}n_i} = \eta + \beta E_i. \tag{3.30}$$

From (3.30), one obtains the F-D distribution function, which reads

$$f(E_i) = \frac{n_i}{g_i} = \frac{1}{1 + \mathrm{e}^{(\eta + \beta E_i)}}, \tag{3.31}$$

where $\eta = -E_\mathrm{f}/k_\mathrm{B}T$ is the reduced Fermi energy, E_f is the Fermi energy or chemical potential, and $\beta = 1/k_\mathrm{B}T$. Dropping the index i in (3.31), the F-D distribution function can be expressed by

$$f_0(E) = \frac{1}{1 + \mathrm{e}^{(E - E_\mathrm{f})/k_\mathrm{B}T}}. \tag{3.32}$$

To explain the physical significance of the F-D distribution function given by (3.32), one must refer to Figure 3.2a, which shows the F-D distribution function $f_0(E)$ versus energy E for three different temperatures, and Figure 3.2b, which

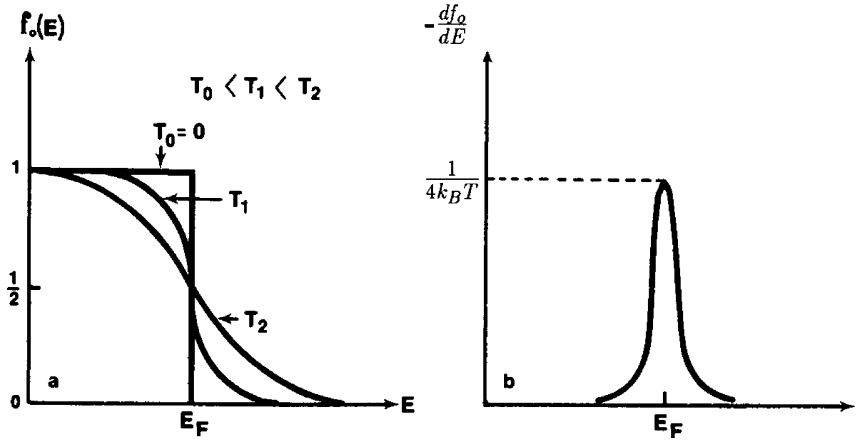


FIGURE 3.2. Fermi–Dirac distribution function $f_0(E)$ and its derivative, $\frac{df_0}{dE}$, versus energy E.

shows a plot of $\mathrm{d}f_0(E)/\mathrm{d}E$ versus E at temperature T. As shown in Figure 3.2a, at $T = 0$ K, $f_0(E) = 1$ for $E < E_\mathrm{f}$, and $f_0(E) = 0$ for $E > E_\mathrm{f}$. This means that the probability of finding a particle with energy smaller than the Fermi energy (i.e., $E < E_\mathrm{f}$) is equal to unity, which implies that all the quantum states below the Fermi level E_f are completely occupied at $T = 0$ K. On the other hand, the probability of finding a particle with energy greater than the Fermi energy (i.e., $E > E_\mathrm{f}$) is zero, which implies that all the quantum states above E_f are empty at $T = 0$ K. These results are in sharp contrast to the results predicted by the classical M-B statistics, which shows that the kinetic energy ($E = \frac{3}{2}k_\mathrm{B}T$) for electrons is zero at $T = 0$ K. The fact that the kinetic energy of electrons is not zero at $T = 0$ K can be explained using the F-D statistics. It is clear from the F-D distribution function that even at $T = 0$ K, electrons will fill all the quantum states up to the Fermi level E_f. In fact, one can show that the average kinetic energy of electrons in a metal at $T = 0$ K is equal to $\frac{3}{5}E_\mathrm{f}(0)$, where $E_\mathrm{f}(0)$ is the Fermi energy at $T = 0$ K. As discussed previously, this result is in sharp contrast to the zero kinetic energy predicted by the classical M-B statistics at $T = 0$ K.

As shown in Figure 3.2a, for $T > 0$ K, $f_0(E) = \frac{1}{2}$ at $E = E_\mathrm{f}$, which shows that the probability of finding an electron at the Fermi level is 50%. For $E < E_\mathrm{f}$, $f_0(E)$ is greater than $\frac{1}{2}$, and f_0 is smaller than $\frac{1}{2}$ for $E > E_\mathrm{f}$. The results show that for $T > 0$ K, the quantum states below the Fermi level are partially empty, and the quantum states above the Fermi level are partially filled. It is these partially filled states (electrons) and partially empty quantum states (holes) that are responsible for the electronic conduction in semiconductors. This can also be explained using Figure 3.2b. It shows that $\mathrm{d}f_0/\mathrm{d}E$ is a delta function centered at E_f, and only those quantum states that are a few $k_\mathrm{B}T$ above and below the Fermi level may contribute to the electrical conduction in a semiconductor or metal.

To find the Fermi energy and the electron density in a metal at $T = 0$ K, it is necessary to know the density-of-states function $g(E)$ in the conduction band. This is due to the fact that the density of electrons in an energy band depends on the availability of the quantum states in that energy band and the probability of a quantum state being occupied by an electron. The density of quantum states for electrons in the conduction band and holes in the valence bands of a semiconductor can be derived using the phonon density-of-states function derived in Section 2.5. The only difference between these two particle systems is that, in deriving the electron density-of-states function, the spin degeneracy (= 2) due to Pauli exclusion principle must be considered. Thus, by taking into account the spin degeneracy of electrons, the density of quantum states per unit energy interval between the constant-energy surfaces of E and $E + \mathrm{d}E$ in k-space can be found from (2.34) by simply replacing the phonon frequency ω by the electron energy E, and the phonon wave vector q by the electron wave vector k, which yields

$$g(E) = 2\left(\frac{1}{2\pi}\right)^3 \int \frac{\mathrm{d}S_k}{|\nabla_k E|}, \tag{3.33}$$

where dS_k is the surface element in k-space and $\nabla_k E$ is the gradient of energy, which is directly related to the group velocity of electrons. The factor 2 on the right-hand side of (3.33) accounts for the spin degeneracy. The density-of-states function for the free-electron case can be derived as follows.

The energy of free electrons is given by

$$E = \frac{\hbar^2 k^2}{2m_0}. \tag{3.34}$$

Thus, using the energy dispersion relation given by (3.34) and assuming a spherical constant-energy surface, one obtains $\nabla_k E = \hbar^2 k/m_0$, and the surface integral in k-space is given by $\int dS_k = 4\pi k^2$. Now, substituting these two expressions into (3.33), one obtains the density-of-states function per unit volume as

$$g(E) = \left(\frac{1}{4\pi^3}\right)\frac{4\pi k^2}{\hbar^2 k/m_0} = \left(\frac{4\pi}{h^3}\right)(2m_0)^{3/2}E^{1/2}. \tag{3.35}$$

Equation (3.35) shows that the density-of-states function $g(E)$ for the free-electron case with a parabolic energy band is proportional to the square root of the energy, as illustrated in Figure 3.3. This result can also be applied to describe the density-of-states functions in the conduction and valence bands of a semiconductor provided that the free-electron mass used in (3.35) is replaced by either the electron density-of-states effective mass ($m_{dn}{}^*$) in the conduction band or hole density-of-states effective mass ($m_{dp}{}^*$) in the valence bands.

The equilibrium electron density in a metal can be calculated using (3.32) and (3.35). The density of electrons with energy between E and $E + dE$ in the

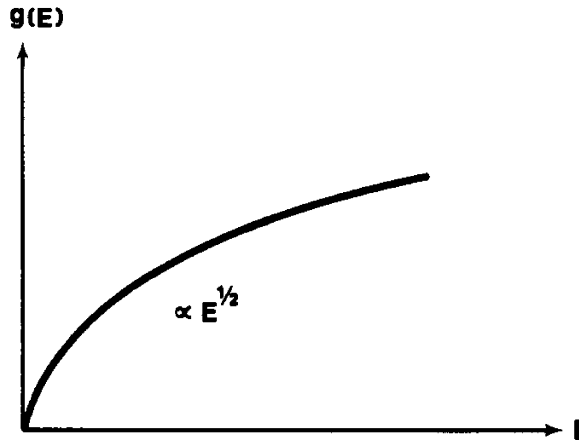


FIGURE 3.3. Density of quantum states $g(E)$ versus energy E for the free-electron case.

conduction band can be expressed by

$$\mathrm{d}n = f_0(E)g(E)\,\mathrm{d}E = \left(\frac{4\pi}{h^3}\right)(2m_0)^{3/2}\frac{E^{1/2}\,\mathrm{d}E}{1+\mathrm{e}^{(E-E_\mathrm{f})/k_\mathrm{B}T}}. \tag{3.36}$$

Thus, the total density of electrons in a metal can be obtained by integrating (3.36) from $E = 0$ to $E = \infty$, which yields

$$n_0 = \int \mathrm{d}n = \left(\frac{4\pi}{h^3}\right)(2m_0)^{3/2}\int_0^\infty \frac{E^{1/2}\,\mathrm{d}E}{1+\mathrm{e}^{(E-E_\mathrm{f})/k_\mathrm{B}T}}. \tag{3.37}$$

At $T = 0$ K, (3.37) reduces to

$$n_0 = \int_0^{E_\mathrm{f}(0)} \left(\frac{4\pi}{h^3}\right)(2m_0)^{3/2}E^{1/2}\,\mathrm{d}E, = \left(\frac{8\pi}{3h^3}\right)(2m_0)^{3/2}E_\mathrm{f}(0)^{3/2}. \tag{3.38}$$

The Fermi energy at $T = 0$ K can be obtained by solving (3.38), and the result is

$$E_\mathrm{f}(0) = \left(\frac{h^2}{8m_0}\right)\left(\frac{3n_0}{\pi}\right)^{2/3}. \tag{3.39}$$

Note that in (3.39), n_0 is the free-electron density, m_0 is the free-electron mass, and h is Planck's constant. Now, substituting the value of $n_0 = 1 \times 10^{22}$ cm^{-3} and $m_0 = 9.1 \times 10^{-31}$ kg into (3.39), one finds that the Fermi energy is given by $E_\mathrm{f}(0) = 10$ eV at $T = 0$ K for a metal. This implies that the average kinetic energy of electrons can still be quite large even at $T = 0$ K (see Problem 3.3). This result is in sharp contrast with the prediction given by the classical M-B statistics, which predicts that the average kinetic energy of the classical particles is equal to zero at $T = 0$ K.

In a metal, both the electron density and the Fermi energy depend very weakly on the temperature. For example, using the F-D statistics, one can show that the first-order correction in the temperature dependence of the Fermi energy is given by

$$E_\mathrm{f}(T) \approx E_\mathrm{f}(0)\left[1 - \frac{\pi^2}{12}\left(\frac{T}{T_\mathrm{f}}\right)^2\right], \tag{3.40}$$

where $E_\mathrm{f}(0)$ is the Fermi energy at $T = 0$ K, given by (3.39), and $T_\mathrm{f} = E_\mathrm{f}(0)/k_\mathrm{B}$ is the Fermi temperature. Using (3.40) and $E_\mathrm{f}(0) = 10$ eV, the Fermi temperature T_f was found to be 11,600 K. Thus, it is clear that the Fermi energy in a metal indeed depends very weakly on temperature. In fact, if the values of $E_\mathrm{f}(0)$ and T_f given above were used in (3.40), the second term in the square brackets of (3.40) would indeed be negligible for a metal. On the contrary, both the Fermi energy and the carrier density for an intrinsic or lightly doped semiconductor are in general a strong function of temperature. This will be discussed further in Chapter 5.

3.4. Bose–Einstein Statistics

In this section the distributon function for photons and phonons is derived using the B-E statistics. The general characteristics of photons and phonons include the following: (i) they are indistinguishable and each is with a quantized energy value and wave number, and (ii) the occupation number in each quantum state for these particles is not restricted by the Pauli exclusion principle. To derive the B-E distribution function, consider a linear array of n_i particles and $(g_i - 1)$ partitions, which are necessary to divide these particles into g_i quantum states. It is not difficult to see that the number of ways of arranging the n_i particles among g_i quantum states is equal to the number of independent permutations of particles and partitions. Since there are a total of $(n_i + g_i - 1)$ particles plus partitions, they can be arranged linearly in $(n_i + g_i - 1)!$ ways. However, permutations of particles among themselves or partitions among themselves do not count as independent arrangements; one must take into account the number of ways of permuting particles among themselves (i.e., $n_i!$) and the number of ways of permuting partitions among themselves ($(g_i - 1)!$). Therefore, the total number of independent ways of arranging n_i particles among g_i quantum states in the E_i energy level can be written as

$$W_i = \frac{(n_i + g_i - 1)!}{n_i!(g_i - 1)!}. \tag{3.41}$$

In contrast to the F-D statistics, n_i can be greater than g_i in the B-E statistics. Therefore, the total number of independent ways of arranging $n_1, n_2, \ldots, n_i, \ldots, n_q$ particles among $E_1, E_2, E_i, \ldots, E_q$ energy levels with $g_1, g_2, g_i, \ldots, g_q$ quantum states pertaining to each corresponding energy level is given by

$$\ln W(n_1, n_2, \ldots, n_q) = \sum_{i=1}^{q} \ln \left[\frac{(n_i + g_i - 1)!}{n_i!(g_i - 1)!} \right]. \tag{3.42}$$

Now differentiating (3.42) with respect to n_i and using Stirling's approximation and Lagrange multipliers, one obtains

$$\frac{\mathrm{d} \ln W}{\mathrm{d}n_i} \approx \ln \left[\left(\frac{g_i}{n_i} \right) + 1 \right] = \alpha + \beta E_i. \tag{3.43}$$

From (3. 43), the B-E distribution function can be expressed as

$$f(E_i) = \frac{n_i}{g_i} = \frac{1}{(e^{\alpha + \beta E_i} - 1)}, \tag{3.44}$$

where α and β are constants to be determined. In the B-E statistics, since the number of particles in each quantum state is not restricted by the Pauli exclusion principle, the constant α can be set equal to zero, and β is equal to $1/k_{\mathrm{B}}T$. Thus, dropping the index i in (3.44), the B-E distribution function can be expressed as

$$f(E) = \frac{1}{(\mathrm{e}^{\hbar\omega/k_{\mathrm{B}}T} - 1)}, \tag{3.45}$$

where $E = h\nu = \hbar\omega$ is the energy of a photon or a phonon, h is Planck's constant, and ν is the phonon or photon frequency. It is important to compare the B-E distribution function given by (3.45) with Planck's blackbody radiation formula, which describes the photon distribution over a wide range of optical spectrum. According to Planck's blackbody radiation law, the number of photons per unit volume having frequencies between ν and $\nu + d\nu$ is given by

$$Q_{eq}d\nu = \frac{(8\pi\nu^2/c^3)\,d\nu}{(e^{h\nu/k_BT} - 1)}. \tag{3.46}$$

Comparing (3.45) and (3.46) one finds that the denominators in both equations are identical, implying that the distribution function for photons given by Planck's blackbody radiation law is indeed identical to the distribution function for the photons and phonons given by the B-E statistics. It is noted that similar to the F-D distribution function, the B-E distribution function will reduce to the classical M-B distribution function at high temperatures (i.e., for $h\nu \geq 4k_BT$) at which the lattice phonons are dominated by the optical phonons.

3.5. Statistics for the Shallow-Impurity States in a Semiconductor

The F-D distribution function derived in Section 3.3 is applicable to electrons in the conduction band and holes in the valence bands of a semiconductor. However, the distribution of electrons in the shallow-donor states or holes in the shallow-acceptor states in the forbidden gap of a semiconductor is somewhat different from that in the conduction or the valence band states. The difference stems mainly from the fact that in the forbidden gap a shallow-donor state can be occupied by one electron with either spin-up or spin-down. Once a donor impurity state is occupied by one electron, it becomes a neutral donor state. As a result, no additional electron is allowed to occupy this donor impurity state. The distribution function of electrons in the shallow-donor state can be derived as follows.

Consider a hydrogenic shallow-donor impurity state located a few k_BT below the conduction band edge with density N_d and ionization energy E_d. If W_d is the total number of ways of arranging n_d electrons in the N_d donor states with one electron per state, then there are $2N_d$ ways of putting the first electron in the N_d donor states, taking into account the spin degeneracy (2) of electrons. Similarly, the number of ways of arranging the second electron in the $(N_d - 1)$ donor states is equal to $2(N_d - 1)$, since there are only $(N_d - 1)$ empty donor states available for electron occupation This procedure can be repeated either until all the n_d electrons are filled up or until there are only $2(N_d - n_d + 1)$ ways of arranging the last electron in the remaining donor impurity state. Since electrons are indistinguishable, the permutations among them do not count as independent arrangements. Therefore,

the total number of ways of distributing n_d electrons among N_d donor states is given by

$$W_d = 2(N_d)2(N_d - 1)2(N_d - 2)\cdots 2(N_d - n_d + 1)/n_d! = \frac{2^{n_d} N_d!}{n_d!(N_d - n_d)!}. \tag{3.47}$$

Comparing (3.47) and (3.26) reveals that an extra factor of 2^{n_d} is included in (3.47) to account for the Coulombic nature of the shallow-donor impurity states. Applying the same procedure as described in the previous sections to (3.47), the distribution function of electrons in a shallow-donor state can be written as

$$f(E_d) = \frac{n_d}{N_d} = \frac{1}{1 + \frac{1}{2}\, e^{(E_d - E_f)/k_B T}}. \tag{3.48}$$

From (3.48), it is noted that a degenerate factor of $\frac{1}{2}$ appears in the exponential term of the denominator. This degenerate factor is to account for the Coulombic interaction between the electron and the ionized donor impurity state. A more generalized expression for the electron distribution function in a shallow-donor state can be obtained by replacing the degenerate factor $\frac{1}{2}$ in (3.48) by g_D^{-1}, where g_D is the degeneracy factor of the shallow-donor state. Values ofg_d may vary between 2 and 12 depending on the nature of the shallow-donor states and the conduction band structure of a semiconductor. For semiconductors with a single spherical conduction band, $g_d = 2$, and for the multivalley semiconductor such as silicon, g_d is equal to 12.

The above derivation can also be applied to find the distribution function of holes in the shallow-acceptor states. Using a similar procedure for electrons, it can be shown that the distribution function of holes in a shallow-acceptor state is given by

$$f(E_a) = \frac{p_a}{N_a} = \frac{1}{1 + g_A\, e^{(E_f - E_a)/k_B T}}. \tag{3.49}$$

Equation (3.49) is the distribution function for holes in the shallow-acceptor states, which can be used in calculating the hole density in the shallow-acceptor states of a semiconductor. Note that g_A is the degenerate factor for the acceptor states; for the valence bands with a heavy-hole band and a light-hole band, the value of g_A is equal to 4.

The M-B and F-D distribution functions and the density-of-states function derived in this chapter are very important for calculating the density of electrons and holes in a semiconductor. While the M-B statistics are applicable to both the ideal gas system and nondegenerate semiconductors, the F-D statistics are used mainly for metals and degenerate semiconductors. On the other hand, the B-E statistics are used primarily for calculating the average energy and population distribution of phonons and photons.

Problems

3.1. Using (3.20), show that the average velocity of a classical particle is given by

$$\langle v \rangle = (8k_B T/\pi m_0)^{(1/2)}.$$

3.2. Using the M-B statistics, find the average kinetic energy of classical particles. What is the root-mean-square value of the velocity? The average kinetic energy of a classical particle can be obtained using the following equation:

$$\langle E \rangle = \frac{\int_0^\infty E g(E) \exp(-E/k_B T) \mathrm{d}E}{\int_0^\infty g(E) \exp(-E/k_B T) \mathrm{d}E},$$

where $g(E)$ is the density-of-states function given by (3.35).

3.3. Using the F-D statistics and Problem 3.2, derive a general expression for the average kinetic energy of electrons, and show that at $T = 0$ K, the average energy of electrons $\langle E \rangle$ is equal to $(\frac{3}{5})E_f(0)$. Compare the average kinetic energies predicted by both the M-B and F-D statistics at $T = 0$ K, and explain the physical meanings of their difference.

3.4. Calculate the Fermi energy (in eV) for electrons in sodium and copper at $T = 0$ K. Assuming one electron per atom, calculate the equivalent Fermi temperature for each case.

3.5. Plot the F-D distribution function $f_0(E)$ and its derivative $\partial f_0(E)/\partial E$ as a function of electron energy for $T = 0$, 300, 600, and 1000 K.

3.6. Derive (3.40). *Hint*: The temperature dependence of the Fermi energy for a metal can be derived using the integral.

$$I = \int_0^\infty f_0(E)(\mathrm{d}G(E)/\mathrm{d}E)\, \mathrm{d}E,$$

where $f(E)$ is the F-D distribution function and $G(E)$ is a well-behaved function that vanishes at $E = 0$. Using integration by parts and Taylor's series expansion in $G(E)$ with respect to $E = E_f$, the above integral reduces to

$$I = G(E_f) + \left(\frac{\pi^2}{6}\right)(k_B T)^2 \left(\frac{\mathrm{d}^2 G(E)}{\mathrm{d}E^2}\right)|_{E_f} + \cdots,$$

where

$$G(E) = \int_0^E g(E)\, \mathrm{d}E, \quad \frac{\mathrm{d}G(E)}{\mathrm{d}E} = g(E), \qquad \frac{\mathrm{d}^2 G(E)}{\mathrm{d}E^2} = \frac{\mathrm{d}g(E)}{\mathrm{d}E}.$$

Here $g(E)$ is the density-of-states function given by (3.35). It is noted that the temperature dependence of free-electron density n_0 in a metal can be

expressed by

$$n_0 = \int_0^{\infty} f_0(E)g(E)\,\mathrm{d}E$$
$$= \int_0^{E_\mathrm{f}} g(E)\,\mathrm{d}E + \left(\frac{\pi^2}{6}\right)(k_\mathrm{B}T)^2\left(\frac{\mathrm{d}g(E)}{\mathrm{d}E}\right)_{E_\mathrm{f}}.$$

At $T = 0$ K,

$$n_0 = \int_0^{E_\mathrm{f}(0)} g(E)\,\mathrm{d}E.$$

3.7. Using the results given in Problem 3.6, show that the electronic specific heat for a metal is given by

$$C_\mathrm{e} = \frac{\mathrm{d}U}{\mathrm{d}T} = \left(\frac{\pi^2 n_0 k_\mathrm{B}}{2}\right)\left(\frac{T}{T_\mathrm{f}}\right),$$

where U is the average energy of electrons at any given temperature, given by

$$U = \int_0^{\infty} Eg(E)f(E)\,\mathrm{d}E = U_0 + \left(\frac{\pi^2}{6}\right)(k_\mathrm{B}T)^2 g(E_\mathrm{f}(0)).$$

3.8. Derive (3.48) and (3.49).

Bibliography

E. Band, *An Introduction to Quantum Statistics*, D. Van Nostrand, Princeton, NJ (1955).
J. S. Blakemore, *Semiconductor Statistics*, Pergamon Press, New York (1960).
R. W. Gurney, *Introduction to Statistical Mechanics*, McGraw-Hill, New York (1949).
J. P. McKelvey, *Solid State and Semiconductor Physics*, Chapter 5, Harper & Row, New York (1966).
R. C. Tolman, *The Principles of Statistical Mechanics*, Oxford University Press, London (1938).
S. Wang, *Solid State Electronics*, Chapter 1, McGraw-Hill, New York (1966).

4
Energy Band Theory

4.1. Introduction

In this chapter the one-electron energy band theories for crystalline solids are presented. The importance of energy band theories for a crystalline solid is due to the fact that many important physical and optical properties of a solid can be readily explained using its energy band structure. In general, the energy band structure of a solid can be constructed by solving the one-electron Schrödinger equation for electrons in a crystalline solid that contains a large number of interacting electrons and atoms. To simplify the difficult task of solving the Schrödinger equation for the many-body problems in a crystal, the effects that arise from the motion of atomic nuclei must be neglected (i.e., it is assumed that the nuclei are at rest in the equilibrium positions at each lattice site). Under this condition, the nuclear coordinates enter the problem only as a constant parameter. However, even though the problem is confined as a purely electronic one, there are still the many-electron problems in the system that cannot be solved explicitly. Therefore, it is necessary to apply additional approximations in solving the Schrödinger equation for electrons in a crystalline solid.

One of the most fruitful methods developed for solving the many-electron problems in a crystal is the one-electron approximation. In this method the total wave functions of electrons are chosen as a linear combination of the individual wave functions in which each wave function involves only the coordinates of one electron. It is this approximation that forms the basic framework for calculating the energy band structure of a solid. This method can be described by assuming that each electron sees, in addition to the potential of the fixed charges (i.e., positive ions), only some average potential due to the charge distribution of the rest of the electrons in the solid. Therefore, the movement of each electron is essentially independent of the other electrons throughout the crystal lattice. By means of the one-electron approximation, the solution of the many-electron problems is reduced to (1) finding equations that are satisfied by the one-electron wave functions and (2) obtaining adequate solutions for the electron wave functions and electron energies in the crystal under consideration.

Section 4.2 presents the basic quantum concepts and wave mechanics that are essential for dealing with systems of atomic scale and for solving the electron wave functions and energy band structures in crystalline solids. Section 4.3 describes the basic constraints imposed on the electron wave functions that are attributed to the translational symmetry of the periodic crystal. For example, suitable electron wave functions in a crystal must obey the Bloch theorem. According to this theorem, the electron wave functions in a periodic crystal consist of a plane wave modulated by a Bloch function that has the same periodicity as the crystal potential. Section 4.4 depicts the Kronig–Penney model for the one-dimensional (1-D) periodic crystal lattice. Section 4.5 describes the nearly free electron (NFE) approximation for a three-dimensional (3-D) crystal lattice. The NFE method can be used to find the electronic energy states for the outer-shell valence electrons in which the periodic potential of the crystal can be treated as a small perturbation. Section 4.6 presents the tight-binding approximation (or a linear combination of atomic orbits (LCAO)). The LCAO method may be employed to calculate the electronic states for the inner shell core electrons in a crystalline solid. The solutions of Schrödinger equations and the density of states functions for low-dimensional systems (0-D, 1-D, 2-D, quasi-1-D, and quasi-2-D) will also be discussed in this section. Section 4.7 describes the energy band structures for some elemental and compound semiconductors. In general, the calculations of energy band structures for semiconductors are carried out using more rigorous and sophisticated methods than those described in this chapter. The effective mass concept for electrons and holes in a semiconductor is presented in Section 4.8.

4.2. Basic Quantum Concepts and Wave Mechanics

In this section several important historical experimental observations dealing with blackbody radiation, optical spectra emitted by atoms, and the wave like nature of particles that could not be explained by the classical mechanics, and the success of quantum mechanics in describing the behavior of systems with atomic dimensions will be discussed.

4.2.1. Planck Blackbody Radiation Formula

For an ideal radiator, called the blackbody, the spectrum or the wavelength dependence of the emitted radiation is described by Planck's blackbody radiation law. Various attempts to explain the observed blackbody radiation spectrum were made in the later half of the nineteenth century. Rayleigh and Jeans first proposed the concept of blackbody radiation, based on classical mechanics, that the heat absorbed by a material should cause vibration of atoms within the solid. The vibrating atoms were modeled as harmonic oscillators with a spectrum of normal mode frequency $\nu = \omega/2\pi$, and a continuum of allowed energies distributed in accordance with statistical considerations. The emitted radiation was in essence equated to a sampling of the energy distribution inside the solid. This classical

theory was in good agreement with experimental observation at long wavelengths but failed at short wavelengths. In 1901, Max Planck provided a detailed theoretical fit to the observed blackbody spectrum. The explanation was based on the hypothesis that the vibrating atoms in a material could radiate or absorb energy only in discrete packets. Specifically, for a given atomic oscillator vibrating at a frequency v, Planck postulated that the energy of the oscillator was restricted to the quantized values

$$E_n = nhv = n\hbar\omega, \quad n = 0, 1, 2, 3 \ldots, \tag{4.1}$$

where $h = 6.628 \times 10^{-34}$ J/s ($\hbar = h/2\pi$) is Planck constant. Planck's blackbody radiation formula for describing the photon emission spectra is given by

$$S(v) = \int Q_{\text{eq}} \mathrm{d}v = \int \frac{(8\pi v^2/c^3)\mathrm{d}v}{(\mathrm{e}^{hv/k_{\text{B}}T} - 1)}. \tag{4.2}$$

It is noted from the above equation that for atomic-dimension systems the classical view, which always allows a continuum of energies, is experimentally incorrect. Extremely small discrete steps in energy, or energy quantization, can occur in a photon, and is a central feature of quantum mechanics. A comparison of the blackbody radiation formula given by (4.2) with the Bose–Einstein (B-E) distribution function given by (3.45) reveals that blackbody radiation indeed obeys the B-E statistics.

4.2.2. Bohr Model for the Hydrogen Atom

Another experimental observation that puzzled scientists in the nineteenth century was the sharp, discrete spectral lines emitted by heated gases. In 1913, Niels Bohr proposed a model explaining the discrete nature of the spectra emitted by heated gases. Building on Planck's hypothesis and Rutherford's atomic model, Bohr suggested that the electrons in an atom are restricted to certain well-defined orbits, or, equivalently, assumed that the orbiting electrons could take on only certain (quantized) values of angular momentum L.

For the simple hydrogen atom with $Z = 1$ and a circular electron orbit, the Bohr postulate of angular momentum can be expressed by

$$L_n = m_0 v r_n = n\hbar, \quad n = 1, 2, 3 \ldots, \tag{4.3}$$

Since the electron orbits are assumed stable, the centripetal force on the electron must be balanced by the Coulomb attractive force. Thus, one obtains

$$\frac{m_0 v^2}{r_n} = \frac{q^2}{4\pi\varepsilon_0 r_n^2}. \tag{4.4}$$

Solving (4.3) and (4.4) yields

$$r_n = \frac{4\pi\varepsilon_0 (n\hbar)^2}{m_0 q^2}. \tag{4.5}$$

Next, the total energy of an electron (E_n) is equal to the sum of kinetic energy (K.E.) and potential energy (P.E.). The kinetic energy is given by

$$\text{K.E.} = \frac{1}{2} m_0 v^2 = \frac{1}{2} \frac{q^2}{4\pi\varepsilon_0 r_n}, \tag{4.6}$$

and the potential energy is given by

$$\text{P.E.} = \frac{-q^2}{4\pi\varepsilon_0 r_n}. \tag{4.7}$$

Note that the potential energy vanishes for $r_n \to \infty$.

From (4.6) and (4.7), the total electron energy is given by

$$E_n = \text{P.E.} + \text{K.E.} = \frac{1}{2} \frac{-q^2}{4\pi\varepsilon_0 r_n}. \tag{4.8}$$

Substituting r_n given by (4.5) into (4.8) yields

$$E_n = -\frac{m_0 q^4}{2(4\pi\varepsilon_0 n\hbar)^2} = \frac{-13.6}{n^2}\,(\text{eV}), \tag{4.9}$$

where $n = 1, 2, 3\ldots$. Equation (4.9) shows that the ionization energy of the first Bohr orbit with $n = 1$ is $E_1 = -13.6$ eV. The allowed energy transitions in the hydrogen atom as predicted by Bohr's model are found in excellent agreement with the obsverved spectral lines. Although Bohr's model given by (4.9) successfully predicted the hydrogen spectrum, the model failed to predict the spectra of more complex atoms such as helium. Nevertheless, the Bohr theory reinforced the concept of energy quantization and failure of the classical mechanics in dealing with systems on an atomic scale. The quantization of angular momentum in Bohr's model clearly extended the quantum concept in dealing with systems of atomic dimensions.

4.2.3. The Wave–Particle Duality

In 1925 de Broglie suggested that since electromagnetic radiation (waves) exhibited particle-like (photon) properties, particles should also exhibit wave like properties. De Broglie further hypothesized that, parallel to the photon momentum calculation, the wavelength characteristic of a given particle with momentum p can be calculated from $p = h/\lambda$, where λ is the wavelength of the electromagnetic radiation. Based on de Broglie's wave–particle duality hypothesis, the momentum of a particle (or wave) can be written as

$$p = m_0 v = \hbar k. \tag{4.10}$$

Although pure conjecture at the time, the de Broglie hypothesis was quickly confirmed by the well-established fact of the wave–particle duality of electromagnetic radiation.

Based on the experimental evidence of blackbody radiation, the Bohr atom, and the wave–particle duality, one is led to the conclusion that classical mechanics

does not accurately describe the action of particles in systems of atomic dimension. Experiments point to a quantization of observables (energy, angular momentum, etc.) and to the inherent wave like nature of all matter.

4.2.4. Schrödinger Equations

In 1926, Schrödinger established a unified scheme valid for describing both the microscopic and macroscopic universes. The formulation, called wave mechanics, which incorporated the physical notion of quantization first advanced by Planck and the wave like nature of matter hypothesized by de Broglie, was subsequently developed by Schrödinger to treat the electron systems in crystalline materials. There are five basic postulates in Schrödinger wave mechanics for a single-particle system:

(a) There exists a wave function $\Psi = \Psi(r, t)$, where $r = x, y, z$, from which one can ascertain the dynamic behavior of the system and all desired system variables. Note that $\psi(r, t)$ may be a complex quantity with real and imaginery parts and is in general a function of space coordinates, $r = x, y, z$, and time, t.
(b) The wave function $\psi(r, t)$ for a given system and specified system constraints is determined by solving the time-dependent Schrödinger equation, which is given by

$$\frac{-\hbar^2}{2m}\nabla^2\Psi + V(r)\Psi = -\frac{\hbar}{i}\frac{\partial\Psi}{\partial t}, \tag{4.11}$$

where $V(r)$ is the potential energy of the system, and $i = \sqrt{-1}$.
(c) Both ψ and $\nabla\psi$ must be finite, continuous, and single-valued for all values of r and t.
(d) If ψ^* is the conjugate of ψ, then $\psi^*\psi \mathrm{d}r^3$ represents the probability that the particle will be found in the volume element $\mathrm{d}r^3$. Thus,

$$\int_v \Psi^*\Psi \mathrm{d}r^3 = 1. \tag{4.12}$$

Equation (4.12) implies that the probability of finding a particle over the entire space is equal to unity.
(e) The expectation value of a system variable such as momentum p and position r can be calculated from the mathematical operator

$$\langle\alpha\rangle = \int \Psi\alpha\Psi^* \mathrm{d}r^3. \tag{4.13}$$

To deal with electrons in crystalline solids, the time-independent Schrödinger equation is used to solve the electron wave functions and energy states in such solids. If the electron in the crystal under consideration has a fixed total energy E, then the quantum-mechanical formulation of the problem can be greatly simplified. Using (4.13), the expectation value of energy $\langle E\rangle$ is equal to a constant E, and the

right-hand side of (4.11) becomes

$$-\frac{i}{\hbar}\frac{\partial \psi}{\partial t} = E\psi. \tag{4.14}$$

Using the separation of variables method, the time-dependent wave functions $\psi(r, t)$ given in (4.11) can be expressed in terms of the product of the time-varying phase factor $e^{-iEt/\hbar}$ and the spatially dependent wave functions $\phi(r)$ as

$$\psi(r, t) = \phi(r)e^{-iEt/\hbar}. \tag{4.15}$$

Now, substituting (4.14) and (4.15) into (4.11), one obtains

$$-\left(\frac{\hbar^2}{2m}\right)\nabla^2\phi(r) + V(r)\phi(r) = E\phi(r). \tag{4.16}$$

Equation (4.16) is called the time-independent Schrödinger equation, and $\phi(r)$ is a function only of the space coordinate, r. This time-independent Schrödinger equation is the basis for solving the one-electron energy band theory and related problems in crystalline materials.

4.3. The Bloch–Floquet Theorem

The Bloch–Floquet theorem states that the most generalized solution for a one-electron time-independent Schrödinger equation in a periodic crystal lattice is given by

$$\phi_k(r) = u_k(r)e^{ik\cdot r}, \tag{4.17}$$

where $u_k(r)$ is the Bloch function, which has the spatial periodicity of the crystal potential, and $k(=2\pi/\lambda)$ is the wave vector of the electron. The one-electron time-independent Schrödinger equation for which $\phi_k(r)$ is a solution is given by (4.16) and can be rewritten as

$$-\left(\frac{\hbar^2}{2m}\right)\nabla^2\phi_k(r) + V(r)\phi_k(r) = E_k\phi_k(r), \tag{4.18}$$

where $V(r)$ is the periodic crystal potential, which arises from the presence of ions at their regular lattice sites, and has the periodicity of the crystal lattice given by

$$V(r) = V(r + \boldsymbol{R}_j). \tag{4.19}$$

Note that $\boldsymbol{R}_j$ is the translational vector in the direct lattice defined by (1.3). To prove the Bloch theorem, it is necessary to consider the symmetry operation, which translates an eigenfunction in a periodic crystal lattice via the translational vector $\boldsymbol{R}_j$. This translational operation can be expressed by

$$T_j f(r) = f(r + \boldsymbol{R}_j). \tag{4.20}$$

The periodicity of a crystal lattice can be verified from the fact that $f(r)$ is invariant under the symmetry operations of T_j. Since the translational operator T_j commutes

with the Hamiltonian H, it follows that

$$T_j H \phi_k = H T_j \phi_k. \tag{4.21}$$

Since ϕ_k is an eigenfunction of T_j, one may write

$$T_j \phi_\kappa(r) = \phi_k(r + \boldsymbol{R}_j) = \sigma_j \phi_k(r), \tag{4.22}$$

where σ_j is a phase factor and an eigenvalue of T_j. The phase factor σ_j can be expressed by

$$\sigma_j = \mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{R}_j}, \tag{4.23}$$

where $\boldsymbol{k}$ is the wave vector of electrons, which can be a complex number in a periodic crystal. If one performs two successive translational operations (i.e., $T_j T_i$) on the wave function ϕ_k, one obtains from (4.22) and (4.23) the following relationship:

$$T_j T_i \phi_k = T_j \sigma_i \phi_k = \mathrm{e}^{ik\cdot(R_i + R_j)} \phi_k(r). \tag{4.24}$$

From (4.17), the Bloch function $u_k(r)$ can be written as

$$u_k(r) = \mathrm{e}^{-ik\cdot r} \phi_k(r). \tag{4.25}$$

Now solving (4.22), (4.24), and (4.25), one obtains

$$\begin{aligned} T_j u_k(r) &= u_k(r + R_j) = T_j \left[\mathrm{e}^{-ik\cdot r} \phi_k(r)\right] \\ &= \mathrm{e}^{-ik\cdot(r+R_j)} T_j \phi_k(r) = \mathrm{e}^{-ik\cdot(r+R_j)} \mathrm{e}^{ik\cdot R_j} \phi_k(r) \\ &= \mathrm{e}^{-ik\cdot r} \phi_k(r) = u_k(r). \end{aligned} \tag{4.26}$$

Thus, from the symmetry operations given by (4.26) one obtains

$$u_k(r + R_j) = u_k(r), \tag{4.27}$$

which shows that the Bloch function $u_k(r)$ has indeed the same periodicity in space as the crystal potential $V(r)$. Therefore, the general solution of (4.18) is given by (4.17). From (4.17), it is noted that the electron wave function in a periodic crystal lattice is a plane wave modulated by the Bloch function. The Bloch function $u_k(r)$ is invariant under translation. It should be pointed out here that the exact shape of $u_k(r)$ depends on the electron energy E_k and the crystal potential $V(r)$ of a crystalline solid. Thus, the Bloch theorem described in this section can be applied to solve the electron wave functions and energy band structures (i.e., E_k vs. k relation) for the crystalline solids with periodic potential.

4.4. The Kronig–Penney Model

In this section, the one-electron Schrödinger equation is used to solve the electron wave functions and energy states for a one-dimensional (1-D) periodic lattice. The periodic potential $V(x)$ for the 1-D lattice is shown in Figure 4.1a. The

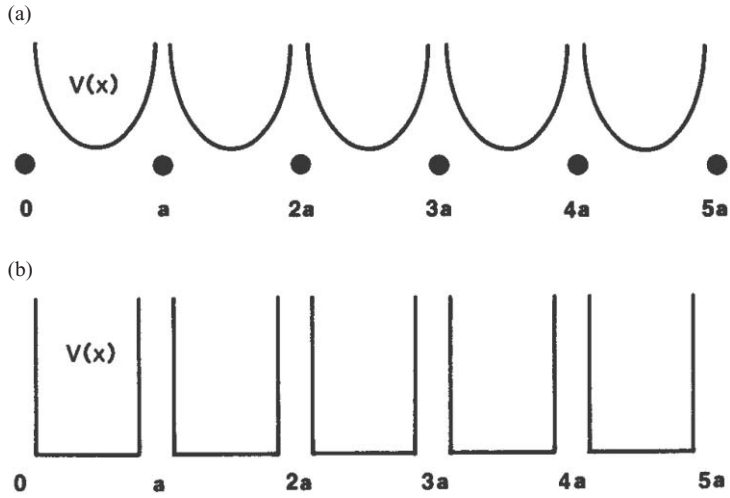


FIGURE 4.1. (a) A one-dimensional (1-D) periodic potential distribution. (b) The Kroning–Penney model for a 1-D periodic lattice.

Kronig–Penney model shown in Figure 4.1b is used to replace the periodic potential of a 1-D crystal lattice with a delta function at each lattice site. In this model, it is assumed that $V(x)$ is zero everywhere except at the atomic site, where it approaches infinity in such a way that the integral of $V(x)\,\mathrm{d}x$ over the potential barrier remains finite and equal to a constant C. Inside the potential barrier, the electron wave functions must satisfy the one-electron Schrödinger equation, which is given by

$$\left(\frac{\hbar^2}{2m}\right)\frac{\mathrm{d}^2\phi_k}{\mathrm{d}x^2} + [E - V(x)]\phi_k = 0, \tag{4.28}$$

where $V(x)$ is the periodic potential with period a. According to the Bloch–Floquet theorem discussed above, the general solution of (4.28) is given by

$$\phi_k(x) = u_k(x)\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{4.29}$$

Note that between the potential barriers (i.e., $0 < x < a$), $V(x) = 0$, and (4.28) becomes

$$\frac{\partial^2\phi_k}{\partial x^2} + k_0^2\phi_k = 0, \tag{4.30}$$

where

$$k_0^2 = \frac{2mE}{\hbar^2}, \tag{4.31}$$

and k_0 is the wave vector of free electrons. Since the solution of the electron wave functions given by (4.29) is valid everywhere in the periodic lattice, one can substitute (4.29) into (4.30) to obtain an equation that contains only the Bloch

function $u_k(x)$, namely,

$$\frac{\mathrm{d}^2 u_k}{\mathrm{d}x^2} + 2ik\frac{\mathrm{d}u_k}{\mathrm{d}x} + \left(k_0^2 - k^2\right)u_k = 0. \tag{4.32}$$

This is a second-order differential equation with constant coefficients, and the roots of its characteristic equation are equal to $-i(k \pm k_0)$. Thus, the general solution of (4.32) for $u_k(x)$ can be expressed as

$$u_k(x) = \mathrm{e}^{-ik\cdot x}(A\cos k_0 x + B\sin k_0 x), \tag{4.33}$$

where A and B are constants, which can be determined from the periodic boundary conditions. The first boundary condition can be obtained by noting the fact that both $u_k(x)$ and $\phi_k(x)$ are invariant under translation. Thus, one can write

$$u_k(0) = u_k(a), \tag{4.34}$$

where a is the period of the crystal potential (i.e., $V(x) = V(x + a)$). To calculate the change in the slope of the electron wave functions across the infinitely thin potential barrier at the atomic site, one can integrate (4.28) from $x = 0_-$ on the left-hand side of the potential barrier to $x = 0_+$ on the right-hand side of the potential barrier at $x = 0$. This yields

$$\int_{0_-}^{0_+} \left\{ \frac{\partial^2 \phi_k}{\partial x^2} + \left(\frac{2m}{\hbar^2}\right)[E - V(x)]\phi_k \right\} \mathrm{d}x = 0, \tag{4.35}$$

or

$$\phi_k'(0_+) - \phi_k'(0_-) = \left(\frac{2m}{\hbar^2}\right) C\phi_k(0), \tag{4.36}$$

where C is defined by (4.43).

Equation (4.36) is obtained using the fact that as $x \to 0$ inside the potential barrier, integration of $E\mathrm{d}x$ over the barrier width is equal to 0, and the change in the slope of the electron wave functions ($\phi_k' = \mathrm{d}\phi_k/\mathrm{d}x$) across the potential barrier is given by (4.36). From (4.29) and (4.36), one obtains the derivative of u_k as

$$u_k'(0_+) - u_k'(0_-) = \left(\frac{2mC}{\hbar^2}\right) u_k(0). \tag{4.37}$$

Now, replacing $0_+ = 0$ and $0_- = a$ in (4.37), the second boundary condition for $u_k(x)$ is given by

$$u_k'(0) = u_k'(a) + \left(\frac{2mC}{\hbar^2}\right) u_k(0). \tag{4.38}$$

Note that the first derivative of $u_k(x)$ is identical on the left-hand side of each potential barrier shown in Figure 4.1b. Next, substituting the two boundary conditions given by (4.34) and (4.38) into (4.33), one obtains two simultaneous equations for A and B:

$$A(\mathrm{e}^{-ika}\cos k_0 a - 1) + B(\mathrm{e}^{-ika}\sin k_0 a) = 0, \tag{4.39}$$

$$A\left[-ik(1 - \mathrm{e}^{-ika}\cos k_0 a) + \left(\mathrm{e}^{-ika} k_0 \sin k_0 a - \frac{2mC}{\hbar^2}\right)\right] + B\left[k_0 + \mathrm{e}^{-ika}(ik\sin k_0 a - k_0 \cos k_0 a)\right] = 0. \tag{4.40}$$

In order to have a nontrivial solution for (4.39) and (4.40), the determinant of the coefficients of A and B in both equations must be set equal to 0, which yields

$$\begin{vmatrix} [\mathrm{e}^{-ika}\cos k_0 a - 1] & \mathrm{e}^{-ika}\sin k_0 a \\ -ik(1-\mathrm{e}^{-ika}\cos k_0 a) + \left(\mathrm{e}^{-ika}k_0 \sin k_0 a - \dfrac{2mC}{\hbar^2}\right) & \times\left[k_0 + \mathrm{e}^{-ika}(ik\sin k_0 a - k_0 \cos k_0 a)\right] \end{vmatrix} = 0 \tag{4.41}$$

Solving (4.41), one obtains

$$\cos ka = \left(\frac{P}{k_0 a}\right)\sin k_0 a + \cos k_0 a, \tag{4.42}$$

where $P = mCa/\hbar^2$, and C is defined by

$$C = \lim_{\substack{V(x)\to\infty \\ \mathrm{d}x\to 0}} \left[\int_{0_-}^{0_+} V(x)\mathrm{d}x\right]. \tag{4.43}$$

Equation (4.42) has a real solution for the electron wave vector $\boldsymbol{k}$ if the value of the right-hand side of (4.42) lies between -1 and $+1$. Figure 4.2 shows a plot of the right-hand-side term of (4.42) versus $k_0 a$ for a fixed value of P. It is noted that the solution of (4.42) consists of a series of alternate allowed and forbidden regions,

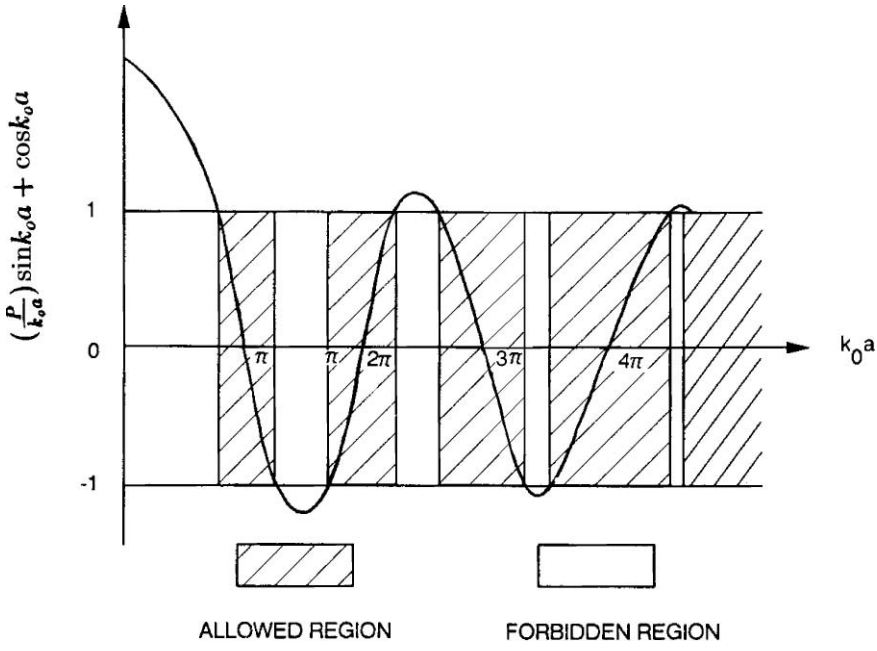


FIGURE 4.2. A plot of the magnitude of the right-hand side of (4.26) versus $k_0 a$ for a one-dimensional periodic lattice.

with the forbidden regions becoming narrower as the value of k_0a becomes larger. We now discuss the physical meaning of Figure 4.2.

It is noted that the magnitude of P is closely related to the binding energy of electrons in the crystal. For example, if P is zero, then one has the free-electron case, and the energy of the electrons is a continuous function of the wave vector k, as given by (4.31). On the other hand, if P approaches infinity, then the energy of the electrons becomes independent of k. This corresponds to the case of an isolated atom. In this case, the values of electron energy are determined by the condition that $\sin k_0a$ in (4.42) must be set equal to 0 as P approaches infinity, which implies $k_0a = n\pi$. Thus, the electron energy levels are quantized for this case, and are given by

$$E_n = \frac{\hbar^2 k_0^2}{2m_0} = \frac{n^2\pi^2\hbar^2}{2m_0a^2}, \tag{4.44}$$

where $n = 1, 2, 3 \ldots$. In this case, the electrons are completely bound to the atom, and their energy levels become discrete. If P has a finite value, then the energy band scheme of electrons is characterized by the alternate allowed and forbidden energy regions, as shown in Figure 4.2. The allowed regions are the regions in which the magnitude of the right-hand side in (4.42) lies between -1 and $+1$, while the forbidden regions are the regions in which the magnitude of the right-hand side is greater than 1. It is further noticed from this figure that the forbidden region becomes smaller and the allowed region becomes larger as the value of k_0a increases.

Figure 4.3 shows the plot of electron energy as a function of P. As shown in this figure, the origin, where $P = 0$, corresponds to the free-electron case, and the energy of electrons is continuous in k-space. In the region where P has a finite value,

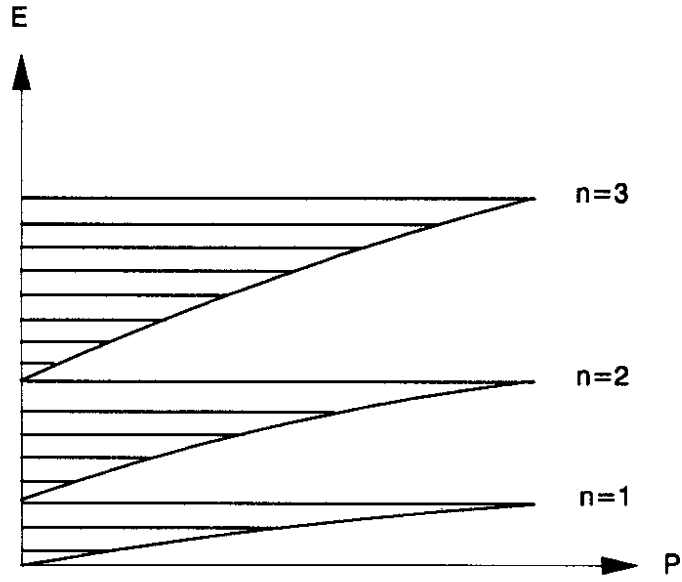


FIGURE 4.3. The energy versus P for a one-dimensional (1-D) periodic lattice.

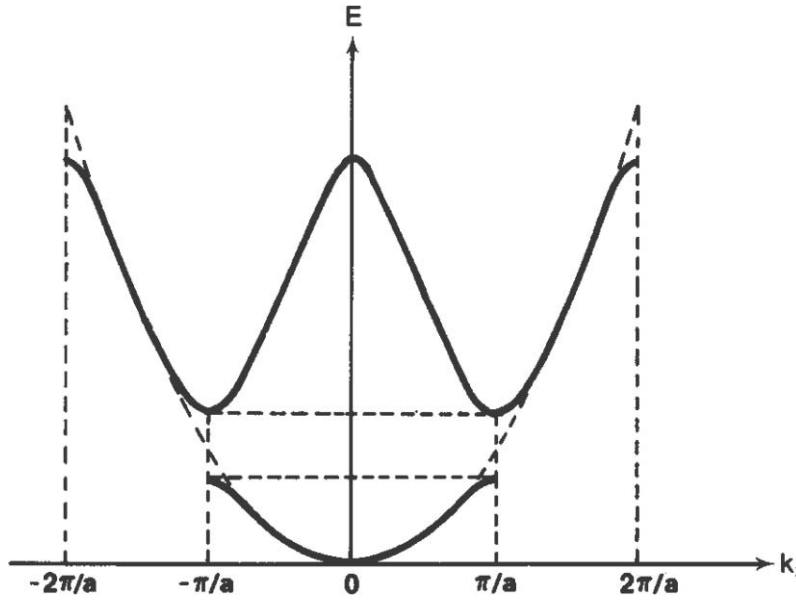


FIGURE 4.4. The energy band diagram for a one-dimensional (1-D) periodic potential.

the energy states of electrons are characterized by a series of allowed (shaded area) and forbidden regions. As P approaches infinity, the energy of electrons becomes discrete (or quantized), which corresponds to the case of an isolated atom with atomic spacing $a \to \infty$.

Based on the Kronig–Penney model discussed above, a schematic energy band diagram for the 1-D periodic lattice is illustrated in Figure 4.4, which is plotted in the extended zone scheme. The values of the wave vector k are given by $-n\pi/a, \ldots, -\pi/a, 0, +\pi/a, \ldots, n\pi/a$. The first Brillouin zone, known as the unit cell of the reciprocal lattice, is defined by the wave vectors with values varying between $-\pi/a$ and $+\pi/a$. Figure 4.4 illustrates two important physical aspects of the energy band diagram: (i) at the zone boundaries where $k = \pm n\pi/a$ and $n = 1, 2, 3, \ldots$, there exists an energy discontinuity, and (ii) the width of allowed energy bands increases with increasing electron energy, and the width of forbidden gaps decreases with increasing electron energy.

If the energy band diagram (i.e., E vs. k) is plotted within the first Brillouin zone, then it is called the reduced zone scheme. The reduced zone scheme (i.e., $-\pi/a \leq k \leq \pi/a$) is more often used than the extended zone scheme because for any values of the wave vector $\boldsymbol{k}'$ in the higher zones there is a corresponding wave vector $\boldsymbol{k}$ in the first Brillouin zone, and hence it is easier to describe the electronic states and the related physical properties using the reduced zone scheme. The relation between $\boldsymbol{k}'$ and $\boldsymbol{k}$ can be obtained via the translational symmetry

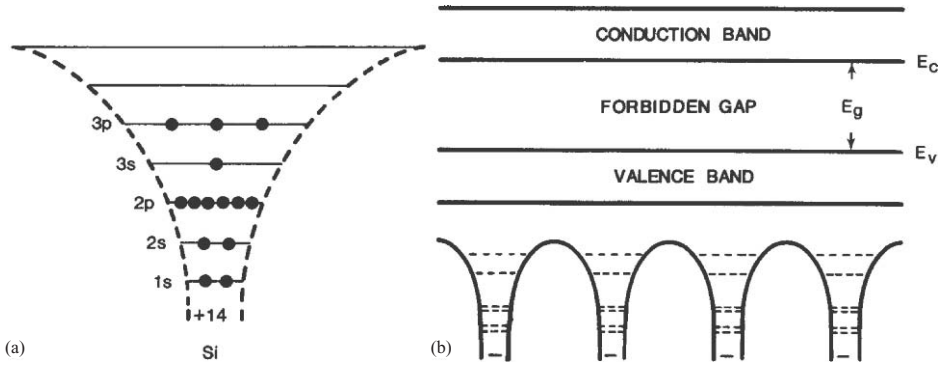


FIGURE 4.5. (a) Energy band diagrams for an isolated silicon atom, and (b) a one-dimensional silicon lattice.

operation, which is given by

$$\boldsymbol{k}' = \boldsymbol{k} \pm 2n\pi/a, \tag{4.45}$$

where $\boldsymbol{k}'$ represents the wave vector in the higher zones, $\boldsymbol{k}$ is the corresponding wave vector in the first Brillouin zone, $n = 1, 2, 3 \ldots$, and a denotes the lattice constant of the crystal. Thus, the reduced zone scheme contains all the information relating to the electronic states in the crystalline solids.

The Kronig–Penney model described above can be employed to construct the energy band diagrams of an isolated silicon atom and an artificial 1-D periodic silicon lattice. Figures 4.5a and b show the discrete energy level schemes for such an isolated silicon atom and the energy band diagram for a 1-D silicon lattice, respectively. As shown in Figure 4.5a, electrons in the 3s and 3p shells are known as the valence electrons, while electrons in the 1s, 2s, and 2p orbits are called the core electrons. When the valence electrons are excited into the conduction band, the conductivity of a semiconductor increases. It is noted that as the spacing of silicon atoms reduces to a few angstroms, the discrete energy levels shown in Figure 4.5a broaden into energy bands, and each allowed energy band is separated by a forbidden band gap. In this energy band scheme the highest filled band (i.e., 3s and 3p states for silicon) is called the valence band, while the lowest empty band is called the conduction band. In a semiconductor, a forbidden band gap always exists between the conduction and the valence bands, while in metals the energy bands are usually continuous. For most semiconductors, the band gap energies may vary between 0.1 and 6.2 eV.

The main difference in the energy band scheme between the 1-D and 2- or 3-D crystal lattices is that in the 1-D case, an energy discontinuity always exists at the zone boundary, and hence the energy band is characterized by a series of alternate allowed and forbidden bands. However, in the 3-D case, the energy band discontinuity may or may not exist, since the values of $k_{\max}$ at the zone boundaries along different crystal orientations may be different, as is clearly illustrated in

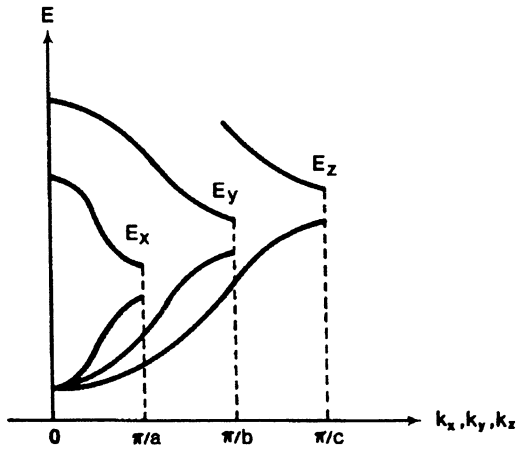


FIGURE 4.6. Energy band diagrams in the reduced zone scheme for a three-dimensional (3-D) rectangular lattice, assuming $a \neq b \neq c$.

Figure 4.6. This will lead to an overlap of energy states at the zone boundaries and hence the possible disappearance of the band gap in the 3-D energy band diagram. It should be mentioned that the electron wave functions in a 3-D periodic crystal lattice are of the Bloch type and can be described by (4.17). In the next section we shall describe the nearly free electron (NFE) approximation for constructing the energy band scheme of valence electrons in a semiconductor. It is noted that the NFE approximation can provide only a qualitative description of the energy band schemes for the valence electrons in a 3-D crystal lattice. To obtain true energy band structures for semiconductors and metals, more rigorous and sophisticated methods such as the pseudopotential and orthogonalized plane wave methods must be used in calculations of the energy band structures for these materials.

4.5. The Nearly Free Electron Approximation

In Section 4.4, it was shown that when the value of P in (4.42) is small compared to $k_0 a$, the behavior of electrons in the 1-D periodic lattice should resemble that of the free-electron case, in which the energy band is continuous in k-space. In a semiconductor, the outer-shell valence electrons are loosely bound to the atoms, and the effect of the periodic crystal potential on the electron wave functions can be treated as a perturbing potential. In this case, the nearly free electron (NFE) approximation can be applied to deal with the valence electrons.

In order to apply the NFE approximation to a 3-D crystal lattice, the periodic potential must be treated as a small perturbation. In doing so, one assumes that the perturbing potential is small compared to the average energy of electrons. The problem can then be solved using the quantum-mechanical stationary perturbation theory. From wave mechanics, the stationary perturbation method can be derived using the first- and second-order approximations in the time-independent Schrödinger

equation. In the NFE approximation, it is assumed that the total Hamiltonian H consists of two parts, H_0 and H', with H_0 being the unperturbed Hamiltonian and H' the perturbed Hamiltonian. Thus, one can write

$$H = H_0 + aH', \quad \text{for } a \leq 1. \tag{4.46}$$

The unperturbed one-electron Schrödinger equation is given by

$$H_0\phi_{n0} = E_{n0}\phi_{n0}, \tag{4.47}$$

where ϕ_{n0} and E_{n0} are the unperturbed eigenfunctions and eigenvalues, respectively. The perturbed Schrödinger equation is given by

$$H\phi_n = E_n\phi_n. \tag{4.48}$$

From stationary perturbation theory, the solutions of the electron wave functions and energies in (4.48) can be expressed in terms of power series expansions, which are given, respectively, by

$$\phi_n = \phi_{n0} + a\phi_{n1} + a^2\phi_{n2} + \cdots, \quad \text{where } a \leq 1, \tag{4.49}$$

$$E_n = E_{n0} + aE_{n1} + a^2E_{n2} + \cdots. \tag{4.50}$$

The new perturbed wave functions ϕ_{nj} ($j = 1, 2, 3, \ldots$) given in (4.49) and (4.50) can be expressed in terms of a linear combination of the unperturbed wave functions ϕ_{l0} as

$$\phi_{nj} = \sum_{l=0}^{\infty} b_{lj}\phi_{lj}. \tag{4.51}$$

Now, substituting (4.46), (4.49), and (4.50) into (4.48) and equating the coefficients of the a and a^2 terms on both sides of (4.49) and (4.50), one obtains

$$H_0\phi_{n1} + H'\phi_{n0} = E_{n0}\phi_{n1} + E_{n1}\phi_{n0}, \tag{4.52}$$

$$H_0\phi_{n2} + H'\phi_{n1} = E_{n0}\phi_{n2} + E_{n1}\phi_{n1} + E_{n2}\phi_{n0}. \tag{4.53}$$

Note that (4.52) contains the coefficients of the a term, and (4.53) contains the coefficients of the a^2 term. For simplicity one can set a equal to 1. Consequently, the first-order correction of energy, E_{n1}, and wave function, ϕ_{n1}, is obtained by multiplying both sides of (4.52) by the unperturbed conjugate wave function ϕ^*_{m0} and integrating the equation over the entire volume. This yields

$$\int \phi^*_{m0}\left[H_0\left(\sum_{l=0}^{\infty} b_{l1}\phi_{l0}\right) + H'\phi_{n0}\right] \mathrm{d}r^3 = \int \phi^*_{m0}\left[E_{n0}\left(\sum_{l=0}^{\infty} b_{l1}\phi_{l0}\right) + E_{n1}\phi_{n0}\right] \mathrm{d}^3r. \tag{4.54}$$

Integrating (4.54) using the orthonormality of the wave functions ϕ_{n0} and the

Hermitian property of H_0, one obtains

$$b_{m1}E_{m0} + \int \phi_{m0}^{*} H' \phi_{n0} \mathrm{d}^3 r = E_{n0} b_{m1} \quad \text{for } m \neq n, \tag{4.55}$$

and

$$E_{n1} = \int \phi_{n0}^{*} H' \phi_{n0} \mathrm{d}r^3 = H'_{nn} \quad \text{for } m = n. \tag{4.56}$$

Solving (4.55) and (4.56) yields

$$b_{m1} = \frac{H'_{mn}}{(E_{n0} - E_{m0})} \quad \text{for } m \neq n, \tag{4.57}$$

$$E_{\mathrm{n1}} = 0 \qquad \text{for } m = n. \tag{4.58}$$

In (4.57), H'_{mn} is called the matrix element, and is defined by the second term on the left-hand side of (4.55). Thus, the new electron wave function ϕ_n with the first-order correction using the stationary perturbation theory is given by

$$\phi_n = \phi_{n0} + \phi_{n1} = \phi_{n0} + \sum_{\substack{m=0 \\ m \neq n}}^{\infty} \frac{H'_{mn}\phi_{m0}}{(E_{n0} - E_{m0})}, \tag{4.59}$$

where the matrix element H'_{mn} can be expressed by

$$H'_{mn} = \int \phi_{m0}^{*} H' \phi_{n0} \mathrm{d}r^3, \tag{4.60}$$

where H' is the perturbing Hamiltonian. Equation (4.59) can be used to find the wave functions of valence electrons in a periodic crystal lattice using the NFE approximation. In order to find the lowest-order correction of the electron energy due to the perturbing potential H', it is usually necessary to carry out the expansion to the second-order correction given by (4.50). The reason for the second-order correction in energy calculations is that the perturbed Hamiltonian H' has a vanishing diagonal matrix element such that the first-order correction in energy is equal to 0 (i.e., $E_{n1} = 0$). This can be explained by the fact that the perturbed Hamiltonian H' is usually an odd function of the coordinates, and hence H'_{nn} is equal to 0. From (4.51), the perturbed wave functions for the first- and second-order corrections are given, respectively, by

$$\phi_{n1} = \sum_{l=0}^{\infty} b_{l1}\phi_{l0}, \tag{4.61}$$

$$\phi_{n2} = \sum_{l=0}^{\infty} b_{l2}\phi_{l0}. \tag{4.62}$$

Now, substituting (4.61) and (4.62) into (4.53) and using the same procedure as described above for the first-order correction of electron wave functions, one

obtains the second-order correction of energy, which is

$$E_{n2} = \sum_{\substack{m=0 \\ m \neq n}}^{\infty} \frac{|H'_{nm}|^2}{(E_{n0} - E_{m0})}. \tag{4.63}$$

Using (4.63), the expression for the electron energy corrected to the second order is given by

$$E_n = E_{n0} + \sum_{\substack{m=0 \\ m \neq n}}^{\infty} \frac{|H'_{nm}|^2}{(E_{n0} - E_{m0})}. \tag{4.64}$$

Equations (4.59) and (4.64) are the new wave functions and energies of electrons derived from the quantum-mechanical stationary perturbation theory. The results may be used in the NFE approximation to find the wave functions and energies of the outer-shell electrons of a crystalline solid. As mentioned earlier, the valence electrons in a semiconductor are loosely bound to the atoms, and hence the periodic crystal potential seen by these valence electrons can be treated as a small perturbing Hamiltonian. The unperturbed one-electron Schrödinger equation is described by

$$\frac{-\hbar^2}{2m_0} \nabla^2 \phi_k^0(r) = E_k^0 \phi_k^0(r), \tag{4.65}$$

which has the solutions of free-electron wave functions and energies given, respectively, by

$$\phi_k^0(r) = \sqrt{\frac{1}{NV}} \mathrm{e}^{ik \cdot r}, \tag{4.66}$$

$$E_k^0 = \frac{\hbar^2 k^2}{2m_0}, \tag{4.67}$$

where N is the total number of unit cells in the crystal, V is the volume of the unit cell, $\phi_k^0(r)$ are the free-electron wave functions, and E_k^0 is the free-electron energy. The preexponential factor given by (4.66) is the normalization constant. The one-electron Schrödinger equation in the presence of a periodic crystal potential $V(r)$ is given by

$$\left(-\frac{\hbar^2}{2m^*} \right) \nabla^2 \phi_k(r) + V(r) \phi_k(r) = E_k \phi_k(r), \tag{4.68}$$

where m^* is the effective mass of electrons in the crystal. The crystal potential $V(r)$ can be expressed in terms of the Fourier expansion in the reciprocal space, which is given by

$$V(r) = \sum_{K_j} v(K_j) \mathrm{e}^{-iK_j \cdot r}, \tag{4.69}$$

where K_j is the reciprocal lattice vector and $v(K_j)$ is the Fourier coefficient of the periodic potential $V(r)$.

The new electron wave functions and energies can be obtained by finding the matrix element $H_{k'k}$ due to the periodic crystal potential $V(r)$ using the stationary perturbation method described above. Now substituting (4.69) into (4.60), the matrix element due to the periodic potential $V(r)$ is given by

$$H_{kk'} = \int \phi_{k'}^{*}(r)|V(r)|\phi_k(r)\mathrm{d}r^3$$
$$= \left(\frac{1}{NV}\right)\int \mathrm{e}^{-ik'\cdot r}\left(\sum_{K_j} v(K_j)\mathrm{e}^{-iK_j\cdot r}\right)\mathrm{e}^{ik\cdot r}\mathrm{d}^3r. \tag{4.70}$$

Note that the integral on the right-hand side of (4.70) will vanish unless $k - k' = K_j$, where K_j is the reciprocal lattice vector. Thus, by substituting $k - k' = K_j$ in (4.70) and carrying out the integration one obtains

$$H_{k'k} = v(K_j). \tag{4.71}$$

Now, substituting (4.71) into (4.59) yields the new electron wave function, which is

$$\phi_k(r) = \sqrt{\frac{1}{NV}}\mathrm{e}^{ik\cdot r}\left[1 + \sum_{K_j}\frac{v(K_j)\mathrm{e}^{-iK_j\cdot r}}{\left(E_k^0 - E_{k'}^0\right)}\right]. \tag{4.72}$$

It is interesting to note that the term inside the square brackets on the right-hand side of (4.72) has the periodicity of the crystal potential $V(r)$, and may be designated as the Bloch function $u_k(r)$. Thus, the new electron wave functions given by (4.72) indeed satisfy the Bloch-type wave functions defined by (4.17).

The expression of electron energy can be obtained by substituting (4.71) into (4.64), yielding

$$E_k = E_k^0 + \sum_{K_j}\frac{|v(K_j)|^2}{\left(E_k^0 - E_{k'}^0\right)}. \tag{4.73}$$

It is seen that the expressions for the electron wave functions and energies given by (4.72) and (4.73) become infinite if $E_k^0 = E_{k'}^0$, and hence the perturbation approximation is no longer valid. This condition occurs at the zone boundaries, and the electron energy corresponding to this condition is given by

$$E_k^0 = \frac{\hbar^2k^2}{2m_0} = \frac{\hbar^2(k - K_j)^2}{2m_0} = E_{k'}^0. \tag{4.74}$$

Solving (4.74) yields

$$k\cdot K_j = \frac{|K_j|^2}{2}. \tag{4.75}$$

Here the relation $k' = k - K_j$ is used in (4.74). Equation (4.75) represents exactly the Bragg diffraction condition in a crystalline solid, which occurs at the zone boundaries. Failure of the perturbation theory at the zone boundaries is due to the fact that the periodic crystal potential $V(r)$ at zone boundaries is no longer small, and hence cannot be treated as a small perturbing potential. In fact, the Bragg diffraction condition results in a very severe perturbation of electron wave functions

at the zone boundaries. Therefore, to find a proper solution for the electron energy and wave functions at the zone boundaries, it is necessary to reconstruct a new perturbed wave function, which is a linear combination of an incident- and a reflected-plane wave. Using a linear combination of the incident- and reflected-plane waves, one can construct a new electron wave function at the zone boundary, which is given by

$$\phi_k^0(r) = A_0 e^{ik\cdot r} + A_1 e^{ik'\cdot r}, \tag{4.76}$$

where $k' = k - K_j$. Substituting (4.76) into (4.65) yields

$$\left\{\frac{\hbar^2 k^2}{2m} + [V(r) - E_k]\right\} A_0 e^{ik\cdot r} + \left\{\frac{\hbar^2 k'^2}{2m} + [V(r) - E_k]\right\} A_1 e^{ik'\cdot r} = 0. \tag{4.77}$$

Now, multiplying (4.77) by $e^{-ik\cdot r}$ and integrating the equation over the entire space, one obtains

$$A_0\left(E_k^0 - E_k\right) - A_1 v^*(K_j) = 0, \tag{4.78}$$

where $E_k^0 = \frac{\hbar^2 k^2}{2m_0}$, and $v^*(K_j)$, the conjugate of the Fourier coefficient, is given by

$$v^*(K_j) = \int_0^\infty e^{-ik\cdot r} V(r) e^{ik'\cdot r} d^3 r. \tag{4.79}$$

Similarly, multiplying (4.77) by $e^{-ik'\cdot r}$ and integrating over the entire space, one obtains

$$A_0 v(K_j) - A_1\left(E_k - E_{k'}^0\right) = 0, \tag{4.80}$$

where $E_{k'}^0 = \frac{\hbar^2 k'^2}{2m_0}$, and

$$v(K_j) = \int_0^\infty e^{-ik'\cdot r} V(r) e^{ik\cdot r} d^3 r \tag{4.81}$$

is the Fourier coefficient of the periodic crystal potential $V(r)$. A nontrivial solution exists in (4.78) and (4.80) only if the determinant of the coefficients of A_0 and A_1 is set equal to zero, namely,

$$\begin{vmatrix} \left(E_k^0 - E_k\right) & -v^*(K_j) \\ v(K_j) & -\left(E_k - E_{k'}^0\right) \end{vmatrix} = 0. \tag{4.82}$$

Now, solving (4.82) for E_k yields

$$E_k = \frac{1}{2}\left\{\left(E_k^0 + E_{k'}^0\right) \pm \left[\left(E_k^0 - E_{k'}^0\right)^2 + 4v^*\left(K_j\right)\cdot v(K_j)\right]^{1/2}\right\}. \tag{4.83}$$

Equation (4.83) shows that a forbidden gap exists at the zone boundaries, and the width of the forbidden gap is determined by the value of $4v^*(K_j)\cdot v(K_j)$ inside the square brackets of (4.83), which is determined by the Fourier coefficient

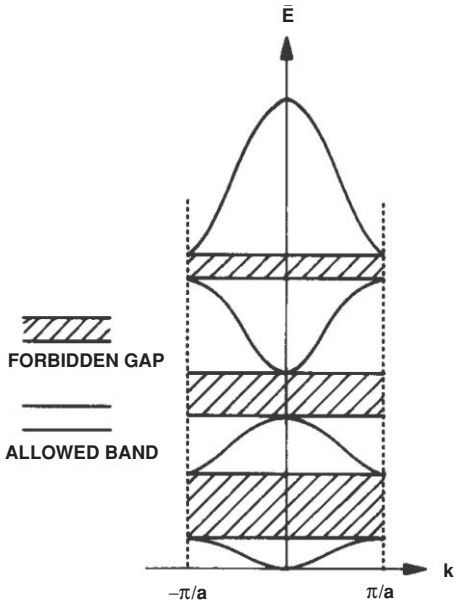


FIGURE 4.7. The energy band diagram in a reduced zone scheme showing the discontinuity of the energy at the zone boundaries.

of the periodic crystal potential. In general, the energy band gap will increase with increasing value of the Fourier coefficient $|v(K_j)|$. Figure 4.7 shows the schematic energy band diagram in the reduced zone scheme derived from the NFE approximation. It is interesting to note that the energy band scheme derived from NFE approximation is similar to that obtained from the Kronig–Penney model for the 1-D periodic lattice. Furthermore, the electron wave functions derived from the NFE approximation indeed satisfy the Bloch condition. The results show that, except at the zone boundaries where an energy discontinuity (or a band gap) occurs, the energy band scheme derived from the NFE approximation resembles that of the free-electron case (with $v(K_j) = 0$) discussed earlier.

The NFE approximation presented in this section provides a qualitative description of the electronic states for the outer-shell valence electrons of a 3-D crystal lattice. However, in order to obtain true energy band structures for a real crystal, a more rigorous and sophisticated method, such as the pseudopotential or the orthogonalized plane wave method, must be employed in the energy band calculations. Both methods have been widely used in the energy band calculations of semiconductors.

4.6. The Tight-Binding Approximation

In this section energy band calculation using the tight-binding approximation or the linear combination of atomic orbits (LCAO) method is described. The LCAO method, which was first proposed by Bloch, is often used to calculate the electronic states of core electrons in a crystalline solid. It is generally known that core

electrons are tightly bound to the individual atoms, which interact with one another within the crystal lattice. In this case, the construction of electron wave functions is achieved using the LCAO method, and the energy bands of electrons are calculated for the corresponding periodic crystal potential. The atomic orbitals are centered on one of the constituent atoms of the crystal. The resulting wave functions are then substituted into the Schrödinger equation, and the energy values are calculated by a procedure similar to that of the NFE approximation described in Section 4.5. In order to apply the LCAO method to core electrons in a crystalline solid, the solution for the free atomic orbital wave functions must be obtained first. This is discussed next.

If $\phi_n(r - R_j)$ represents the atomic orbital wave functions centered at the lattice site R_j, then the wave functions of the crystal orbits $\phi_k(r)$ corresponding to the wave vector $\boldsymbol{k}$ may be represented by a Bloch sum, which is

$$\phi_k(r) = \sum_j C_j(k)\phi_n(r - R_j). \tag{4.84}$$

The summation in (4.84) extends over all the constituent atoms of the crystal. The coefficient $C_j(k)$, which satisfies the Bloch condition, can be written as

$$C_j(k) = \mathrm{e}^{ik\cdot R_j}. \tag{4.85}$$

Now substituting (4.85) into (4.84) one obtains

$$\phi_k(r) = \sum_j \mathrm{e}^{ik\cdot r}\mathrm{e}^{-ik\cdot(r-R_j)}\phi_n(r - R_j) = \mathrm{e}^{ik\cdot r}U_{k,n}(r). \tag{4.86}$$

To satisfy the Bloch condition, the summation given by (4.86) must have the periodicity of the crystal lattice.

The LCAO method is clearly an approximation to the true crystal orbitals. This method is adequate when the interatomic spacing is large enough such that overlapping among the atomic orbital wave functions $\phi_n(r - R_j)$ is negligible. Thus, the LCAO method is most suitable for the tightly bound core electrons, and is frequently referred to as the tight-binding approximation. Using this method to derive the wave functions and energy band schemes for the core electrons of a crystalline solid is discussed next.

If $\phi_n(r - R_j)$ represents a set of atomic orbital wave functions that satisfy the free-atom Schrödinger equation, then one can write

$$-\left(\frac{\hbar^2}{2m^*}\right)\nabla^2\phi_n(r - R_j) + V_{n0}(r - R_j)\phi_n(r - R_j) = E_{n0}\phi_n(r - R_j), \tag{4.87}$$

where $V_{n0}(r - R_j)$ is the free atomic potential of the R_jth atom. The wave functions for the crystal orbitals may be expressed in terms of a Bloch sum, which is given by

$$\phi_k(r) = \left(\frac{1}{NV}\right)^{1/2}\mathrm{e}^{ik\cdot r}\mathrm{e}^{-ik(r-R_j)}\phi_n(r - R_j) = \left(\frac{1}{NV}\right)^{1/2}\mathrm{e}^{ik\cdot r}u_k(r), \tag{4.88}$$

where $u_k(r)$ is the Bloch function. In (4.88), the atomic wave functions are normalized (i.e., N represents the total number of atoms in the crystal). The factor $(1/NV)^{1/2}$ is the normalization constant for the Bloch sum if overlapping of the atomic orbitals centered at different atomic sites is negligible. Thus, (4.88) is a good approximation for the crystal orbitals, provided that the energy levels of the atomic orbits are nondegenerate and overlapping between the orbital wave functions of the neighboring atoms is negligible. This condition can be expressed by

$$\int \phi_n^*(r - R_j)\phi_n(r - R_i)\mathrm{d}r^3 = \delta_{ij}. \tag{4.89}$$

Note that in (4.89), $\delta_{ij} = 0$ if $i \neq j$. Now, substituting (4.88) into (4.87), multiplying (4.87) by the conjugate wave functions $\phi_n^*(r - R_i)$, and integrating the equation over the entire space, one obtains the energy

$$\begin{aligned} E_k &= \int \phi_k^*(r)H\phi_k(r)\mathrm{d}r^3 \\ &= \left(\frac{1}{NV}\right)\left\{\int \sum_{ij} \mathrm{e}^{ik(R_j - R_i)}\phi_n^*(r - R_i)\left[-\frac{\hbar^2\nabla^2}{2m^*} + V_{n0}(r - R_j)\right]\phi_n(r - R_j)\mathrm{d}r^3\right. \\ &\qquad \left. + \int \sum_{ij} \mathrm{e}^{jk\cdot(R_j - R_i)}\phi_n^*(r - R_j)V'(r - R_j)\phi_n(r - R_j)\mathrm{d}r^3\right\}. \end{aligned} \tag{4.90}$$

Using (4.89), (4.90) can be rewritten as follows:

$$E_k = E_{n0} - \alpha_n - \sum_{R_{ij}} \beta_n(R_{ij})\mathrm{e}^{ik\cdot R_{ij}}, \tag{4.91}$$

where $R_{ij} = R_j - R_i$, and

$$E_{n0} = \left(\frac{1}{NV}\right)\int \phi_n^*\left[-\frac{\hbar^2\nabla^2}{2m^*} + V_{n0}\right]\phi_n \mathrm{d}^3r, \tag{4.92}$$

$$\alpha_n = -\int \phi_n^2(r - R_i)V'(r - R_j)\mathrm{d}r^3, \tag{4.93}$$

$$\beta_n = -\int \phi_n^*(r - R_i)V'(r - R_j)\phi_n(r - R_j)\,\mathrm{d}r^3, \tag{4.94}$$

$$V(r - R_j) = V_{n0}(r - R_j) + V'(r - R_j). \tag{4.95}$$

As shown in Figure 4.8, $V_{n0}(r - R_j)$ is the unperturbed atomic potential centered at R_j, and $V'(r - R_j)$ is the perturbed crystal potential due to atoms other than the R_jth atom.

In general, the atomic orbital wave functions $\phi_n(r)$ fall off exponentially with the distance r, and hence overlapping of each atomic orbital wave function $\phi_n(r)$ is assumed to be negligibly small. Therefore, it is expected that the contribution to β_n will come from a rather restricted range of r. Furthermore, it is also expected that β_n will decrease rapidly with increasing distance between the neighboring atoms. Figure 4.8 illustrates the potential $V'(r - R_j)$, which plays the role of the

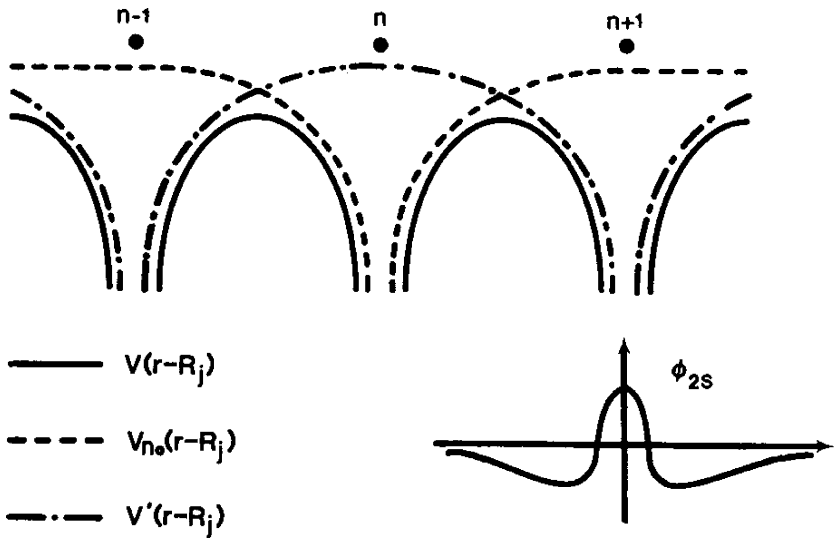


FIGURE 4.8. The crystal potential used in the tight-binding approximation.

perturbing potential and is practically zero in the vicinity of R_j. The LCAO method may be applied to construct the energy band structures of the s-like states for a simple cubic lattice and a body-centered cubic lattice. This is discussed next.

4.6.1. The s-like States for a Simple Cubic Lattice

The LCAO method is first applied to the calculations of the energy band structure of the s-like states for a simple cubic lattice. In a simple cubic lattice, there are six nearest-neighbor atoms located at an equal distance a from any chosen atomic site. Therefore, the value of $\beta_n(a)$, given by (4.94), is the same for all six nearest-neighbor atoms. Since the perturbing potential $V'(r)$ is negative, and the atomic wave functions are of the same sign in the region of overlapping, values for both α_n and $\beta_n(a)$ are positive. Thus, the energy dispersion relation (E vs. k) for s-like states of a simple cubic lattice can be derived by substituting $R_{ij} = (a, 0, 0), (0, a, 0), (0, 0, a), (-a, 0, 0), (0, -a, 0), (0, 0, -a)$ into (4.91), which yields

$$\begin{aligned} E_k &= E_0 - \alpha_n - \beta_n \left(e^{ik_x a} + e^{ik_y a} + e^{ik_z a} + e^{-ik_x a} + e^{-ik_y a} + e^{-ik_z a}\right) \\ &= E_{n0} - \alpha_n - 2\beta_n(\cos\, k_x a + \cos\, k_y a + \cos\, k_z a). \end{aligned} \tag{4.96}$$

Equation (4.96) shows the E–k relation for the s-like states of a simple cubic lattice. Figures 4.9a and b shows the energy band diagrams plotted in the k_x-direction and the k_x–k_y plane, respectively, as calculated from (4.96). The width of the energy band for this case is equal to $12\beta_n$. It is of interest to note that the shape of the E–k plot is independent of the value of α_n or β_n used, but depends only on the geometry of the crystal lattice. Two limiting cases deserve special mention, namely, (i) near

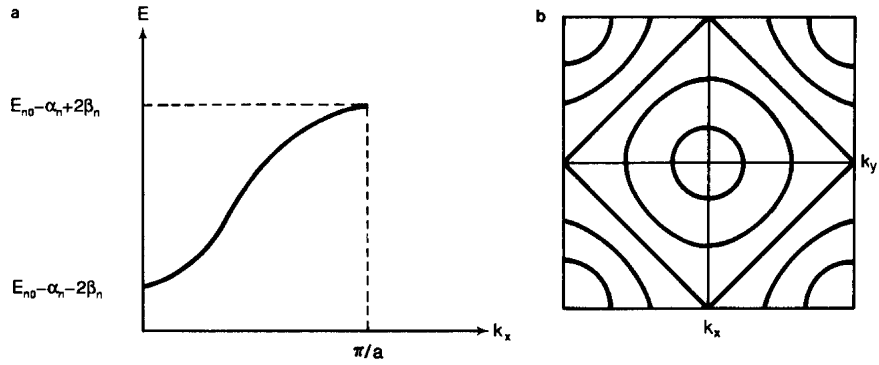


FIGURE 4.9. Energy band diagram for the *s*-like states of a simple cubic lattice: (a) one-dimensional and (b) two-dimensional energy band diagrams.

the top of the band and (ii) near the bottom of the band. First, near the bottom of the band, the value of k is very small, and the cosine terms in (4.96) may be expanded for small ka (i.e., $\cos ka \approx (1 - k^2a^2/2)$). If only the first-order term is retained, then the energy E is found to vary with k^2 near the bottom of the band. This result is identical to the free-electron case. Under this condition, the E–k relation for the s-like states in a simple cubic lattice is reduced to

$$E_k = E_{n0} - \alpha_n - 6\beta_n + \beta_n k^2 a^2. \tag{4.97}$$

From (4.97), the electron effective mass m^* for small ka can be expressed as

$$m^* = \hbar^2 \left(\frac{\partial^2 E_k}{\partial k^2} \right)^{-1} = \frac{\hbar^2}{2\beta_n a^2}, \tag{4.98}$$

which shows that the constant-energy surface near the bottom of the band is parabolic (i.e., $E_k = \hbar^2 k^2 / 2m^*$), and the effective mass of electrons is a scalar quantity. Similarly, the E–k relation near the top of the band (i.e., $k \approx \pi/a$) can be obtained by expanding $\cos(ka)$ in (4.96) at $k_x = k_y = k_z = \pi/a$. This is carried out by substituting $k_x = \pi/a - k'_x$, $k_y = \pi/a - k'_y$, and $k_z = \pi/a - k'_z$ into (4.96), where k'_x, k'_y, k'_z, are small wave vectors, which yields

$$E_{k'} = C + \frac{\hbar^2 k'^2}{2m^*}, \tag{4.99}$$

where C is a constant, and m^* is given by

$$m^* = -\frac{\hbar^2}{2\beta_n a^2}. \tag{4.100}$$

Equation (4.100) shows that the electron effective mass m^* is negative near the top of the band. It is noted that the effective masses given by (4.98) and (4.100) represent the curvatures of the bottom and top of the *s*-like energy band, respectively. The effective mass is an important physical parameter in that it

measures the curvature of the $(E-k)$ energy band diagram. It is noted that a positive m^* means that the band is bending upward, and a negative m^* implies that the band is bending downward. Moreover, an energy band with a large curvature corresponds to a small effective mass, and an energy band with a small curvature represents a large effective mass. The effective mass concept is important since the mobility of electrons in a band is inversely proportional to the effective mass of electrons. For example, by examining the curvature of the energy band diagram near the bottom of the conduction band one can obtain qualitative information concerning the effective mass and the mobility of electrons in the conduction band. A detailed discussion of the effective masses for electrons (or holes) in the bottom (or top) of an energy band is given in Section 4.8.

4.6.2. The s-like States for a Body-Centered Cubic Lattice

For a body-centered cubic (BCC) lattice, there are eight nearest-neighbor atoms for each chosen atomic site, which are located at $R_{ij} = (\pm a/2, \pm a/2, \pm a/2)$. If one substitutes these values in (4.101), the E–k relation for the s-like states of the BCC crystal can be expressed as

$$E_k = E_{n0} - \alpha_n - 8\beta_n[\cos(k_x a/2)\cos(k_y a/2)\cos(k_z a/2)]. \qquad (4.101)$$

In (4.101), values of k must be confined to the first Brillouin zone in order to have nondegenerate energy states. Using (4.101), the 2-D constant-energy contour plotted in the first quadrant of the k_x–k_y plane for the s-like states of a body-centered cubic lattice is shown in Figure 4.10. Although the constant-energy surfaces are spherical near the zone center and zone boundaries, the constant-energy contours depart considerably from the spherical shape for other values of k. For small values of k near the zone center and for large values of k near the zone boundaries, the electron energy E is proportional to k^2, and the effective mass of electrons can be

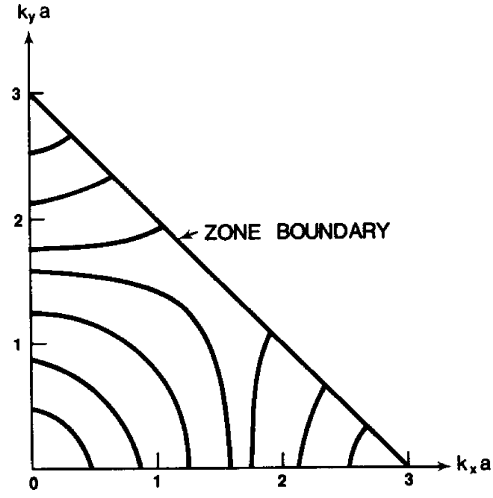


FIGURE 4.10. Constant-energy contours for a two-dimensional body-centered cubic (BCC) lattice. The $E-k$ relation is given by $E_k = E_{n0} - \alpha_n - 8\beta_n \cos(k_x a/2) \times \cos(k_y a/2)$ and $k_z = 0$.

derived from (4.101), which yields

$$m^* = \frac{\hbar^2}{8a^2\beta_n}. \tag{4.102}$$

From (4.101), it can be shown that the total width of the allowed energy band for the s-like states in a body-centered cubic crystal lattice is equal to $16\beta_n(a)$.

It is clear from the above examples that the tight-binding approximation is indeed applicable for calculating the energy states of the core electrons, such as the s-like states in the cubic crystals.

4.7. Energy Band Structures for Some Semiconductors

Calculations of energy band structures for the elemental (Si, Ge) and III-V compound semiconductors (e.g., GaAs, InP) have been widely reported in the literature. As a result, a great deal of information is available for the band structures of semiconductors from both the theoretical and experimental sources. In most cases, theoretical calculations of the energy band structures for these semiconductor materials are guided by the experimental data from the optical absorption, photoluminescence, and photoemission experiments in which the fundamental absorption process is closely related to the density of states and the transitions from the initial to the final states of the energy bands. The energy band structures for some elemental and III-V compound semiconductors calculated from the pseudopotential method are discussed in this section. In general, the exact calculations of the energy band structures for semiconductors are much more complex than those of the NFE approximation and the LCAO method described in this chapter. In fact, both of these approximations can provide only a qualitative description of the energy bands in a crystalline solid. For semiconductors, the two most commonly used methods for calculating the energy band structures are the pseudopotential and the orthogonalized plane wave methods. They are discussed briefly as follows.

The main difficulty of band calculations in a real crystal is that the only wave functions that satisfy the boundary conditions imposed by the Bloch theorem in a simple manner are plane waves, but plane wave expressions do not converge readily in the interior of an atomic cell. The pseudopotential method is based on the concept of introducing the pseudopotential for a crystal that will lead to the same energy levels as the real crystal potential but do not have the same wave functions. The pseudopotential technique can greatly improve the convergence of the series of the plane waves that represent the pseudowave functions of electrons in a crystal. In many cases it is convenient to choose the pseudopotential to be a constant within the ion core. The parameters of the pseudopotential can be determined from the spectroscopic data for the individual atom. Results of the empirical pseudopotential energy band calculations for some elemental and compound semiconductors with diamond and zinc blende structures are shown in Figure 4.11.[1] Figure 4.12 shows the various symmetry points displayed at the zone center (Γ) and along the (100) axis (X) and (111) axis (L) inside the first Brillouin zone of a diamond lattice. The

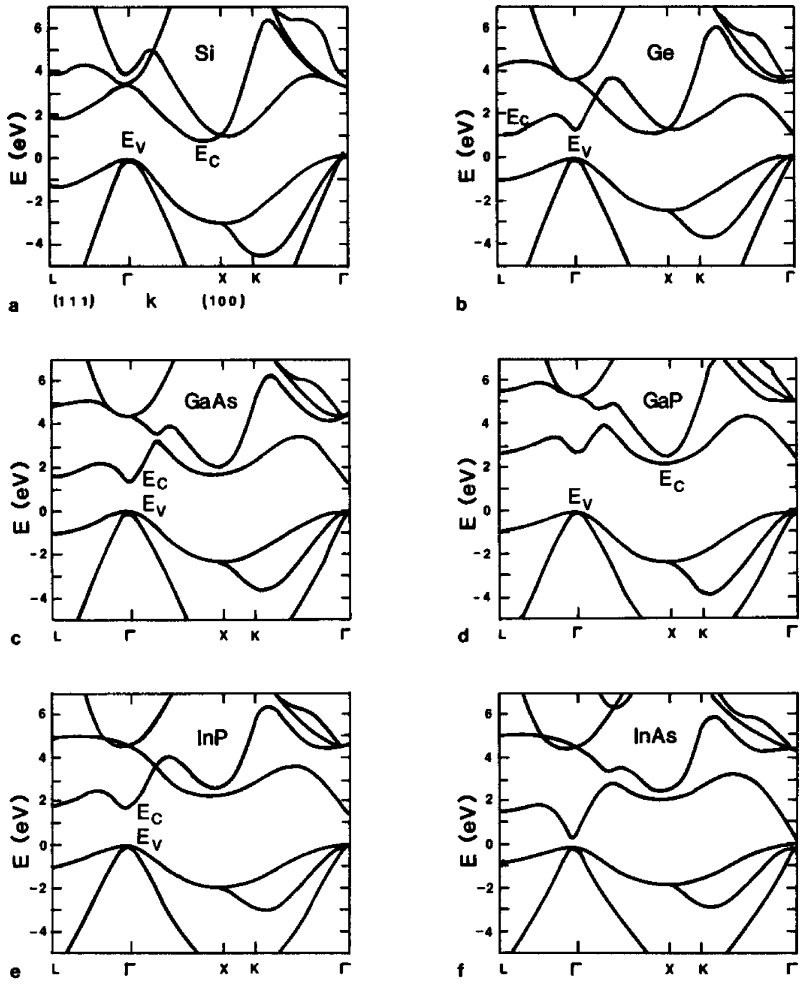


FIGURE 4.11. The energy band structures for some semiconductors with diamond and zinc blende structures. After Cohen and Bergstrasser,[1] by permission.

first symmetry point, Γ, is the symmetry point located at the Brillouin zone center. The conduction band minimum and the valence band maximum located at the Γ-point in the zone center are designated as E_c and E_v, respectively. It is noted that the conduction band is defined as the lowest empty band, while the valence band is defined as the highest filled band at $T = 0$ K. In most semiconductors, there exists a forbidden gap between the conduction and valence bands, and the values of the energy band gap may vary from 0.1 to about 6.2 eV for the semiconductors. If the conduction band minimum and the valence band maximum are located at the same k-value in the first Brillouin zone, such as the Γ-point at the zone center, then the semiconductor is called the direct band gap semiconductor. Most of the III-V

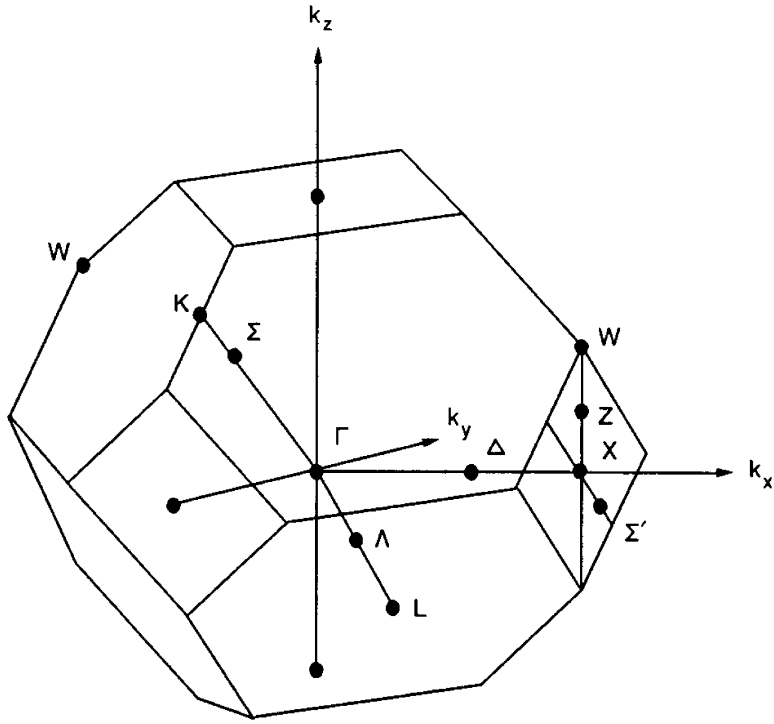


FIGURE 4.12. Symmetry points in the first Brillouin zone of a diamond lattice.

compound semiconductors, such as GaN, GaAs, InP, InAs, and InSb, belong to this category. Direct band gap semiconductors have been widely used in photonic device applications such as laser diodes, LEDs, and photodetectors because their band structures allow for direct optical transitions. They are also widely used in high-speed and high-frequency device applications due to the small electron effective mass and high electron mobility in these materials. If the conduction band minimum and the valence band maximum are not located at the same k-value in the first Brillouin zone, then the semiconductor is referred to as an indirect band gap semiconductor. Elemental semiconductors such as silicon and germanium belong to this category. Table 4.1 lists the energy band gaps and the effective masses of electrons and holes for the elemental and compound semiconductors.

The conduction band of a diamond or a zinc blende crystal usually consists of several subbands or satellite bands. For example, the conduction band minimum of a germanium crystal is located at the zone boundaries along the {111} axes, while for silicon it is located near the zone boundaries along the {100} axes; these are shown in Figures 4.11b and a, respectively. It is noted that the constant-energy surfaces for electrons in silicon and germanium are ellipsoidal energy surfaces, while the constant-energy surface near the conduction band minimum is spherical for GaAs and other III-V compound semiconductors. Figure 4.13 shows a more

TABLE 4.1. Energy Band Gaps and Effective Masses for Elemental and Compound Semiconductors at 300 K.

Element	E_g (eV)	Electron mass (m_e^*/m_0)	Hole mass (m_e^*/m_0)
Si	1.12	$m_t^* = 0.19, m_l^* = 0.97$	$m_{lh}^* = 0.16, m_{hh}^* = 0.50$
Ge	0.67	$m_t^* = 0.082, m_l^* = 1.6$	$m_{lh}^* = 0.04, m_{hh}^* = 0.30$
GaAs	1.43	0.068	$m_{lh}^* = 0.074, m_{hh}^* = 0.62$
AlN	6.10 (WZ)	$m_t^* = 0.33, m_l^* = 0.32$	$m_{hht} = 0.73, m_{hhl} = 3.52;$
	6.15 (ZB)		$m_{lh} = 0.471$
GaN	3.51 (WZ)	$m_t^* = 0.22, m_l^* = 0.20$	$m_{hht} = 0.39, m_{hhl} = 2.04$
	3.35 (ZB)		$m_{lht} = 0.39, m_{lhl} = 0.74$
InN	0.78 (WZ)	$m_t^* = 0.07, m_l^* = 0.06$	$m_{hht} = 0.14, m_{hhl} = 2.09$
	0.70 (ZB)		$m_{hht} = 0.13, m_{hhl} = 0.50$
AlAs	2.16	$m_l = 2.0$	$m_{lh}^* = 0.15, m_{hh}^* = 0.76$
GaP	2.26	$m_l = 1.12, m_t^* = 0.22$	$m_{hh}^* = 0.79, m_{lh}^* = 0.14$
GaSb	0.72	0.045	$m_{hh}^* = 0.62, m_{hl}^* = 0.074$
InP	1.29	0.08	$m_{hh}^* = 0.85, m_{hl}^* = 0.089$
InAs	0.33	0.023	$m_{hh}^* = 0.60, m_{hl}^* = 0.027$
InSb	0.16	0.014	$m_{hh}^* = 0.60, m_{hl}^* = 0.027$
CdS	2.42	0.17	0.60
CdSe	1.70	0.13	0.45
CdTe	1.50	0.096	0.37
ZnSe	2.67	0.14	0.60
ZnTe	2.35	0.18	0.65
ZnS	3.68	0.28	—
PbTe	0.32	0.22	0.29

$+m_t^*$ denotes transverse effective mass, m_l^* longitudinal effective mass, m_{lh}^* light-hole mass, m_{hh}^* heavy-hole mass, and m_0 free-electron mass (9.1×10^{-31} kg).
WZ:Wurtzite structure, ZB: Zincblende structure.

detailed energy band structure of GaAs calculated from the pseudopotential method.[2] The Γ-conduction band minimum is located at the zone center, the L-conduction band valleys are located at $(2\pi/a)(1/2, 1/2, 1/2)$ along the (111) axes, and the X-conduction band valleys are located at the zone boundaries along the (100) axes. The separation between the L-valley and the Γ-band minimum is equal to 0.29 eV. The valence band maxima of the heavy- and light-hole bands are located at the Γ-point in the Brillouin zone center. Therefore, both silicon and germanium are indirect bandgap semiconductors, while GaN, GaAs, InP, and InAs are direct bandgap semiconductors. For silicon, the conduction band minima consist of six ellipsoids of constant-energy surfaces along the {100} axes with the center of each ellipsoidal energy surface located about three-fourths of the distance from the zone center to the zone boundary. For germanium, the conduction band minima consist of eight ellipsoidal constant-energy surfaces along the {111} axes with the center of each ellipsoid located at the zone boundary. Thus, for germanium there are eight half-ellipsoidal conduction band valleys inside the first Brillouin zone. For GaAs, the constant-energy surface of the Γ-conduction band

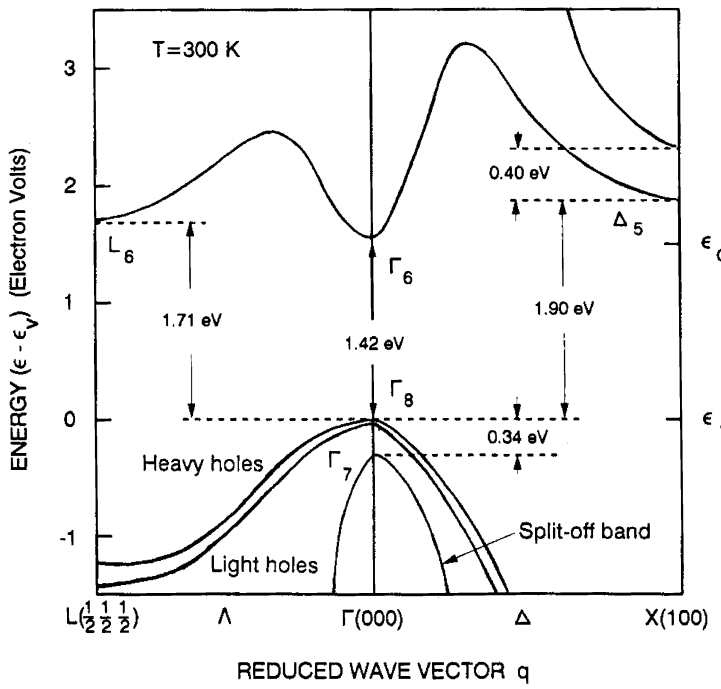


FIGURE 4.13. Detailed energy band diagram for a GaAs crystal calculated from the pseudopotential method, showing both the conduction and valence bands along the (100) and (111) crystal orientation. After Chelikowski and Cohen,[2] by permission.

minimum is spherical, and is located at the zone center. The energy dispersion (i.e., E vs. k) relation for electrons near the bottom of the conduction band can be expressed by

$$E(k) = E_c + \frac{\hbar^2 k^2}{2m_n^*} \tag{4.103}$$

for the spherical constant energy surface, and

$$E(k) = E_c + \frac{\hbar^2}{2}\left(\frac{k_l^2}{m_l} + \frac{k_t^2}{m_t}\right) \tag{4.104}$$

for the ellipsoidal constant-energy surface, where m_l and m_t denote the longitudinal and transverse effective masses of electrons in the conduction band, respectively.

The valence bands of silicon, germanium, and GaAs crystals consist of the heavy- and light-hole bands that are degenerate at $k = 0$. In addition, a spin-orbit split-off band is located at a few tens of meV below the top of the heavy- and light-hole bands. This can be best described using the band structure shown in Figure 4.13 for a GaAs crystal. In this figure, it is shown that the heavy- and light-hole bands are degenerate at the top of the valence band and may be represented

by a parabolic band with different curvatures. The valence band with a smaller curvature (i.e., with a larger hole effective mass) is usually referred to as the heavy-hole band, and the valence band with a larger curvature (i.e., with a smaller hole effective mass) is known as the light-hole band. The effective masses of the light- and heavy-hole bands for Si, Ge, and GaAs are also given in Table 5.3. In general, the energy versus wave vector relation (E vs. k) for the heavy- and light-hole bands near the top of the valence bands is nonparabolic and can be expressed by

$$E(k) = E_{\mathrm{v}} - \frac{\hbar^2 k^2 s(k)}{2m_{\mathrm{p}}^*}, \tag{4.105}$$

where $s(k)$ is given by

$$s(k) = A \pm \left[B^2 + C^2\left(k_x^2 k_y^2/k^4 + k_x^2 k_z^2/k^4 + k_y^2 k_z^2/k^4\right)\right]^{1/2}. \tag{4.106}$$

Note that A, B, and C in (4.106) are constants (see Problem 4.10); the plus and minus signs correspond to the heavy- and light-hole bands, respectively. It should be noted that the constant-energy surfaces near the top of the valence bands are warped and nonparabolic for Si, Ge, GaAs, and other III-V compound semiconductors.

Another important feature of the III-V semiconductor technology is the ability to grow the lattice-matched ternary or quaternary compound semiconductor epitaxial layers on either the GaAs or InP semi-insulating substrates (e.g., $In_xGa_{1-x}P$, $Al_xGa_{1-x}As$, and $In_xGa_{1-x}As_yP_{1-y}$ on GaAs; $In_xGa_{1-x}As$ and $In_xAl_{1-x}As$ on InP substrates). Using these ternary and quaternary compound semiconductors, it is possible to change many important optical, physical, and electrical properties of the III-V compound semiconductors, such as the band gap energy and electron mobility, for a wide variety of applications. In addition, many novel device structures can be fabricated using the binary/ternary superlattice and quantum well heterojunction structures (e.g., $Al_xGa_{1-x}As$/GaAs, InGaAs/AlGaAs). These features are extremely important for many applications in detectors, lasers, LEDs, and high-speed devices using III-V compound semiconductor epitaxial layers grown by the MOCVD and MBE techniques. Figure 4.14 shows the energy band gap versus lattice constant for Si, Ge, II-VI, and III-V binary compound semiconductors.[3] The solid lines denote the direct band gap materials and the dashed lines are for the indirect band gap materials. A mixture of AlP/GaP to form $Al_xGa_{1-x}P$, AlAs/GaAs to form $Al_xGa_{1-x}As$, AlSb/GaSb to form $Al_xGa_{1-x}Sb$ ternary compounds, and InP/GaAs/InAs to form $In_xGa_{1-x}As_yP_{1-y}$ quaternary compound semiconductor along the vertical line of Figure 4.14 yields lattice-matched epitaxial layers grown on the GaP, GaAs, InP, and GaSb substrates, respectively. By tailoring the energy band gap of these III-V alloy systems, it is possible to produce detectors, LEDs, and lasers with wavelengths covering the visible to infrared spectral range. Wide band gap semiconductors such as AlN, SiC, and GaN have been widely investigated in recent years, enabling the fabrication of various electronic devices for microwave, high-temperature, and high-power applications. Furthermore, GaN-based ternary compounds such as $Al_xGa_{1-x}N$ and $In_xGa_{1-x}N$ with the energy band gaps varying from 0.7 to 6.2 eV have been developed for UV detectors, laser diodes, and LED

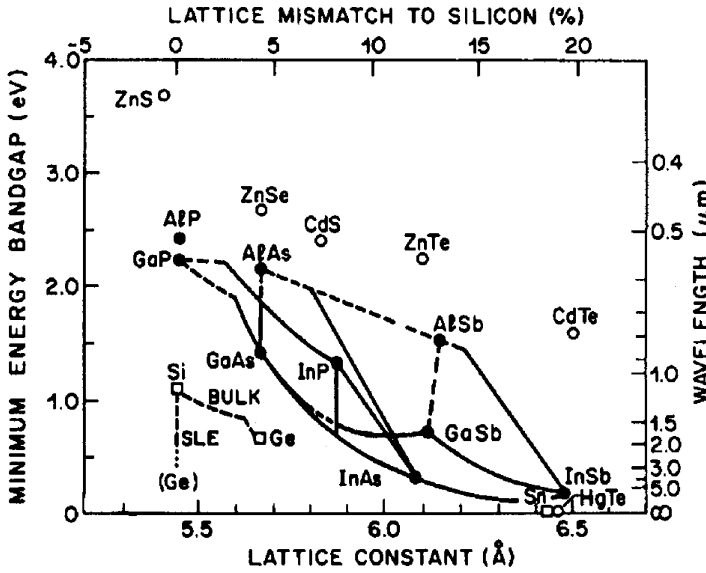


FIGURE 4.14. The energy band gap versus lattice constant for III-V binary compound semiconductors. Solid lines denote the direct band gap materials and dashed lines the indirect band gap materials. Vertical lines are for the lattice-matched ternary compound semiconductors on a selected binary compound semiconductor substrate. After Hansen,[3] reprinted by permission from John Wiley & Sons Inc.

applications. Figures 4.15a and b show the energy band gap versus alloy composition x for $GaAs_xP_{1-x}$ and $Al_xGa_{1-x}As$ ternary compound semiconductors, which illustrate the band gap variation from $E_g = 1.42$ to 2.25 eV and 2.19 eV, respectively, as x varies from 1 to 0. The variation of band gap with alloy composition in a III-V ternary material system can be estimated using an empirical formula given by

$$E_g(x) = E_g(0) + bx + cx^2, \tag{4.107}$$

where b is a fitting parameter, and c is called the bowing parameter, which may be calculated theoretically or determined experimentally. For the $Al_xGa_{1-x}As$ material system, the energy band gap for the Γ-, X-, and L-valleys as a function of alloying composition x can be expressed as

$$E_g^{\Gamma}(x) = 1.425 + 1.247x + 1.147(x - 0.45)^2, \tag{4.108a}$$

$$E_g^{X}(x) = 1.90 + 0.125x + 0.143x^2, \tag{4.108b}$$

$$E_g^{L}(x) = 1.708 + 0.642x. \tag{4.108c}$$

It is noted that $Al_xGa_{1-x}As$ becomes indirect band gap material for $x \geq 0.43$, and $GaAs_xP_{1-x}$ becomes indirect band gap material for $x \geq 0.45$. In general, many

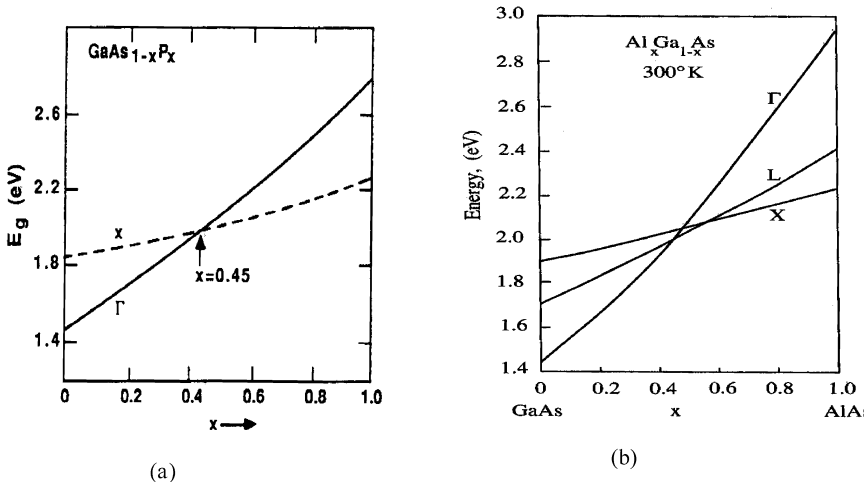


FIGURE 4.15. Energy band gap E_g versus alloy composition x for (a) $GaAs_xP_{1-x}$ and (b) $Al_xGa_{1-x}As$ material systems.

physical parameters of ternary compounds are determined by the parameters of the constituent binaries and vary roughly linearly with the composition. For example, in a ternary compound semiconductor, the lattice constant varies linearly with the composition; this also holds true for the quaternary alloys.

The energy band structures presented in this section are extremely important for understanding the physical, optical, and electrical properties of semiconductor materials and devices. The energy band structures for semiconductors presented in this section will be used in explaining the physical and transport properties of a wide variety of semiconductor devices to be discussed throughout this book.

4.8. The Effective Mass Concept for Electrons and Holes

As described in Section 4.1, the most generalized solution of the electron wave functions in a periodic crystal is a plane wave modulated by the Bloch function, $u_k(r)$. For the time-dependent electron wave functions, this can be written as

$$\phi_k(r,\ t) = u_k(r)\, e^{i(k\cdot r-\omega t)}. \tag{4.109}$$

Since the wave function for a Bloch-type wave packet extends over the entire crystal lattice, the group velocity for such a wave packet is given by

$$v_g = \frac{d\omega}{dk} = \left(\frac{1}{\hbar}\right)\nabla_k E(k). \tag{4.110}$$

Note that the electron energy $E(k) = \hbar\omega$ is used in (4.110) to define the group velocity, v_g. According to (4.110), the group velocity of an electron wave packet is in the direction perpendicular to the constant-energy surface at a given wave vector

k in k-space. The group velocity can be determined by the gradient of energy with respect to the wave vector k.

If a Lorentz force $\boldsymbol{F}$, which may be due to either an electric field or a magnetic field, is applied to the electrons inside a crystal, then the wave vector of electrons will change with the applied Lorentz force according to the following relation:

$$\boldsymbol{F} = -q\left(\boldsymbol{\mathcal{E}} + \boldsymbol{v}_{\mathrm{g}} \times \boldsymbol{B}\right) = \hbar\left(\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t}\right) = \hbar\dot{k}, \tag{4.111}$$

where $\boldsymbol{\mathcal{E}}$ is the electric field, and $\boldsymbol{B}$ is the magnetic flux density. The product $\hbar\dot{k}$ is referred to as the change of crystal momentum. Equation (4.111) shows that the external applied force acting on an electron tends to change the crystal momentum or the electron wave vector in a crystal lattice. The electron effective mass in a crystal lattice can be defined by

$$\boldsymbol{F} = m_n^*\boldsymbol{a} = m_n^*\left(\frac{\mathrm{d}\boldsymbol{v}_{\mathrm{g}}}{\mathrm{d}t}\right). \tag{4.112}$$

Solving (4.110) through (4.112), one obtains an expression of acceleration for electrons due to the applied Lorentz force, which is given by

$$\boldsymbol{a} = \frac{\mathrm{d}\boldsymbol{v}_g}{\mathrm{d}t} = \left(\frac{1}{\hbar}\right)\left(\frac{\mathrm{d}\nabla_k E}{\mathrm{d}k}\right)\left(\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t}\right) = \left(\frac{1}{\hbar^2}\right)\left(\frac{\mathrm{d}^2 E}{\mathrm{d}k^2}\right)\cdot\boldsymbol{F}. \tag{4.113}$$

Solving (4.112) and (4.113), one obtains an expression for the reciprocal effective mass tensor for electrons, whose component is given by

$$\left(m_n^*\right)_{ij}^{-1} = \left(\frac{1}{\hbar^2}\right)\left(\frac{\partial^2 E(k)}{\partial k_i\,\partial k_j}\right), \tag{4.114}$$

where $i, j = 1, 2, 3\ldots$ are the indices used to define the crystal orientations. From (4.114), it is noted that the reciprocal effective mass is directly proportional to the curvature of the energy band structure in the E versus k plot. A large curvature near the conduction band minimum implies a small effective mass of electrons, and vice versa. For example, a comparison of the curvatures of the energy band diagrams near the bottom of the conduction band for silicon and GaAs (see Figure 4.11) shows that silicon has a smaller curvature than GaAs near the conduction band minimum, and hence has a larger electron effective mass than that of a GaAs crystal.

Another important concept to be discussed in this section is concerned with holes in the valence bands of a semiconductor. A hole in the valence band marks the absence of a valence electron or the creation of an empty state in the valence band. Furthermore, the motion of a hole can be regarded as the motion of a missing electron in the valence band. Since most of the holes reside near the top of the valence band maximum in which the curvature of the E versus k diagram is always negative, which implies a negative electron effective mass, it is appropriate to replace the missing electrons by the positively charged holes. This arrangement greatly simplifies the treatment of electronic conduction in the valence band of a semiconductor. By using the concept of holes, which have a positive effective

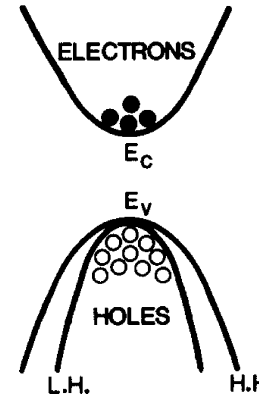


FIGURE 4.16. Electrons near the bottom of the conduction band and holes near the top of the valence band.

mass and a positive charge, the inverse hole effective mass can be derived from the expression

$$\frac{\mathrm{d}\boldsymbol{v}_{\mathrm{g}}}{\mathrm{d}t} = -\left(\frac{1}{m_{\mathrm{n}}^*}\right)\boldsymbol{F} = \left(\frac{1}{\hbar^2}\right)\nabla_{k'}^2 E_{k'} \cdot \boldsymbol{F} = \left(\frac{1}{m_{\mathrm{h}}^*}\right)\boldsymbol{F}, \tag{4.115}$$

which yields

$$\frac{1}{m_{\mathrm{h}}^*} = \left(\frac{1}{\hbar^2}\right)\nabla_{k'}^2 E_{k'}, \tag{4.116}$$

where $\boldsymbol{F}$ is the Lorentz force experienced by a hole. Thus, a hole in the valence band may be considered as a charged particle with a positive charge q and a positive effective mass m_{h}^*. Figure 4.16 shows the electrons near the bottom of the conduction band and holes near the top of the valence band.

The effective mass concept presented above is particularly useful for describing the transport properties of a semiconductor. In a semiconductor, most of the electrons reside near the bottom of the conduction band, and holes are located near the top of the valence bands. If the energy band structures near the bottom of the conduction band and the top of the valence bands have spherical constant-energy surfaces, then the effective masses for both electrons and holes are given by a scalar quantity. If one assumes that both the conduction band minimum and the valence band maximum are located at $k = 0$ (i.e., at the zone center (Γ-point)), then the $E-k$ relation can be expressed by

$$E_k = E_{\mathrm{c}} + \frac{\hbar^2 k^2}{2m_{\mathrm{n}}^*} \tag{4.117}$$

for electrons in the conduction band, and

$$E_{k'} = E_{\mathrm{v}} - \frac{\hbar^2 k'^2}{2m_{\mathrm{h}}^*} \tag{4.118}$$

for holes in the valence bands. Both the heavy- and light-hole bands degenerate into a single band at the top of the valence band edge.

Equations (4.117) and (4.118) may be used to describe the E–k relation for electrons near the bottom of the conduction bands and holes near the top of the valence bands with parabolic band structure. These relations are valid for direct band gap semiconductors such as GaAs, InP, and InAs, in which the constant-energy surfaces near the conduction band minimum and the valence band maximum are assumed parabolic. If the constant-energy surface near the band edge is nonparabolic, then an effective mass tensor given by (4.116) should be used instead. For silicon and germanium, the constant-energy surface near the bottom of the conduction band is ellipsoidal, and the electron effective mass may be expressed in terms of its transverse and longitudinal effective masses (i.e., m_t^* and m_l^*). Both these masses can be determined using the cyclotron resonance experiment performed at very low temperature. The effective masses of electrons and holes for some practical semiconductors are listed in Table 4.1. Using the effective mass concept for electrons in the conduction band and holes in the valence bands, one can treat both the electrons and holes as quasifree particles, which in turn greatly simplify the mathematics of solving the carrier transport problems in a semiconductor.

4.9. Energy Band Structures and Density of States for Low-Dimensional Systems

In this section the band structure and the density of states for a heterostructure superlattice are discussed. In addition, the density of states functions for the low-dimensional systems (0-D,1-D, 2-D, Q1-D, Q2-D systems) are also presented. With the advent of molecular beam epitaxy (MBE) and metal-organic chemical vapor deposition (MOCVD) growth techniques, it is now possible to grow high-quality III-V semiconductor epitaxial layers composed of alternating material systems (e.g., AlGaAs/GaAs, InAlAs/InGaAs) with a thickness of few atomic layers. As a result, extensive studies of the fundamental properties of superlattices, such as energy band structures and carrier transport in the growth direction of the superlattice layers, have been widely reported in recent years. Novel devices such as semiconductor lasers, infrared detectors, LEDs, and modulators using quantum well/superlattice structures have been developed. Unlike the three-dimensional (3-D) system in which the size of the sample in the x, y, z directions is much larger than the de Broglie wavelength (i.e., $L_x, L_y, L_z \gg \lambda_e$), the thickness of a two-dimensional (2-D) system along the growth direction is smaller than the de Broglie wavelength ($d \leq \lambda_e$). For a GaAs crystal, this corresponds to a layer thickness of 25 nm or less at 300 K. In a 2-D system, carrier confinement occurs along the growth direction in which the layer thickness is comparable to the de Broglie wavelength, but retains quasifree-electron behavior within the plane of the superlattice.

A superlattice structure is formed when thin layers ($d \leq 25\,\text{nm}$) of a larger band gap semiconductor (e.g., AlGaAs) and a smaller band gap semiconductor (e.g., GaAs) are grown alternately on a conducting or a semi-insulating substrate. The periodic structure formed by alternate deposition of thin epitaxial layers of

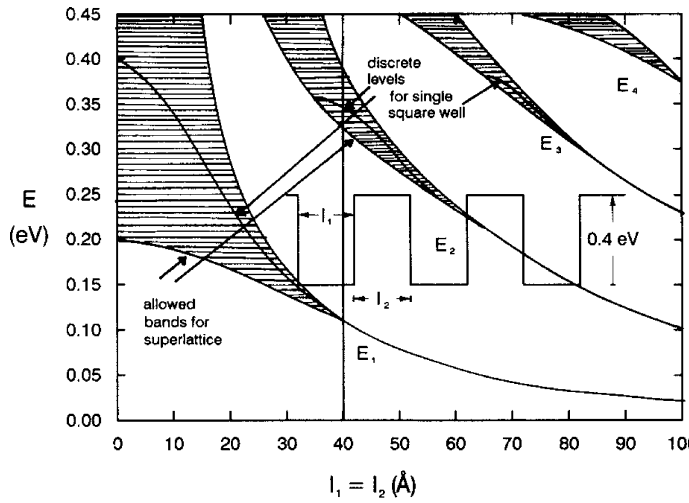


FIGURE 4.17. Calculated widths of minibands and intermittent gaps as a function of the period length for a symmetrical well/barrier heterostructure (e.g., AlGaAs/GaAs). After Esaki,[4] by permission.

two different band gap materials produces a periodic potential similar to the 1-D Kronig–Penney potential discussed in Section 4.3. A potential barrier is formed between a larger band gap material (AlGaAs) and a smaller band gap material (GaAs), while a potential well is formed in the smaller band gap material sandwiched between two wide band gap materials. The energy band diagram for the superlattice is similar to that of free electrons exposed to a periodic crystal potential, except that now the periodic potential is imposed on Bloch electrons with an effective mass m_n^*. Depending on the width of the superlattice, the energy states inside the quantum well could be discrete bound states or minibands. Figure 4.17 shows the calculated widths of minibands and intermittent gaps as a function of the period length (i.e., $l = l_1 + l_2$) for a symmetric barrier/quantum well structure with a barrier height of 0.4 eV.[4] It is noted that for an equal barrier/well width (i.e., $l_1 = l_2 = 4\,\text{nm}$) superlattice, the lowest band is extremely narrow and lies 100 meV above the bottom of the quantum well. The second miniband extends from 320 to 380 meV, while higher bands overlap above the top of the potential barrier.

Figure 4.18 shows (a) the first and second minibands inside the conduction band of a superlattice along the growth direction (i.e., the z-direction), (b) minibands and minigaps in the k_z-direction inside the Brillouin zone, and (c) energy (E_1 and E_2) versus wave vector k in the k_x- and k_y-directions (i.e., in the plane of the superlattice). It is seen that within the conduction band, we observe a subband structure of minibands across the potential barrier and the quantum well; the higher minibands extend beyond the height of potential barriers. The lower minibands inside the well are separated by the minigaps in the direction of superlattice periodicity (i.e., the z-direction). Within the plane of the superlattice layers (i.e., the x–y plane), the electron wave functions experience only the regular periodic lattice potential. Therefore, the energy dispersion relations (i.e., E vs. k_x and k_y)

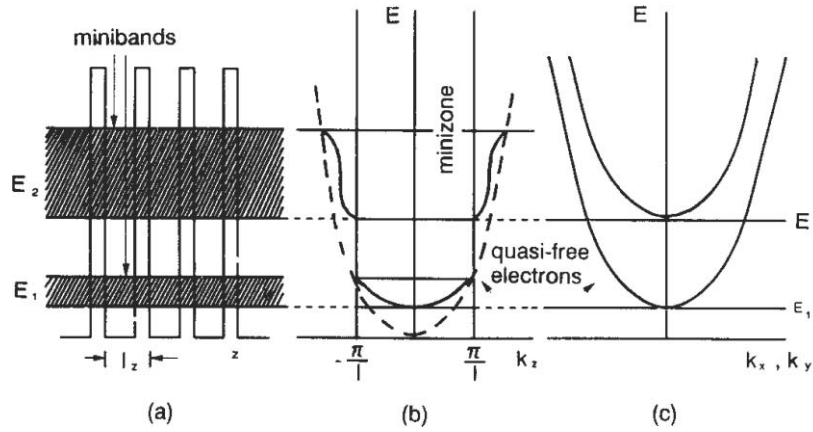


FIGURE 4.18. (a) Minibands in the growth direction, z, of the superlattice layers, (b) minibands and gaps in the k_z direction (perpendicular to the superlattice layer) inside the Brillouin zone, and (c) energy dispersion curves in the k_x and k_y directions (i.e., in the plane of the superlattice later), which show the continuous states within the Brillouin zone for the E_1 and E_2 levels.

are similar to that of the unperturbed crystal lattice except for mixing the states in the z-direction, which results in lifting the lowest energy states at $k = 0$ above E_c of the bulk well material as shown in Figures 4.18b and c. The second miniband results in a second shifted parabola along the k_x- and k_y-directions. It is seen that the E versus k relation in the k_x–k_y plane is continuous, while a minigap between the first and second minibands appears in the direction perpendicular to the superlattice (k_z). Formation of the miniband in a superlattice can be realized when the wave functions of carriers in the neighboring quantum wells of a multilayer heterostructure overlap significantly. The energy levels broaden into minibands with extended Bloch states. These minibands are expected to lead to the transport of carriers perpendicular to the superlattice layers, which include tunneling, resonant tunneling, ballistic and miniband transport.

Calculations of energy band structures in a superlattice can be carried out by several methods. These include the pseudopotential, tight-binding (LCAO), and envelope-function (i.e., $k \cdot p$) methods. Among these methods, the envelope-function approach is most widely used due to its simplicity. With several refinements, this method can become quite effective in dealing with many problems such as band mixing, the effects of external fields, impurities, and exciton states. A detailed description of the envelope-function approximation for calculating the energy bands in superlattice heterostructure devices has been given by Altarelli.[5,6]

The density of states in the minibands of a superlattice is discussed next. It is shown in Figure 4.19 that the density-of-states function has a staircase character (dashed steps) for the isolated quantum wells (i.e., the barrier width is much larger than the well width).[4] In this case, each level can be occupied by the number of electrons given by its degeneracy multiplied by the number of atoms in the

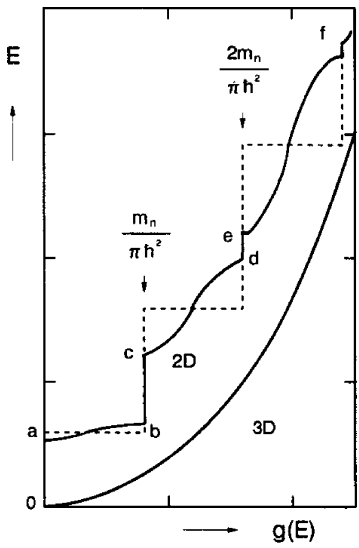


FIGURE 4.19. Staircase density of states for the isolated 2-D quantum wells (dashed line), the superlattice (distorted solid line), and the 3-D system with a parabolic band. After Esaki,[4] by permission.

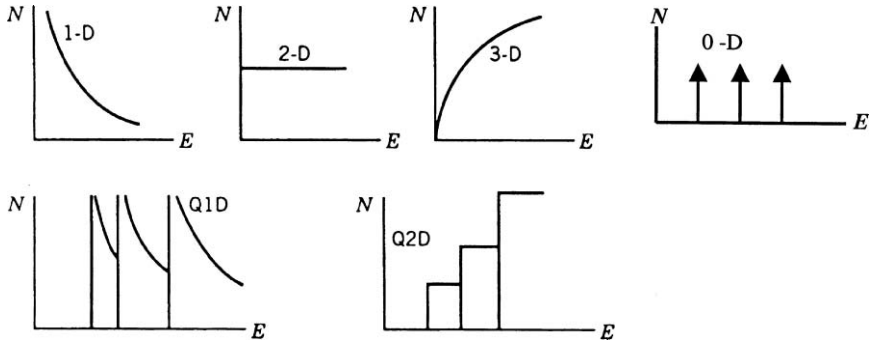


FIGURE 4.20. The density of states functions for the 0-D, 1-D, 2-D, 3-D, Q1-D, and Q2-D systems. After M.J. Kelly.[7]

quantum well. Thus, the two-dimensional (2-D) density of states, $g(E)$, in each discrete level can be described by

$$g(E) = \frac{m_n^*}{\pi\hbar^2}, \tag{4.119}$$

where $g(E)$ is measured in cm^{-2}. Equation (4.119) shows that $g(E)$ for a 2-D system is a constant and independent of energy. When significant overlap occurs, tunneling becomes possible and each energy level splits into minibands, and the staircase behavior (dashed line) changes shape as shown by the solid curly line in Figure 4.19. For comparison, the density-of-states function for a 3-D system is also included in Figure 4.19 for a parabolic band. The density of states functions for other low-dimensional systems has also been published in the literature. Figure 4.20

shows the plots of density of states function versus energy for the 3-D, 2-D, 1-D, Q1-D (quantum wire), and Q2-D (quantum well), and the 0-D (quantum dot) systems. The density of states functions for the low-dimensional systems are given respectively as follows:[7]

$$N(E) = \frac{2L^3(2m^*)^{3/2}}{(2\pi l)^2\hbar^3}E^{1/2}, \qquad \text{(3-D)} \qquad (4.120)$$

$$N(E) = \frac{L^2 2m^*}{\pi\hbar^2}, \qquad \text{(2-D)} \qquad (4.121)$$

$$N(E) = \frac{L(2m^*)^{1/2}}{2\pi\hbar^2}E^{-1/2}, \qquad \text{(1-D)} \qquad (4.122)$$

$$N(E) = \sum_n \frac{L^2(2m^*)}{\pi\hbar^2}H(E - E_n), \qquad \text{(Q2-D)} \qquad (4.123)$$

$$N(E) = \sum_{lm} \frac{L(2m^*)^{1/2}}{2\pi\hbar^2}(E - E_{lm})^{-1/2}H(E - E_{lm}), \qquad \text{(Q1-D)} \qquad (4.124)$$

$$N(E) = \delta(E - E_{lmn}), \qquad \text{(0-D)} \qquad (4.125)$$

where $H(\sigma)$ is the Heaviside function ($H(\sigma) = 1$ for $\sigma > 0$; $H(\sigma) = 0$ for $\sigma < 0$). The electron wave functions ϕ_n, ϕ_{nl}, and ϕ_{nmn}, and the energy levels E_n, E_{ln}, E_{lnm} for the 2-D, 1-D, and 0-D systems are given respectively by[7]

$$\phi_n = \left(\frac{2}{L}\right)^{1/2}\sin\left(\frac{n\pi z}{L}\right), \qquad E_n = \frac{\hbar^2}{2m^*}\left[\frac{n\pi}{L}\right]^2, \qquad \text{(2-D)} \qquad (4.126)$$

$$\phi_{lm} = \left(\frac{2}{L}\right)\sin\left(\frac{l\pi x}{L}\right)\sin\left(\frac{m\pi y}{L}\right),$$

$$E_{l,n} = \frac{\hbar^2}{2m^*}\left[\frac{l\pi}{L} + \frac{n\pi}{L}\right]^2, \qquad \text{(1-D)} \qquad (4.127)$$

$$\phi_{lmn} = \left(\frac{2}{L}\right)^{3/2}\sin\left(\frac{l\pi x}{L}\right)\sin\left(\frac{m\pi y}{L}\right)\sin\left(\frac{n\pi z}{L}\right),$$

$$E_{l,m,n} = \frac{\hbar^2}{2m^*}\left[\frac{l\pi}{L} + \frac{m\pi}{L} + \frac{n\pi}{L}\right]^2. \qquad \text{(0-D)} \qquad (4.128)$$

Equations (4.125) and (4.128) are the density of states function, the wave functions, and energy levels for the 0-D system (quantum dots), respectively. The density of states function is a very important function for calculating many response functions and the transport parameters such as thermoelectric power, thermal conductivity, electrical conductivity, and Hall coefficients, which are all dependent on the density of states at the Fermi energy (E_F). The concept of miniband and the density of states function in a superlattice and in the low-dimensional systems described in this section may be used in the design of various quantum-effect devices using multiple quantum well (QW) and quantum dot (QD) heterostructures grown by

the MBE and MOCVD techniques, as will be dicussed later in Chapters 12, 13, 14, and 16.

Problems

4.1. Using the nearly free-electron approximation for a one-dimensional (1-D) crystal lattice and assuming that the only nonvanishing Fourier coefficients of the crystal potential are $v(\pi/a)$ and $v(-\pi/a)$ in (4.73), show that near the band edge at $k = 0$, the dependence of electron energy on the wave vector k is given by

$$E_k = E_0 + \frac{\hbar^2 k^2}{2m^*},$$

where $m^* = m_0[1 - (32m_0^2a^4/h^4\pi^4)v(\pi/a)^2]^{-1}$ is the effective mass of the electron at $k = 0$.

4.2. The E–k relation of a simple cubic lattice given by (4.79) is derived from the tight-binding approximation. Show that near $k \approx 0$ this relation can be expressed by

$$E_k = E_{n0} + \frac{\hbar^2 k^2}{2m^*},$$

where $m^* = \hbar^2/2\beta_n a^2$.
And for $k \approx \pi/a$, show that the E–k relation is given by

$$E_k = E_{n0} + \frac{\hbar^2 k^2}{2m^*},$$

where $m^* = -\hbar^2/2\beta_n a^2$.

4.3. If the conductivity and the density-of-states effective masses of electrons are defined, respectively, by

$$m^*_{\text{cn}} = 3(1/m^*_{\text{l}} + 2/m^*_{\text{t}})^{-1} \quad \text{and} \quad m^*_{\text{dn}} = v^{2/3}(m^*_{\text{l}}\, m^*_{\text{t}})^{1/3},$$

where m^*_{l} and m^*_{t} denote the longitudinal and transverse effective masses, respectively, find the conductivity effective mass m^*_{cn} and the density-of-states effective mass m^*_{dn} for Si and Ge crystals. Given: $m^*_{\text{t}} = 0.19m_0$, $m^*_{\text{l}} = 0.97m_0$, $v = 6$ for silicon; and $m^*_t = 0.082m_0$, $m^*_{\text{l}} = 1.64m_0$, $v = 4$ for Ge.

4.4. Explain why most of the III-V compound semiconductors such as GaAs, InP, and InSb have smaller electron effective masses than those of silicon and germanium.

4.5. Sketch the constant-energy contours for a two-dimensional square lattice using the expression derived from the tight-binding approximation

$$E(k) = E_0 + B\cos(k_x a/2)\cos(k_y a/2).$$

4.6. Derive expressions for the group velocity (v_{g}), acceleration ($\text{d}v_{\text{g}}/\text{d}t$), and the effective mass (m^*) of electrons using the E–k relation for the two-dimensional square lattice described in Problem 4.5. If $\cos(k_y a/2) = 1$, plot E, v_{g}, $\text{d}v_{\text{g}}/\text{d}t$, and m^* versus k for the one-dimensional (1-D) crystal lattice.

4.7. If the E–k relation for a simple cubic lattice corresponding to an atomic state derived by the tight-binding approximation is given by

$$E(k) = E_0 - E_0' - 2E'(\cos k_1 a + \cos k_2 a + \cos k_3 a),$$

derive the expressions of (i) group velocity, (ii) acceleration, and (iii) the effective mass tensor.

4.8. Repeat Problem 4.7 for a body-centered cubic lattice (s-like states). (See (4.84).)

4.9. Using the tight-binding approximation, derive the E–k relation for the s-like states in a face-centered cubic lattice.

4.10. The E–k relation near the top of the valence band maximum for silicon and germanium is given by

$$E(k) = -\left(\frac{\hbar^2}{2m}\right)\left\{Ak^2 \pm \left[B^2k^4 + C^2\left(k_1^2k_2^2 + k_2^2k_3^2 + k_3^2k_1^2\right)\right]^{1/2}\right\},$$

where E is measured from the top of the valence band edge. Plus refers to the heavy-hole band and minus is for the light-hole band.

	A	B	C
Ge	13.1	8.3	12.5
Si	4.0	1.1	4.1

Using the values of A, B, and C for germanium and silicon given in the above table, plot the constant-energy contours for the heavy- and light-hole bands in silicon and germanium.

4.11. Plot the energy bandgap (E_g) versus temperature (T) for the E_Γ, E_L, and E_X conduction minima of GaAs crystal for $0 < T < 1000$ K. Given:

$$E_\Gamma(T) = 1.519 - \frac{5.405 \times 10^{-4}T^2}{(T + 204)},$$

$$E_L(T) = 1.815 - \frac{6.05 \times 10^{-4}T^2}{(T + 204)},$$

$$E_X(T) = 1.981 - \frac{4.60 \times 10^{-4}T^2}{(T + 204)} \quad \text{(eV)}.$$

4.12. Plot the energy band gap as a function of pressure (P) for the E_Γ, E_L, and E_X conduction minima of GaAs for $0 < P < 50$ bars. At what pressure P will GaAs become an indirect band gap material? Given:

$$E_\Gamma = E_\Gamma(0) + 0.0126P - 3.77 \times 10^{-5}P^2, \quad \text{(eV)}$$

$$E_L = E_L(0) + 0.0055P,$$

$$E_X = E_X(0) - 0.0015P.$$

4.13. Referring to the paper by J. R. Chelikowsky and M. L. Cohen,[2] describe briefly the pseudopotential method for calculating the energy band structures of semiconductors with diamond and zinc blende structures.

4.14. Plot the energy, group velocity, and inverse effective mass of electrons versus the wave vector in the first Brillouin zone of a one-dimensional crystal lattice, using the relation $E = \hbar^2 k^2 / 2m_0$.

4.15. Using the one-dimensional (1-D) Schrödinger equation, derive the expressions of quantized energy states for (i) an infinite square well (with well width $a = 100$ Å), (ii) triangular well, and (iii) parabolic well. Assuming that the quantization occurs in the z-direction and the potential energies for the three cases are given by (i) $U(z) \to \infty$ (ii) $U(z) = q\mathcal{E}z$ (where $\mathcal{E}$ is the electric field inside the triangular well), and (iii) $U(z) = m^*(\omega^2/2)z^2$, calculate the energy levels of the ground state and the first excited state of (i) and (ii). Given: $m^* = 0.067m_0$, $a = 100$ Å, and $\mathcal{E} = 10^5$ V/cm. *Answer*:

$$\text{(i) } E_r = \frac{\hbar^2\pi^2}{8m^*a^2}(r+1)^2,$$

$$r = 0,\ 1,\ 2, \ldots,\ E_0 = 56 \text{ meV},\ E_1 = 224 \text{ meV}$$

$$\text{(ii) } E_r = \left(\frac{\hbar^2 q^2 \mathcal{E}^2}{2m^*}\right)^{1/3} \left[\frac{3\pi}{2}(r+3/4)\right]^{2/3},$$

$$E_0 = 87 \text{ meV},\ E_1 = 153 \text{ meV},$$

$$\text{(iii) } E_r = \hbar\omega(r+1/2).$$

References

1. M. L. Cohen and Bergstrasser, *Phys. Rev.* **141**, 789–796 (1966).
2. J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **14**(2), 556 (1976).
3. M. Hansen, *Constitution of Binary Alloys*, McGraw-Hill, New York (1958).
4. L. Esaki, in: *The Technology and Physics of Molecular Beam Epitaxy* (E. M. C. Parker, ed.), Plenum Press, New York (1985), p. 143.
5. M. Altarelli, *Phys. Rev. B* **32**, 5138 (1985).
6. M. Altarelli, in: *Heterojunctions and Semiconductor Superlattices* (G. Allen et al., eds.), Springer-Verlag, Berlin (1986).
7. M. J. Kelly, in: *Physics and Technology of Submicron Structures* (G. Bauer, F. Kuchar, and H. Heirich, eds.), Springer-Verlag, Berlin (1987), pp. 174–196.

Bibliography

J. S. Blakemore (ed.), *Key Papers in Physics: GaAs*, American Institute of Physics, New York (1987).

F. J. Blatt, *Physics of Electronic Conduction in Solids*, McGraw-Hill, New York (1968).

R. H. Bube, *Electronic Properties of Crystalline Solids*, Academic Press, New York (1974).

J. Callaway, *Energy Band Theory*, Academic Press, New York (1964).

J. Callaway, *Quantum Theory of the Solid State*, Parts A & B, Academic Press, New York (1974).

C. Kittel, *Introduction to Solid State Physics*, 5th ed., Wiley, New York (1976).

H. Kroemer, *Quantum Mechanics for Engineering, Material Science, and Applied Physics,* Prentice Hall, Englewood Cliffs, New Jersey (1994).

R. Kubo and T. Nagamiya, *Solid State Physics*, McGraw-Hill, New York (1969).

Robert F. Pierret, *Advanced Semiconductor Fundamentals,* 2nd edition, Vol. VI, Prentice Hall, New Jersey (2003).
K. Seeger, *Semiconductor Physics*, 3rd ed., Springer-Verlag, Berlin/Heidelberg (1985).
J. C. Slater, *Quantum Theory of Molecules and Solids*, Vols. 1, 2, and 3, McGraw-Hill, New York (1963).
S. Wang, *Solid State Electronics*, McGraw-Hill, New York (1966).
J. M. Ziman, *Principles of the Theory of Solids*, Cambridge University Press, Cambridge (1964).

5
Equilibrium Properties of Semiconductors

5.1. Introduction

In this chapter, the equilibrium properties of semiconductors are presented. The fact that electrical conductivity of a semiconductor can be readily changed by many orders of magnitude through the incorporation of foreign impurities has made the semiconductor one of the most intriguing and unique electronic materials among all the crystalline solids. The invention of germanium and silicon transistors in the early 1950s and the silicon integrated circuits in the 1960s as well as the development of microprocess chips in the 1980s has indeed transformed semiconductors into the most important and indispensable electronic materials of modern times.

Unlike metals, the electrical conductivity of a semiconductor can be changed by many orders of magnitude by simply doping it with acceptor or donor impurities or by using external excitations (e.g., by photoexcitation). At low temperatures, a pure semiconductor may become a perfect electrical insulator, since its valence band is totally filled with valence electrons and the conduction band is completely empty. However, as the temperature rises, a fraction of the valence electrons are excited into the conduction band by the thermal energy, thus creating free holes in the valence band. As a result, the electrical conductivity will increase rapidly with increasing temperature. Therefore, even an intrinsic semiconductor may become a good electrical conductor at high temperatures. In general, the semiconductors may be divided into two categories: the pure undoped semiconductor, which is usually referred to as the intrinsic semiconductor, and the doped semiconductor, which is also called the extrinsic semiconductor. Another distinct difference between a metal and a semiconductor is that the electrical conduction in a metal is due to electrons, while the electrical conduction of a semiconductor may be attributed to electrons, holes, or both carriers. The electrical conduction of an intrinsic semiconductor is due to both electrons and holes, while for an extrinsic semiconductor it is usually dominated by either electrons or holes, depending on whether the semiconductor is doped with shallow donors or shallow acceptor impurities.

To understand the conduction mechanisms in a semiconductor, the equilibrium properties of a semiconductor are first examined. A unique feature of

semiconductor materials is that the physical and transport parameters depend strongly on temperature. For example, the intrinsic carrier concentration of a semiconductor depends exponentially on temperature. Other physical parameters, such as carrier mobility, resistivity, and the Fermi level in a nondegenerate semiconductor, are likewise a strong function of temperature. In addition, both the shallow-level and deep-level impurities may also play an important role in controlling the physical and electrical properties of a semiconductor. For example, the equilibrium carrier concentration of a semiconductor is controlled by the shallow-level impurities, and the minority carrier lifetimes are usually closely related to defects and deep-level impurities in a semiconductor.

In Section 5.2, the general expressions for the electron density in the conduction band and the hole density in the valence band are derived for the cases of the single spherical energy band and the multivalley conduction bands. The equilibrium properties of an intrinsic semiconductor are described in Section 5.3. Section 5.4 presents the equilibrium properties of n-type and p-type extrinsic semiconductors. The change of the conduction mechanism from the intrinsic to n-type (electrons) or p-type (holes) conduction by doping the semiconductors with shallow-donor or shallow-acceptor impurities is discussed in this section. Section 5.5 deals with the physical properties of a shallow-level impurity. Using Bohr's model for a hydrogen-like impurity atom the ionization energy of a shallow-level impurity is derived in this section. In Section 5.6, the Hall effect and the electrical conductivity of a semiconductor are discussed. Finally, the heavy doping effects such as carrier degeneracy and band gap narrowing for degenerate semiconductors are discussed in Section 5.7.

5.2. Densities of Electrons and Holes in a Semiconductor

General expressions for the equilibrium densities of electrons and holes in a semiconductor can be derived using the Fermi–Dirac (F-D) distribution function and the density-of-states function described in Chapter 3. For undoped and lightly doped semiconductors, the Maxwell–Boltzmann (M-B) distribution function is used instead of the F-D distribution function. If one assumes that the constant-energy surfaces near the bottom of the conduction band and the top of the valence band are spherical, then the equilibrium distribution functions for electrons in the conduction band and holes in the valence band may be described in terms of the F-D distribution function. The F-D distribution function for electrons in the conduction band is given by

$$f_{\mathrm{n}}(E) = \frac{1}{[1 + \mathrm{e}^{(E - E_{\mathrm{f}})/k_{\mathrm{B}}T}]}, \tag{5.1}$$

and the F-D distribution function for holes in the valence band can be expressed by

$$f_{\mathrm{p}}(E) = \frac{1}{[1 + \mathrm{e}^{(E_{\mathrm{f}} - E)/k_{\mathrm{B}}T}]}. \tag{5.2}$$

The density-of-states function given by (3.33) for free electrons in a metal can be applied to electrons in the conduction band and holes in the valence band of a semiconductor. Assuming parabolic bands for both the conduction and valence bands and using the conduction and valence band edge as a reference level, the density of states function per unit volume in the conduction band can be expressed by

$$g_{\mathrm{n}}(E - E_{\mathrm{c}}) = \left(\frac{4\pi}{h^3}\right)(2m_{\mathrm{n}}^*)^{3/2}(E - E_{\mathrm{c}})^{1/2}. \tag{5.3}$$

The density of states in the valence band is given by

$$g_{\mathrm{p}}(E_{\mathrm{v}} - E) = \left(\frac{4\pi}{h^3}\right)(2m_{\mathrm{p}}^*)^{3/2}(E_{\mathrm{v}} - E)^{1/2}. \tag{5.4}$$

Figure 5.1 shows a plot of $f_{\mathrm{n}}(E)$, $f_{\mathrm{p}}(E)$, $g_{\mathrm{n}}(E)$, $g_{\mathrm{p}}(E)$, $f_{\mathrm{n}}(E)g_{\mathrm{n}}(E)$, and $f_{\mathrm{p}}(E)g_{\mathrm{p}}(E)$ versus energy E in the conduction and valence bands for $T > 0$ K. The hatched area denotes the electron density in the conduction band and hole density in the valence band, respectively, E_{c} is the conduction band edge; E_{v} is the valence band edge, and E_{g} is the band gap energy. The equilibrium electron density n_0 in the conduction band can be obtained by integrating the product $\mathrm{d}n = f_{\mathrm{n}}(E)g_{\mathrm{n}}(E)\mathrm{d}E$ (i.e., the electron density per unit energy interval) with respect to energy over the entire conduction band using (5.1) and (5.3), which yields

$$\begin{aligned} n_0 &= \int \mathrm{d}n = \int_{E_{\mathrm{c}}}^{\infty} f_{\mathrm{n}}(E)g_{\mathrm{n}}(E - E_{\mathrm{c}})\mathrm{d}E \\ &= \left(\frac{4\pi}{h^3}\right)(2m_{\mathrm{n}}^*)^{3/2}\int_{E_{\mathrm{c}}}^{\infty}\frac{(E - E_{\mathrm{c}})^{1/2}\mathrm{d}E}{[1 + \mathrm{e}^{(E-E_{\mathrm{f}})/k_{\mathrm{B}}T}]} \\ &= \left(\frac{4\pi}{h^3}\right)(2m_{\mathrm{n}}^* k_{\mathrm{B}}T)^{3/2}\int_0^{\infty}\frac{\varepsilon^{1/2}\mathrm{d}\varepsilon}{[1 + \mathrm{e}^{(\varepsilon-\eta)}]} \\ &= N_{\mathrm{c}}F_{1/2}(\eta), \end{aligned} \tag{5.5}$$

where

$$N_{\mathrm{c}} = 2(2\pi m_{\mathrm{n}}^* k_{\mathrm{B}}T/h^2)^{3/2} \tag{5.6}$$

is the effective density of the conduction band states,

$$F_{1/2}(\eta) = \left(\frac{2}{\sqrt{\pi}}\right)\int_0^{\infty}\frac{\varepsilon^{1/2}\mathrm{d}\varepsilon}{[1 + \mathrm{e}^{(\varepsilon-\eta)}]} \tag{5.7}$$

is the Fermi integral of order one-half, $\varepsilon = (E - E_{\mathrm{c}})/k_{\mathrm{B}}T$ is the reduced energy, m_{n}^* is the density-of-states effective mass of electrons, and $\eta = -(E_{\mathrm{c}} - E_{\mathrm{f}})/k_{\mathrm{B}}T$ is the reduced Fermi energy. Equation (5.5) is the general expression for the

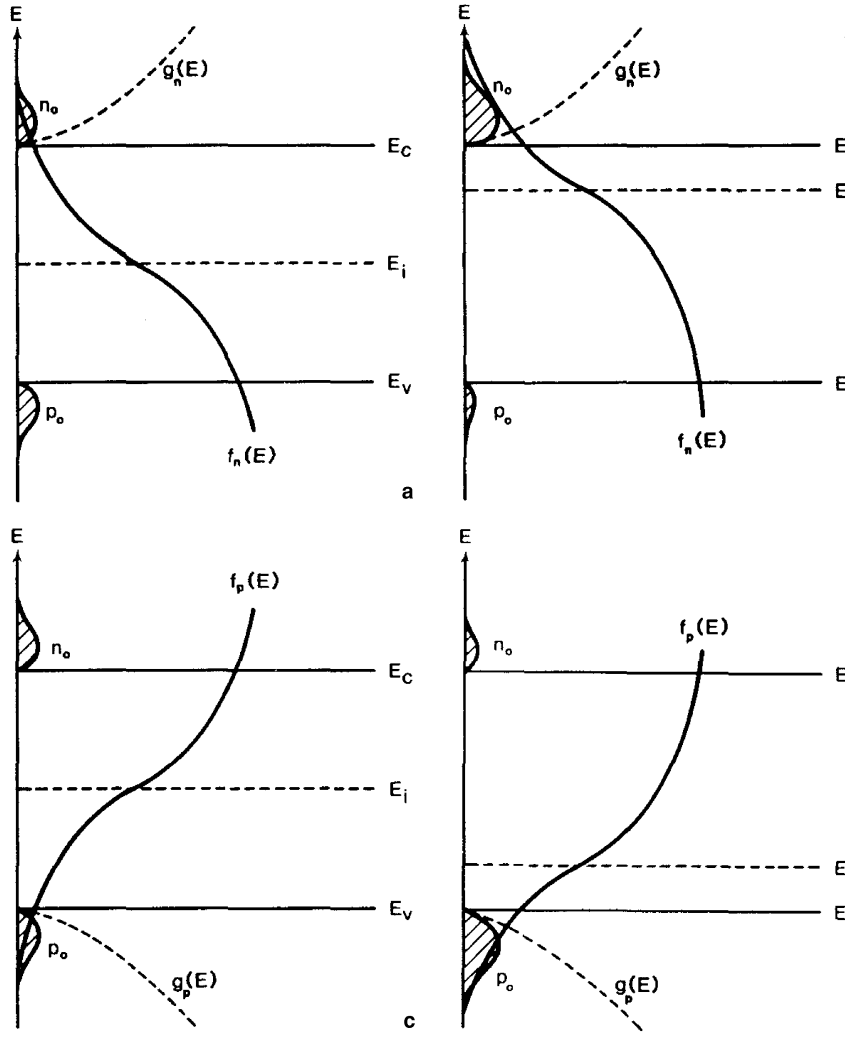


FIGURE 5.1. The Fermi–Dirac distribution function and the density-of-states function for electrons and holes in the conduction and valence bands of a semiconductor respectively, for $T > 0$ K.

equilibrium electron density in the conduction band applicable over the entire doping density range. Since the Fermi integral given by (5.7) can be evaluated only by numerical integration or by using a table of Fermi integrals, it is a common practice to use simplified expressions for calculating the carrier density over a certain range of doping densities in a degenerate semiconductor. The following approximations are valid for a specified range of reduced Fermi

energies:

$$F_{1/2}(\eta) \approx \begin{cases} e^{\eta} & \text{for } \eta < -4, \\ \dfrac{1}{(e^{-\eta} + 0.27)} & \text{for } -4 < \eta < 1, \\ \left(\dfrac{4}{3\sqrt{\pi}}\right)\left(\eta^2 + \dfrac{\pi^2}{6}\right)^{3/4} & \text{for } 1 < \eta < 4, \\ \left(\dfrac{4}{3\sqrt{\pi}}\right)\eta^{3/2} & \text{for } \eta > 4. \end{cases} \tag{5.8}$$

The expression of N_c given by (5.6) can be simplified to

$$N_c = 2.5 \times 10^{19}(T/300)^{3/2}(m_n^*/m_0)^{3/2}, \tag{5.9}$$

where $m_0 = 9.1 \times 10^{-31}$ kg is the free-electron mass, and m_n^* is the density-of-states effective mass for electrons in the conduction band. For $\eta \leq -4$, the Fermi integral of order one-half becomes an exponential function of η, which is identical to the M-B distribution function. In this case, the classical M-B statistics prevail, and the semiconductor is referred to as nondegenerate semiconductor. The density of electrons for the nondegenerate case can be simplified to

$$n_0 = N_c e^{-(E_c - E_f)/k_B T} = N_c e^{\eta}. \tag{5.10}$$

Equation (5.10) is valid for intrinsic or lightly doped semiconductors. For silicon, (5.10) is valid for doping densities less than 10^{19} cm^{-3}. However, for doping densities higher than N_c, (5.5) must be used instead. A simple rule of thumb for checking the validity of (5.10) is that n_0 should be three to four times smaller than N_c. Figure 5.2 shows the normalized electron density versus the reduced Fermi

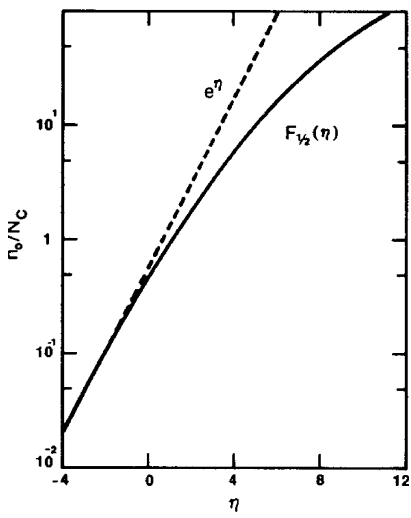


FIGURE 5.2. The normalized electron density, n_0/N_c, versus the reduced Fermi energy, η. The solid line is obtained from (5.5) and the dashed line is calculated from (5.10).

energy as calculated by (5.5) and (5.10). It is evident from this figure that the two curves calculated from the F-D and M-B distribution functions are nearly identical for $\eta \leq -4$ (i.e., the nondegenerate case), but they deviate considerably from each other for $\eta \geq 0$ (i.e., the degenerate case).

The hole density in the valence band can be derived in a similar way from (5.2) and (5.4), and the result is given by

$$p_0 = \left(\frac{4\pi}{h^3}\right)(2m_p^*)^{3/2}\int_{-\infty}^{E_v}\frac{(E_v - E)^{1/2}\mathrm{d}E}{[1 + \mathrm{e}^{(E_f - E)/k_BT}]}, = N_v F_{1/2}(-\eta - \varepsilon_g) \quad (5.11)$$

where $N_v = 2(2\pi m_p^* k_B T/h^2)^{3/2}$ is the effective density of the valence band states, m_p^* is the density-of-states effective mass for holes in the valence band, and $\varepsilon_g = (E_c - E_v)/k_B T$ is the reduced band gap. For the nondegenerate case with $(E_f - E_v) \geq 4k_B T$, (5.11) becomes

$$p_0 = N_v \mathrm{e}^{(E_v - E_f)/k_B T} = N_v \mathrm{e}^{-\eta - \varepsilon_g}, \quad (5.12)$$

which shows that the equilibrium hole density depends exponentially on the temperature and the reduced Fermi energy and energy band gap.

The results derived above are applicable to a single-valley semiconductor with a constant spherical energy surface near the bottom of the conduction band and the top of the valence band maximum, III-V compound semiconductors such as GaAs, InP, and InAs, which have a single constant spherical energy surface near the conduction band minimum (i.e., Γ-band), fall into this category. However, for elemental semiconductors such as silicon and germanium, which have multivalley conduction band minima, the scalar density-of-states effective mass used in (5.6) must be modified to account for the multivalley nature of the conduction band minima. This is discussed next.

The constant-energy surfaces near the conduction band minima for Si, Ge, and GaAs are shown in Figures 5.3a–c, respectively.[1] For silicon, there are six conduction band minima located along the $\{100\}$ axes, while there are eight conduction band minima located at the zone boundaries of the first Brillouin zone along the $\{111\}$ axes for germanium. Furthermore, the constant-energy surfaces near the bottom of the conduction bands are ellipsoidal for Si and Ge and spherical for GaAs. If one assumes that there are ν conduction band minima, then the total density of electrons in ν conduction band minima is given by

$$n_0' = \nu n_0 = \nu N_c F_{1/2}(\eta) = N_c' F_{1/2}(\eta), \quad (5.13)$$

where

$$N_c' = \nu N_c = 2\left(2\pi m_n^* \nu^{2/3} k_B T/h^2\right)^{3/2} = 2\left(2\pi m_{dn}^* k_B T/h^2\right)^{3/2} \quad (5.14)$$

is the effective density of the conduction band states for a multivalley semiconductor with an ellipsoidal constant-energy surface. The density-of-states effective mass of electrons, m_{dn}^*, in (5.14) can be expressed in terms of m_t and m_l by

$$m_{dn}^* = \nu^{2/3}\left(m_t^2 m_l\right)^{1/3}, \quad (5.15)$$

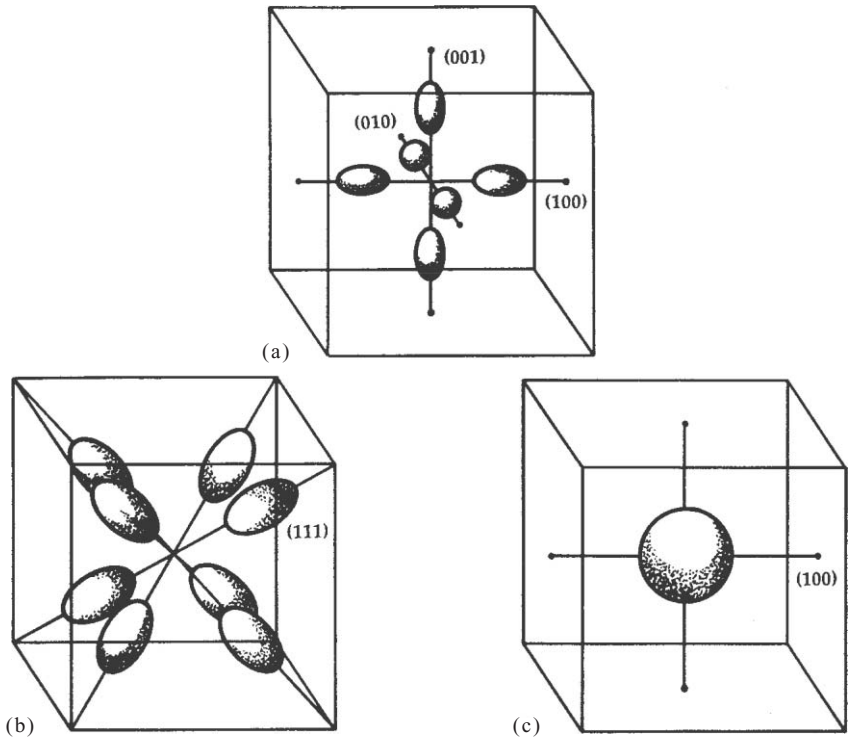


FIGURE 5.3. Constant-energy surfaces near the conduction band edges for (a) Si, (b) Ge, and (c) GaAs.

where m_t and m_l denote the transverse and longitudinal effective masses of electrons along the minor and major axes of the constant ellipsoidal energy surface, respectively. Values of m_t and m_l can be determined using the cyclotron resonant experiment performed at very low temperature. Here ν denotes the number of conduction band valleys in the semiconductor (e.g., $\nu = 6$ for Si and 4 for Ge). For silicon crystal with $\nu = 6$, $m_t = 0.19\, m_0$, and $m_l = 0.98 m_0$, m^*_{dn} was found to be 1.08 m_0. Table 5.1 lists the values of m_t, m_l, m^*_{dn}, and N'_v for Si, Ge, and GaAs at 300 K.

Calculations of hole densities in the valence bands for Si, Ge, and GaAs are different from those of the electron densities in the conduction band. This is because the valence band structures for these semiconductors are similar, consisting of a heavy-hole band and a light-hole band as well as the split-off band, as shown in Figure 5.4. For these semiconductors, the constant-energy surface near the top of the valence bands is nonparabolic and warped. For simplicity, it is assumed that the constant-energy surface near the top of the valence bands is parabolic, and by neglecting the split-off band contribution the hole density in the light- and heavy-hole bands can be expressed as

$$p_0 = p_H + p_L = N'_v F_{1/2}(-\eta - \varepsilon_g), \tag{5.16}$$

TABLE 5.1. Conduction and Valence Band Parameters for Silicon, Germanium, and GaAs.

Parameters	Ge	Si	GaAs
	Conduction band		
ν	4	6	1
m_x/m_0	0.082	0.19	—
m_l/m_0	1.64	0.98	—
m^*_{dn}/m_0	0.561	1.084	0.068
$N'_c(\text{cm}^{-3})$	1.03×10^{19}	2.86×10^{19}	4.7×10^{17}
	Valence band		
m_x/m_0	0.044	0.16	0.082
m_l/m_0	0.28	0.49	0.45
m^*_{dp}/m_0	0.29	0.55	0.47
$N'_c(\text{cm}^{-3})$	5.42×10^{18}	2.66×10^{19}	7.0×10^{18}

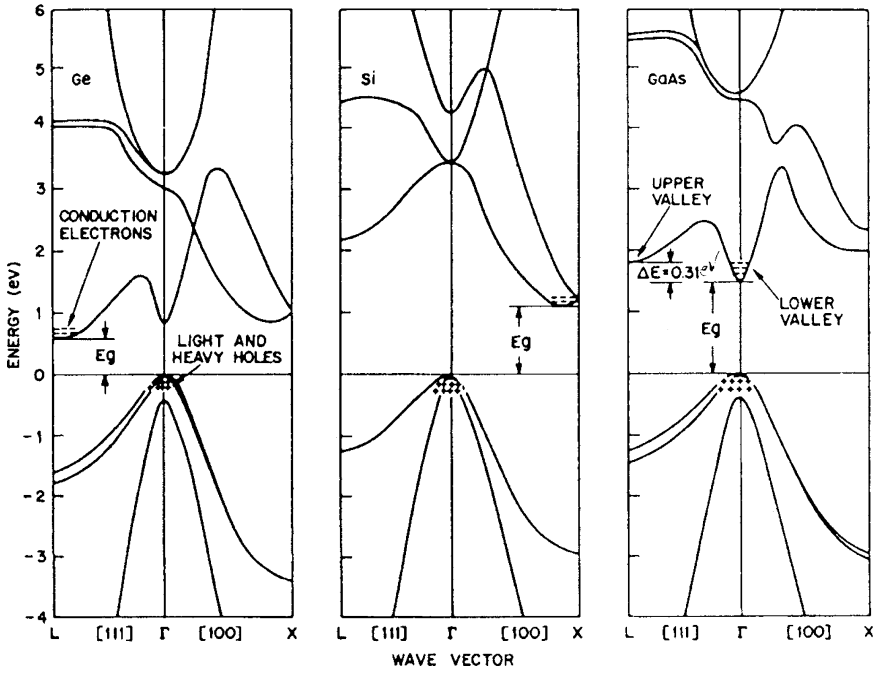


FIGURE 5.4. Energy band structures for Si, Ge, and GaAs along the (111) and (100) axes. For the valence bands, H represents the heavy-hole band and L denotes the light-hole band. Note that both bands are degenerate at $k = 0$. After Sze,[1] reprinted by permission from John Wiley & Sons Inc.

where p_{H} is the heavy-hole density, p_{L} is the light-hole density, $N_{\mathrm{v}}' = 2(2\pi m_{\mathrm{dp}}^* k_{\mathrm{B}} T/h^2)^{3/2}$ is the effective density of the valence band states, and

$$m_{\mathrm{dp}}^* = \left(m_{\mathrm{H}}^{3/2} + m_{\mathrm{L}}^{3/2}\right)^{2/3} \tag{5.17}$$

is the hole density-of-states effective mass; m_{H} and m_L denote the heavy- and light-hole masses, respectively. Values of $m_{\mathrm{H}}, m_L, m_{\mathrm{dp}}^*$, and N_{v}' for Ge, Si, and GaAs are also listed in Table 5.1.

5.3. Intrinsic Semiconductors

A semiconductor may be considered as an intrinsic semiconductor if its thermally generated carrier density (i.e., n_{i}) is much larger than the background doping or residual impurity densities. At $T = 0$ K, an intrinsic semiconductor behaves like an insulator because the conduction band states are totally empty and the valence band states are completely filled. However, as the temperature increases, some of the electrons in the valence band states are excited into the conduction band states by thermal energy, leaving behind an equal number of holes in the valence band. Thus, the intrinsic carrier density can be expressed by

$$n_{\mathrm{i}} = n_0 = p_0, \tag{5.18}$$

where n_0 and p_0 denote the equilibrium electron and hole densities, respectively. Substituting (5.5) and (5.11) into (5.18) yields the intrinsic carrier density

$$\begin{aligned} n_{\mathrm{i}} &= N_{\mathrm{c}} F_{1/2}(\eta) = N_{\mathrm{v}} F_{1/2}(-\eta - \varepsilon_{\mathrm{g}}) \\ &= (N_{\mathrm{c}} N_{\mathrm{v}})^{1/2} [F_{1/2}(\eta) F_{1/2}(-\eta - \varepsilon_{\mathrm{g}})]^{1/2}. \end{aligned} \tag{5.19}$$

In the nondegenerate case, (5.19) becomes

$$\begin{aligned} n_{\mathrm{i}} &= (N_{\mathrm{c}} N_{\mathrm{v}})^{1/2} \mathrm{e}^{-E_{\mathrm{g}}/2k_{\mathrm{B}}T} \\ &= 2.5 \times 10^{19} (T/300)^{3/2} \left(m_{\mathrm{dn}}^* m_{\mathrm{dp}}^* / m_0^2\right)^{3/4} \mathrm{e}^{-E_{\mathrm{g}}/2k_{\mathrm{B}}T}. \end{aligned} \tag{5.20}$$

A useful relationship between the square of the intrinsic carrier density and the product of electron and hole densities, valid for the nondegenerate case, is known as the law of mass action equation, which is given by

$$n_0 p_0 = n_{\mathrm{i}}^2 = N_{\mathrm{c}} N_{\mathrm{v}} \mathrm{e}^{-E_{\mathrm{g}}/k_{\mathrm{B}}T} = K_{\mathrm{i}}(T). \tag{5.21}$$

From (5.20) and (5.21) it is noted that the product $n_0 p_0$ depends only on the temperature, band gap energy, and the effective masses of electrons and holes. Equation (5.21) allows one to calculate the minority carrier density in an extrinsic semiconductor when the majority carrier density is known (e.g., $p_0 = n_{\mathrm{i}}^2/n_0$ for an extrinsic n-type semiconductor).

The intrinsic carrier density depends exponentially on both the temperature and band gap energy of the semiconductor. For example, at $T = 300$ K, values of the band gap energy for GaAs, Ge, and Si are given by 1.42, 0.67, and 1.12 eV,

and the corresponding intrinsic carrier densities are 2.25×10^6, 2.5×10^{13}, and 9.65×10^9 cm^{-3}, respectively. Thus, it is clear that an increase of 0.1 eV in band gap energy can result in a decrease of the intrinsic carrier density by nearly one order of magnitude. This result has a very important practical implication in semiconductor device applications, since the saturation current of a p-n junction diode or a bipolar junction transistor varies with the square of the intrinsic carrier density ($I_0 \propto n_i^2 \propto e^{-E_g/k_B T}$). Therefore, p-n junction devices fabricated from larger band gap semiconductors such as GaAs and GaN are expected to have much lower dark currents than those of smaller band gap semiconductors such as silicon and germanium, and hence are more suitable for high-temperature applications. The Fermi level of an intrinsic semiconductor may be obtained by solving (5.10) and (5.12) for the nondegenerate case, which yields

$$E_f = \frac{(E_c + E_v)}{2} + \left(\frac{k_B T}{2}\right) \ln(N_v/N_c) = E_i + \left(\frac{3k_B T}{4}\right) \ln(m^*_{dp}/m^*_{dn}), \tag{5.22}$$

where E_i is known as the intrinsic Fermi level, which is located in the middle of the forbidden gap at $T = 0$ K. As the temperature rises from $T = 0$ K, the Fermi level, E_f, will move toward the conduction band edge if $m^*_{dp} > m^*_{dn}$, and toward the valence band edge if $m^*_{dp} < m^*_{dn}$, as illustrated in Figure 5.5. The energy band gap of a semiconductor can be determined from the slope ($= -E_g/2k_B$) of the semilog plot of intrinsic carrier density (n_i) versus inverse temperature ($1/T$). The intrinsic carrier density may be determined using either the Hall effect measurements on a bulk semiconductor or the high-frequency capacitance–voltage measurements on a Schottky barrier or a p-n junction diode. The intrinsic carrier

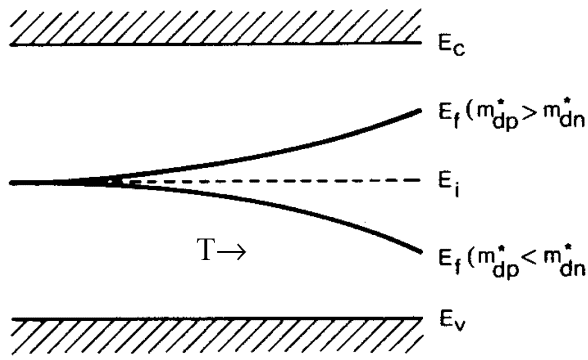


FIGURE 5.5. Fermi level vs. temperature for an intrinsic semiconductor.

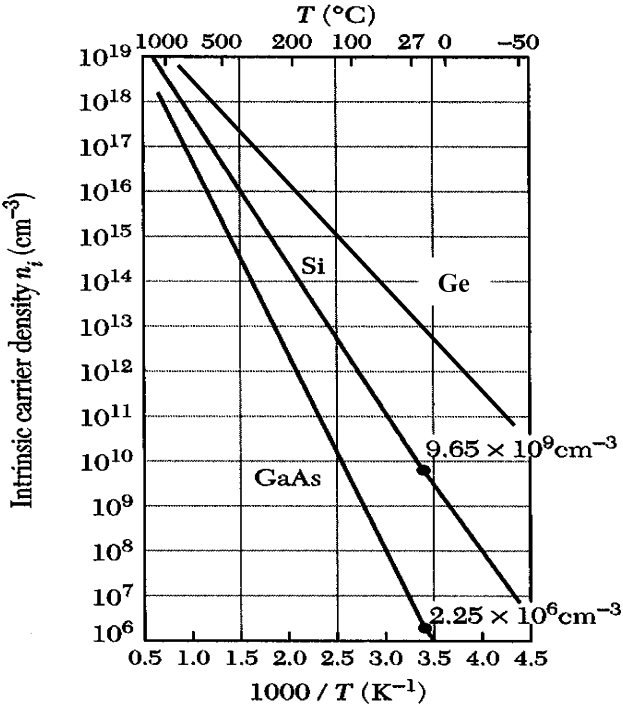


FIGURE 5.6. Intrinsic carrier density vs. inverse temperature for Si, Ge, and GaAs crystals.

densities for Ge, Si, and GaAs as a function of temperature are shown in Figure 5.6. The energy band gaps for these materials can be determined from the slope of the $\ln(n_i T^{-3/2})$ versus $1/T$ plot. The energy band gap determined from n_i is known as the thermal band gap of the semiconductor. On the other hand, the energy band gap of a semiconductor can also be determined using the optical absorption measurements near the absorption edge of the semiconductor. The energy band gap thus determined is usually referred to as the optical band gap of the semiconductor. A small difference between these two band gap values is expected due to the difference in the measurements of optical and thermal band gaps.

Since the energy band gap for most semiconductors decreases with increasing temperature, a correction of E_g with temperature is necessary when the intrinsic carrier density is calculated from (5.20). In general, the variation of energy band gap with temperature can be calculated using an empirical formula given by

$$E_g(T) = E_g(0) - \frac{\alpha T^2}{(T + \beta)}, \tag{5.23}$$

TABLE 5.2. Coefficients for the Temperature-Dependent Energy Band Gap of GaAs, InP, Si, and Ge.

Materials	$E_g(0)$ (eV)	$\alpha(10^{-4}$ eV/K)	β(K)
GaAs	1.519	5.41	204
InP	1.425	4.50	327
Si	1.170	4.73	636
Ge	0.744	4.77	235

where $E_g(0)$ is the energy band gap at $T = 0$ K; values of $E_g(0)$, α (eV/K), and β (K) for GaAs, InP, Si, and Ge are listed in Table 5.2.

5.4. Extrinsic Semiconductors

As discussed in Section 5.3, the electron–hole pairs in an intrinsic semiconductor are generated by thermal excitation. Therefore, for intrinsic semiconductors with an energy band gap on the order of 1 eV or higher, the intrinsic carrier density is usually very small at low temperatures (i.e., for $T < 100$ K). As a result, the resistivity for these intrinsic semiconductors is expected to be very high at low temperatures. This is indeed the case for Si, InP, GaAs, and other large band gap semiconductors. It should be noted that semi-insulating substrates with resistivity greater than $10^7\Omega$-cm can be readily obtained for undoped and Cr-doped GaAs as well as for Fe-doped InP materials. However, high-resistivity semi-insulating substrates are still unattainable for silicon and germanium due to the smaller band gap inherent in these materials, instead, SOI (silicon-on-insulator) wafers formed by oxygen-implantation (SIMOX) or wafer bonding (WB) techniques have been developed for producing the insulating substrates in silicon wafers. Novel devices and integrated circuits have been fabricated on SIMOX and WB silicon wafers for low-power, high-speed, and high-performance CMOS and BICMOS for a wide variety of ULSI applications.

The most important and unique feature of a semiconductor material lies in the fact that its electrical conductivity can be readily changed by many orders of magnitude by simply doping the semiconductor with shallow-donor or shallow-acceptor impurities. By incorporating the doping impurities into a semiconductor, the electron or hole density will increase with increasing shallow-donor or shallow-acceptor impurity concentrations. For example, electron or hole densities can increase from 10^{13} cm^{-3} to more than 10^{20} cm^{-3} if a shallow-donor or shallow-acceptor impurity with an equal amount of impurity densities were added into a silicon crystal. This is illustrated in Figure 5.7 for a silicon crystal.

Figure 5.7a shows an intrinsic silicon crystal with covalent bond structure. In this case, each silicon atom shares the four valence electrons reciprocally with its neighboring atoms to form covalent bonds. The covalent structure also applies

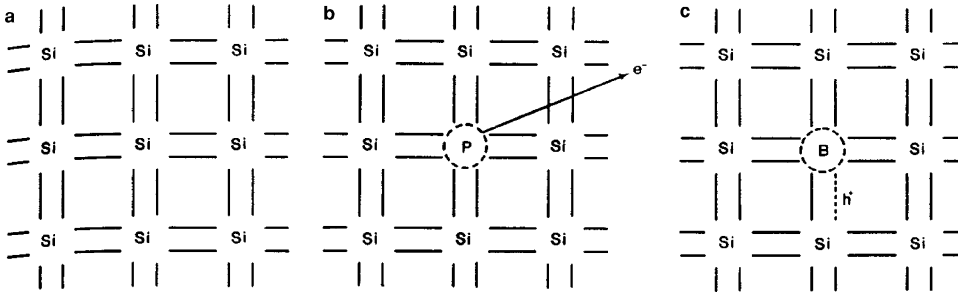


FIGURE 5.7. A covalent bond model for intrinsic and extrinsic silicon.

to other group-IV elements in the periodic table, such as germanium and diamond crystals. Figure 5.7b shows the substitution of a silicon atom by a group-V element such as phosphorus, arsenic, or antimony. In this case, an extra electron from the group-V atom is added to the host silicon lattice. Since this extra electron is loosely bound to the substitutional impurity atom (i.e., with ionization energy of a few tens of meV), it can be easily excited into the conduction band via thermal energy, and hence contributes to free electrons in the conduction band at 300 K. If the electrical conduction is due to electrons, then it is called an n-type semiconductor. A doping impurity that provides an extra electron per impurity atom to the host semiconductor is called a shallow-donor impurity. Thus, group-V elements in the periodic table are usually referred to as shallow-donor impurities for group-IV elemental semiconductors such as Si and Ge. If a group-III element is introduced into a group-IV elemental semiconductor, then there is a deficiency of one electron for each replaced host atom by a group-III impurity atom, leaving an empty state (or creation of a hole) in the valence band, as illustrated in Figure 5.7c. In this case, the conduction process is carried out by holes in the valence bands, and the semiconductor is called a p-type semiconductor. The group-III elements including boron, gallium, and aluminum are common doping impurities for producing p-type doping in the elemental semiconductors. Thus, group-III elements are shallow-acceptor impurities for the elemental semiconductors.

For III-V compound semiconductors (e.g., GaAs, GaP, GaSb, InP, InAs, InSb, etc.) and II-VI compounds (e.g., CdS, CdTe, ZnS, ZnSe, etc.), controlling n- or p-type doping is more complicated than for the elemental semiconductors. For example, n-GaAs can be obtained if the arsenic atoms in the arsenic sublattices are replaced by a group-VI element such as Te or Se, or if the gallium atoms are replaced by a group-IV element such as Ge, Si, or Sn. A p-type GaAs may be obtained if arsenic atoms are replaced by a group-IV element such as Ge or Si, or if gallium atoms are replaced by a group-II element such as Zn or Be. However, in practice, Te, Se, Sn, and Si are often used as n-type dopants, and Zn and Be are widely used as p-type dopants for GaAs and InP materials. As for II-VI compound semiconductors, it is even more difficult to produce n- or p-type

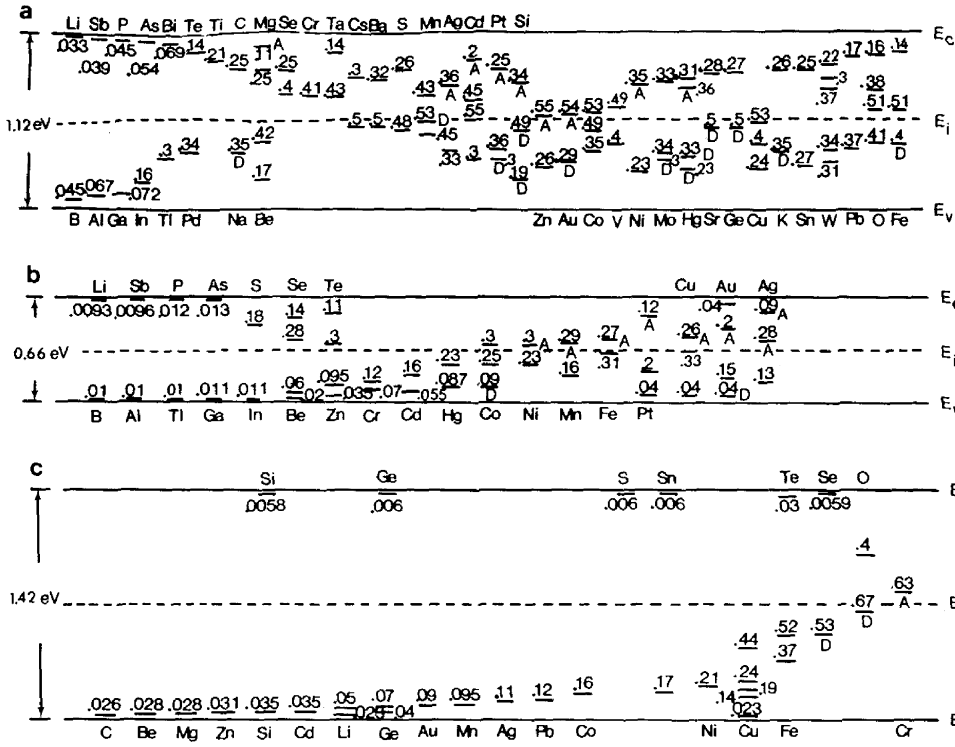


FIGURE 5.8. Ionization energies for various impurity levels in (a) Si, (b) Ge, and (c) GaAs. After Sze,[1] reprinted by permission from John Wiley & Sons Inc.

semiconductors by simply using the doping technique cited above due to the high density of native defects and the nonstoichiometric nature of the II-VI semiconductors. For example, while CdS and ZnSe are always exhibiting n-type conduction, ZnS and ZnTe are always showing p-type conduction. In recent years, nitrogen has been successfully used as p-type dopant to convert n-type ZnSe into p-type ZnSe, which has enabled the fabrication of blue and blue-green ZnSe p-n junction laser diodes and LEDs. The dopant impurities used in controlling the conductivity type of a semiconductor usually have very small ionization energies (i.e., a few tens of meV), and hence these impurities are often referred to as shallow-donor or shallow-acceptor impurities. These shallow-level impurities are usually fully ionized at room temperature for most semiconductors due to the small ionization energy.

Figure 5.8 shows the energy band diagrams and the impurity levels for (a) Si, (b) Ge, and (c) GaAs, respectively. The energy levels shown in the forbidden gap of Si and Ge include all the shallow-donor and acceptor impurities from group-III elements (e.g., B, Al, Ga) and group-V elements (e.g., P, As, Sb), and the deep-level impurity states from normal metals (Au, Cu, Ag) and transition metals

(Fe, Ni, Co). The shallow-level impurities are used mainly for controlling the carrier concentration and the conductivity of semiconductors, while the deep-level impurities are used to control the recombination and hence the minority carrier lifetimes in a semiconductor. As an example, gold is a deep-level impurity and an effective recombination center in n-type silicon; it has an acceptor level with ionization energy of $E^{-}_{\mathrm{Au}} = E_{\mathrm{c}} - 0.55\,\mathrm{eV}$ and a donor level with ionization energy of $E^{+}_{\mathrm{Au}} = E_{\mathrm{v}} + 0.35\,\mathrm{eV}$ in the forbidden gap of silicon. Since the gold acceptor level is the most effective mid-gap recombination center in silicon, gold impurity has been used for controlling the minority carrier lifetimes and hence the switching times in silicon devices.

The temperature behavior of equilibrium carrier density in an extrinsic semiconductor can be determined by solving the charge neutrality equation, using the expressions for the electron and hole densities in the conduction and valence band states as well as in the impurity states derived in Section 5.2. For an extrinsic semiconductor, if both the donor and acceptor shallow impurities are present in a host semiconductor, then the charge neutrality condition in thermal equilibrium is given by

$$\rho = 0 = q(p_0 - n_0 + N_{\mathrm{D}} - n_{\mathrm{D}} - N_{\mathrm{A}} + p_{\mathrm{A}}), \tag{5.24}$$

where p_0 and n_0 are the equilibrium hole and electron densities in the valence and conduction bands, while N_{D} and N_{A} are the donor- and acceptor-impurity densities, respectively. Note that n_{D} and p_{A} are the electron and hole densities in the shallow-donor and shallow-acceptor states, which are given, respectively, by

$$n_{\mathrm{D}} = \frac{N_{\mathrm{D}}}{[1 + g_{\mathrm{D}}^{-1} \mathrm{e}^{(E_{\mathrm{D}} - E_{\mathrm{f}})/k_{\mathrm{B}}T}]}, \tag{5.25}$$

$$p_{\mathrm{A}} = \frac{N_{\mathrm{A}}}{[1 + g_{\mathrm{A}} \mathrm{e}^{(E_{\mathrm{f}} - E_{\mathrm{A}})/k_{\mathrm{B}}T}]}, \tag{5.26}$$

where g_{D} and g_{A} denote the ground-state degeneracy factors for the shallow-donor and shallow-acceptor states.

In general, the temperature dependence of the carrier density and the Fermi level for an extrinsic semiconductor can be predicted using (5.24), (5.25), and (5.26). For an n-type nondegenerate semiconductor, assuming $N_{\mathrm{D}} \gg N_{\mathrm{A}}$ and $N_{\mathrm{A}} \gg p_{\mathrm{A}}$, a general expression for the charge neutrality equation can be obtained by substituting (5.10), (5.12), and (5.25) into (5.24), which yields

$$N_{\mathrm{c}} \mathrm{e}^{-(E_{\mathrm{c}} - E_{\mathrm{f}})/k_{\mathrm{B}}T} = N_{\mathrm{v}} \mathrm{e}^{(E_{\mathrm{f}} - E_{\mathrm{v}})/k_{\mathrm{B}}T} - N_{\mathrm{A}} + \frac{N_{\mathrm{D}}}{1 + g_{\mathrm{D}} \mathrm{e}^{(E_{\mathrm{f}} - E_{\mathrm{D}})/k_{\mathrm{B}}T}}. \tag{5.27}$$

Equation (5.27) is known as the charge neutrality equation for n-type extrinsic semiconductors. The Fermi level E_{f} can be determined by solving (5.27) using an iteration procedure. However, simple analytical solutions may be obtained in three different temperature regimes in which simplification can be made in (5.27). The three temperature regimes, which include the intrinsic, exhaustion, and deionization regimes, are now discussed.

(i) The intrinsic regime. At very high temperatures, when the thermally generated carrier densities in both the conduction and valence bands are much larger than the background doping densities (i.e., $n_i \gg (N_D - N_A)$), the semiconductor becomes an intrinsic semiconductor. In the intrinsic regime, (5.24) reduces to

$$n_0 = p_0 = n_i, \tag{5.28}$$

where

$$n_i = (N_v N_c)^{1/2} e^{-E_g/2k_B T} \tag{5.29}$$

is the intrinsic carrier density. In this regime, the intrinsic carrier density is much larger than the net doping impurity density of the semiconductor. For a silicon specimen with a doping density of 1×10^{16} cm^{-3}, the temperature corresponding to the onset of the intrinsic regime is at $T \geq 800$ K. For a germanium crystal with the same doping density, this occurs at $T \geq 600$ K. Figure 5.6 shows a plot of intrinsic carrier density versus temperature for Ge, Si, and GaAs. Note that the energy band gap can be determined from the slope of the $\ln(n_i T^{-3/2})$ versus $1/T$ plot using (5.29). It is noted that for the same doping density, the intrinsic regime for GaAs will occur at a much higher temperature than that of Si due to the larger energy band gap ($E_g = 1.42$ eV at 300 K) for GaAs. From Figure 5.6 the intrinsic carrier densities at 300 K for Si and GaAs are found to be $n_i = 9.65 \times 10^9$ cm^{-3} for Si, and 2.25×10^6 cm^{-3} for GaAs. The Fermi level as a function of temperature in the intrinsic regime is given by (5.22), which shows a linear dependence with temperature ($E_f = E_i$ at T = 0 K), as shown in Figure 5.5.

(ii) The exhaustion regime. In the exhaustion regime, the shallow-donor impurities in an n-type semiconductor are fully ionized at room temperature, and hence the electron density is equal to the net doping impurity density. Thus, the electron density can be expressed by

$$n_0 \approx (N_D - N_A) = N_c \, e^{-(E_c - E_f)/k_B T}. \tag{5.30}$$

From (5.30), the Fermi level E_f is given by

$$E_f = E_c - k_B T \ln[N_C/(N_D - N_A)]. \tag{5.31}$$

Equation (5.31) is valid only in the temperature regime in which all the shallow-donor impurities are ionized. As the temperature decreases, the Fermi level moves toward the donor level, and a fraction of the donor impurities become deionized (or neutral). This phenomenon is known as carrier freeze-out, which usually occurs in shallow-donor impurities at very low temperatures. This temperature regime is referred to as the deionization regime, which is discussed next.

(iii) The deionization regime. In the deionization regime, the thermal energy is usually too small to excite electrons from the shallow-donor impurity level into the conduction band, and hence a portion of the shallow-donor impurities are filled by electrons while some of the shallow-donor impurities will remain ionized. The kinetic equation, which governs the transition of electrons between

the shallow-donor level and the conduction band, is given by

$$n_0 + (N_D - n_D) \rightleftarrows n_D^0, \tag{5.32}$$

or

$$\frac{n_0(N_D - n_D)}{N_D^0} = K_D(T), \tag{5.33}$$

where N_D^0 denotes the neutral donor density (i.e., $n_D = N_D^0$), and $K_D(T)$ is a constant that depends only on temperature. The charge-neutrality condition in this temperature regime (assuming $p_A \ll N_A$ and $(E_f - E_D) \gg k_B T$) is given by

$$n_0 = p_0 + (N_D - n_D) - N_A \tag{5.34}$$

and

$$(N_D - n_D) = N_D g_D^{-1} \, e^{(E_D - E_f)/k_B T}. \tag{5.35}$$

Substituting (5.35) into (5.33) and assuming $N_D^0 = N_D$, one obtains

$$K_D(T) = g_D^{-1} \, N_c e^{-(E_c - E_D)/k_B T}. \tag{5.36}$$

Now solving (5.33) and (5.34) yields

$$K_D(T) = \frac{n_0(N_D - n_D)}{N_D^0} \approx \frac{n_0(n_0 + N_A)}{(N_D - N_A)}. \tag{5.37}$$

Equation (5.37) is obtained by assuming $n_0 \gg p_0$ and $(N_D - n_D) \gg n_0$, and may be used to determine the ionization energy of the shallow-donor level and the dopant compensation ratio in an extrinsic semiconductor. Two limiting cases, which can be derived from (5.37), are now discussed.

(a) The lightly compensated case ($N_D \gg N_A$ *and* $N_A \ll n_0$). In this case, (5.37) becomes

$$K_D(T) \approx \frac{n_0^2}{(N_D - N_A)}. \tag{5.38}$$

Now solving (5.36) and (5.38), one obtains

$$n_0 = \left[(N_D - N_A) N_c g_D^{-1}\right]^{1/2} e^{-(E_c - E_D)/2k_B T}, \tag{5.39}$$

where $g_D = 2$ is the degeneracy factor for the shallow-donor level. Equation (5.39) shows that the electron density increases exponentially with increasing temperature. Thus, from the slope of the $\ln(n_0 T^{-3/2})$ versus $1/T$ plot in the deionization regime, one can determine the ionization energy of the shallow-donor level. For the lightly compensated case, the activation energy deduced from the slope of this plot is equal to one-half of the ionization energy of the shallow-donor level (i.e., slope $= -(E_c - E_D)/2k_B$).

(b) The highly compensated case ($N_D > N_A \gg n_0$). In this case, (5.37) reduces to

$$K_D = n_0 N_A / (N_D - N_A). \tag{5.40}$$

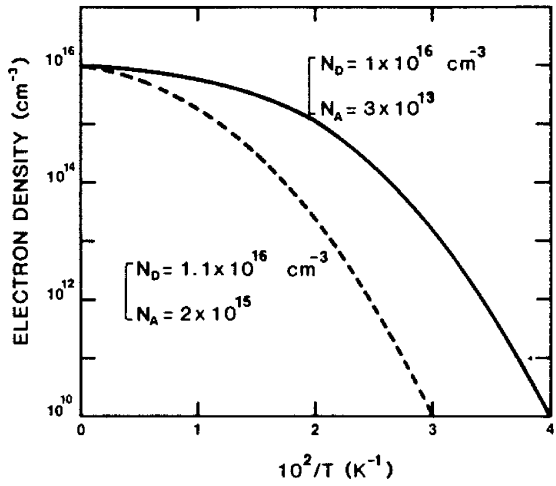


FIGURE 5.9. Electron density versus $10^2/T$ for two n-type silicon samples with different impurity compensations.

Solving (5.36) and (5.40) yields

$$n_0 = [(N_D - N_A)/N_A]N_c g_D^{-1} \, e^{-(E_c - E_D)/k_B T}. \tag{5.41}$$

From (5.41), it is noted that the activation energy determined from the slope of the $\ln(n_0 T^{-3/2})$ versus $1/T$ plot is equal to the ionization energy of the shallow-donor impurity level for the highly compensated case.

Figure 5.9 shows a plot of $\ln(n_0)$ versus $1/T$ for two n-type silicon samples with different doping densities and compensation ratios. The results clearly show that the sample with a higher impurity compensation ratio has a larger slope than the one with a smaller impurity compensation ratio. From the slope of this plot one can determine the activation energy of the shallow-donor impurity level. Therefore, by measuring the majority carrier density as a function of temperature over a wide range of temperature, one can determine simultaneously the values of N_D, N_A, E_g, E_D, or E_A for the extrinsic semiconductors. The above analysis is valid for an n-type extrinsic semiconductor with different impurity compensation ratios. A similar analysis can also be performed for a p-type extrinsic semiconductor.

The resistivity and Hall effect measurements are commonly employed to determine the carrier concentration, carrier mobility, energy band gap, and activation energy of shallow impurity levels as well as the compensation ratio of shallow impurities in a semiconductor. In addition to these measurements, the deep-level transient spectroscopy (DLTS) and photoluminescence (PL) methods are also widely used in characterizing both the deep-level defects and the shallow-level impurities in a semiconductor. Thus, by performing the resistivity and Hall effect measurements, detailed information concerning the equilibrium properties of a semiconductor can be obtained.

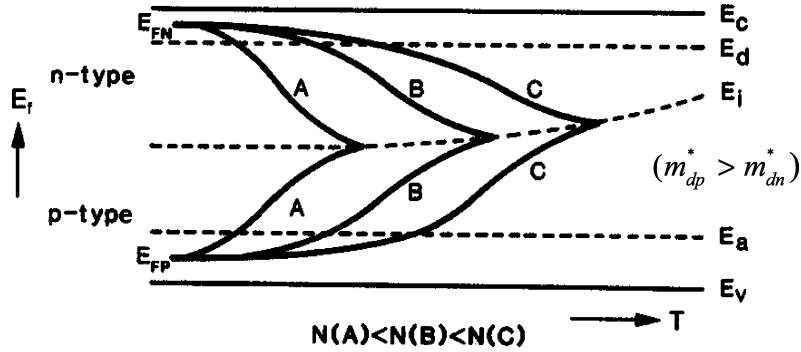


FIGURE 5.10. The Fermi level as a function of temperature for n- and p-type silicon with different compensation ratios (N_D/N_A).

Figure 5.10 shows the plot of Fermi level versus temperature for both n- and p-type silicon with different degrees of impurity compensation ratio. As can be seen in this figure, when the temperature increases, the Fermi level may move either toward the conduction band edge or toward the valence band edge, depending on the ratio of the electron effective mass to the hole effective mass. As shown in this figure, if the hole effective mass is greater than the electron effective mass, then the Fermi level will move toward the conduction band edge at high temperatures. On the other hand, if the electron effective mass is larger than the hole effective mass, then the Fermi level will move toward the valence band edge at high temperatures.

5.5. Ionization Energies of Shallow- and Deep-Level Impurities

Figure 5.8 shows the ionization energies of the shallow-level and deep-level impurities measured in Ge, Si, and GaAs materials. In general, the ionization energies for the shallow-donor impurity levels in these materials are less than 0.1 eV below the conduction band edge for the shallow-donor impurity levels and less than 0.1 eV above the valence band edge for the shallow-acceptor impurity levels. The ionization energy of a shallow-level impurity may be determined using the Hall effect, photoluminescence, or photoconductivity method. For silicon and germanium, the most commonly used shallow-donor impurities are phosphorus and arsenic, and the most commonly used shallow-acceptor impurity is boron. For GaAs and other III-V compounds, Si, Ge, Te, and Sn are the common shallow-donor impurities used as n-type dopants, while Zn and Be are commonly used as p-type dopants.

The shallow-impurity levels in a semiconductor may be treated within the framework of the effective mass model, which asserts that the electron is only loosely bound to a donor atom by a spherically symmetric Coulombic potential, and hence can be treated as a hydrogen-like impurity. Although the ionization energy of a

shallow impurity level can be calculated using the Schrödinger equation for the bound electron states associated with a shallow impurity atom, this procedure is a rather complicated one. Instead of solving the Schrödinger equation to obtain the ionization energy of a shallow impurity state, the simple Bohr model for the hydrogen atom described in Chapter 4 can be applied to calculate the ionization energy of the shallow-donor level in a semiconductor. Although Bohr's model may be oversimplified, it offers some physical insights concerning the nature of the shallow impurity states in a semiconductor. It is interesting to note that the ionization energy of a shallow impurity level calculated from the modified Bohr model agrees reasonably well with the experimental data for many semiconductors. In the hydrogen-like impurity model, the ionization energy of a shallow impurity state depends only on the effective mass of electrons and the dielectric constant of the semiconductor.

To calculate the ionization energy of a shallow impurity level, consider the case of a phosphorus donor atom in a silicon host crystal as shown in Figure 5.7b. Each silicon atom shares four valence electrons reciprocally with its nearest-neighbor atoms to form a covalent bond. The phosphorus atom, which replaces a silicon atom, has five valence electrons. Four of the five valence electrons in the phosphorus atom are shared by its four nearest-neighbor silicon atoms, while the fifth valence electron is loosely bound to the phosphorus atom. Although this extra electron of phosphorus ion is not totally free, it has a small ionization energy, which enables it to break loose relatively easily from the phosphorus atom and become free in silicon crystal. Therefore, one may regard the phosphorus atom as a fixed ion with a positive charge surrounded by an electron with a negative charge. If the ionizaton energy of this bound electron is small, then its orbit will be quite large (i.e., much larger than the interatomic spacing). Under this condition, it is reasonable to treat the bound electron as being embedded in a uniformly polarized medium whose dielectric constant is given by the macroscopic dielectric constant of the host semiconductor. This assumption resembles that of a hydrogen atom embedded in a uniform continuous medium with a dielectric constant equal to unity. Therefore, as long as the dielectric constant of the semiconductor is large enough such that the Bohr radius of the shallow-level impurity ground state is much larger than the interatomic spacing of the host semiconductor, the modified Bohr model can be used to treat the shallow impurity states in a semiconductor.

To apply the Bohr model to a phosphorus impurity atom in a silicon crystal, two parameters must be modified. First, the free electron mass m_0, which is used in a hydrogen atom, must be replaced by the electron effective mass m_{n}^*. Second, the relative permittivity in free space must be replaced by the dielectric constant of silicon, which is $\varepsilon_{\mathrm{s}} = 11.7$. Using (4.9) derived from the Bohr hydrogen model, the ground-state ionization energy for the shallow-donor impurity in silicon may be obtained by setting $n = 1$, and replacing $m_0 = m_{\mathrm{e}}^*$ and $\varepsilon_0 = \varepsilon_0\varepsilon_{\mathrm{s}}$ in (4.9), which yields

$$E_{\mathrm{i}} = \frac{-m_{\mathrm{e}}^* q^4}{32(\pi\varepsilon_0\varepsilon_{\mathrm{s}}h)^2} = -13.6(m_{\mathrm{e}}^*/m_0)\varepsilon_{\mathrm{s}}^{-2}\ \mathrm{eV}. \tag{5.42}$$

From (4.5), the Bohr radius for the ground state of the shallow impurity level is given by

$$r_1 = \frac{4\pi\varepsilon_0\varepsilon_s h^2}{m_e^* q^2} = 0.53\left(m_0/m_e^*\right)\varepsilon_s \text{ Å}. \tag{5.43}$$

Equation (5.42) shows that the ionization energy of a shallow impurity level is inversely proportional to the square of the dielectric constant. On the other hand, (5.43) shows that the Bohr radius varies linearly with the dielectric constant and inversely with the electron effective mass. For silicon, using the values of the electron effective mass $m_e^* = 0.26\ m_0$ and the dielectric constant $\varepsilon_s = 11.7$, the ionization energy calculated from (5.42) is found to be 25.8 meV, and the Bohr radius calculated from (5.43) is 24 Å. For germanium, with $m_e^* = 0.12\ m_0$ and $\varepsilon_s = 16$, the calculated value for E_i is found to be 6.4 meV, and the Bohr radius is equal to 71 Å. The above results clearly illustrate that the Bohr radii for the shallow impurity states in both silicon and germanium are indeed much larger than the interatomic spacing of silicon and germanium. The calculated ionization energies for the shallow impurity states in Si, Ge, and GaAs are generally smaller than the measured values shown in Figure 5.8. However, the agreement should improve for the excited states of these shallow impurity levels (i.e., for $n \geq 1$).

5.6. Hall Effect, Electrical Conductivity, and Hall Mobility

As discussed earlier, the majority carrier density (i.e., n_0 or p_0) and carrier mobility (μ_n or μ_p) are two key parameters that govern the transport and electrical properties of a semiconductor. Both parameters are usually determined using the Hall effect and resistivity measurements.

The Hall effect was discovered by Edwin H. Hall in 1879 during an investigation of the nature of the force acting on a conductor carrying a current in a magnetic field. Hall found that when a magnetic field is applied at right angles to the direction of current flow, an electric field is set up in a direction perpendicular to both the direction of the current and the magnetic field. To illustrate the Hall effect in a semiconductor, Figure 5.11 shows the Hall effect for a p-type semiconductor bar and the polarity of the induced Hall voltage.

As shown in Figure 5.11, the Hall effect is referred to the phenomenon in which a Hall voltage (V_H) is developed in the y-direction when an electric current (J_x) is applied in the x-direction and a magnetic field (B_z) is in the z-direction of a semiconductor bar. The interaction of a magnetic field in the z-direction with the electron motion in the x-direction produces a Lorentz force along the negative y-direction, which is counterbalanced by the Hall voltage developed in the y-direction. This can be written as

$$q\mathcal{E}_y = -qB_z v_x = -B_z J_x/n_0, \tag{5.44}$$

where B_z is the magnetic flux density in the z-direction, and n_0 is the electron density. The current density J_x due to the applied electric field $\mathcal{E}_x$ in the x-direction

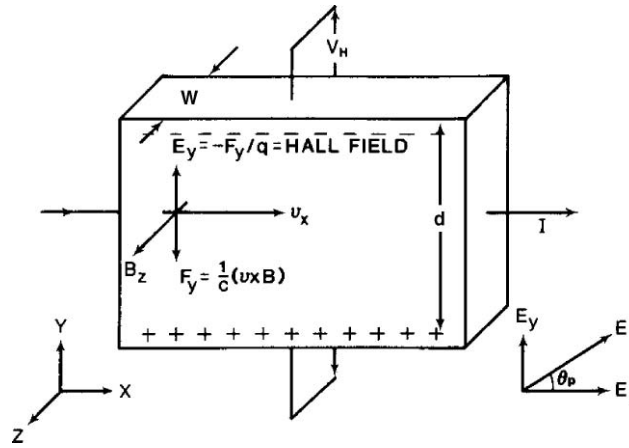


FIGURE 5.11. Hall effect for a p-type semiconductor bar. The Hall voltage V_{H}, the applied electric field $\mathcal{E}_x$, and the applied magnetic field B_z are mutually perpendicular. For an n-type sample, electrons are deflected in the y-direction to the bottom of the sample.

is given by

$$J_x = n_0 q \mu_{\mathrm{n}} \mathcal{E}_x. \tag{5.45}$$

For the small magnetic field case (i.e., $\mu_{\mathrm{n}} B_z \ll 1$), the angle between the current density J_x and the induced Hall field $\mathcal{E}_y$ is given by

$$\tan\theta_{\mathrm{n}} \approx \theta_{\mathrm{n}} = \frac{\mathcal{E}_y}{\mathcal{E}_x} = -B_z \mu_{\mathrm{n}}, \tag{5.46}$$

where θ_{n} is the Hall angle for electrons. The Hall coefficient R_{H} is defined by

$$R_{\mathrm{H}} = \frac{\mathcal{E}_y}{B_z J_x}\Big|_{J_y=0} = \frac{V_{\mathrm{H}} W}{B_z I_x}; \tag{5.47}$$

V_{H} is the Hall voltage, and W is the width of the semiconductor bar. Solving (5.44) through (5.47) yields

$$R_{\mathrm{Hn}} = -\frac{1}{n_0 q}. \tag{5.48}$$

Equation (5.48) shows that the Hall coefficient is inversely proportional to the electron density, and the minus sign in (5.48) is for n-type semiconductors in which the electron conduction prevails. Thus, from the measured Hall coefficient, one can calculate the electron density in an n-type semiconductor. It should be noted that (5.48) does not consider the scattering of electrons by different scattering sources such as ionized impurities, acoustical phonons, or neutral impurities. A relaxation time τ should be introduced when one considers the scattering mechanisms. Thus,

(5.48) is valid as long as the relaxation time constant τ is independent of electron energy. If τ is a function of electron energy, then a generalized expression for the Hall coefficient must be used, namely,

$$R_{\mathrm{Hn}} = -\frac{\gamma_{\mathrm{n}}}{qn_0}, \tag{5.49}$$

where $\gamma_{\mathrm{n}} = \langle\tau\rangle^2/\langle\tau^2\rangle$ is the Hall factor, and $\langle\tau\rangle$ is the average relaxation time. Values of γ_{n} may vary between 1.18 and 1.93, depending on the types of scattering mechanisms involved. Derivation of the Hall factor due to different scattering mechanisms will be discussed further in Chapter 7.
Similarly, the Hall coefficient for a p-type semiconductor can be expressed by

$$R_{\mathrm{Hp}} = \frac{\gamma_{\mathrm{p}}}{qp_0}, \tag{5.50}$$

where p_0 is the hole density and $\gamma_{\mathrm{p}} = \langle\tau\rangle^2/\langle\tau^2\rangle$ is the Hall factor for holes in a p-type semiconductor. Equation (5.50) shows that the Hall coefficient for a p-type semiconductor is positive since a hole has a positive charge. Values of the Hall factor for a p-type semiconductor may vary between 0.8 and 1.9, depending on the types of scattering mechanisms involved. This will also be discussed further in Chapter 7.

For an intrinsic semiconductor, both electrons and holes are expected to participate in the conduction process, and hence the mixed conduction prevails. Thus, the Hall coefficient for a semiconductor in which both electrons and holes are contributing to the conduction can be expressed by

$$R_{\mathrm{H}} = \frac{\mathcal{E}_y}{B_z J_x} = \frac{R_{\mathrm{Hn}}\sigma_{\mathrm{n}}^2 + R_{\mathrm{Hp}}\sigma_{\mathrm{p}}^2}{(\sigma_{\mathrm{n}} + \sigma_{\mathrm{p}})^2} = \frac{\left(p_0\mu_{\mathrm{p}}^2 - n_0\mu_{\mathrm{n}}^2\right)}{q(p_0\mu_{\mathrm{p}} + n_0\mu_{\mathrm{n}})^2}, \tag{5.51}$$

where R_{Hn} and R_{Hp} denote the Hall coefficients for n- and p-type conduction given by (5.49) and (5.50), respectively; and σ_{n} and σ_{p} are the electrical conductivities for the n- and p-type semiconductors, respectively. It is interesting to note from (5.51) that the Hall coefficient vanishes (i.e., $R_{\mathrm{H}} = 0$) if $p_0\mu_{\mathrm{p}}^2 = n_0\mu_{\mathrm{n}}^2$. This situation may in fact occur in an intrinsic semiconductor as one measures the Hall coefficient as a function of temperature over a wide range of temperature in which the conduction in the material may change from n- to p-type conduction at an elevated temperature.

From the above analysis, it is clear that the Hall effect and resistivity measurements are important experimental tools for analyzing the equilibrium properties of semiconductors. It allows one to determine the key physical and material parameters such as majority carrier density, conductivity mobility, ionization energy of the shallow impurity level, conduction type, energy band gap, and the impurity compensation ratio in a semiconductor.

Electrical conductivity is another important physical parameter, which is discussed next. The performance of a semiconductor device is closely related to the

electrical conductivity of a semiconductor. The electrical conductivity for an extrinsic semiconductor is equal to the product of electronic charge, carrier density, and carrier mobility, and can be expressed by

$$\sigma_{\mathrm{n}} = qn_0\mu_{\mathrm{n}} \quad \text{for n-type}, \tag{5.52}$$

$$\sigma_{\mathrm{p}} = qp_0\mu_{\mathrm{p}} \quad \text{for p-type}. \tag{5.53}$$

For an intrinsic semiconductor, the electrical conductivity is given by

$$\sigma_{\mathrm{i}} = \sigma_{\mathrm{n}} + \sigma_{\mathrm{p}} = q(\mu_{\mathrm{n}}n_0 + \mu_{\mathrm{p}}P_0) = q(\mu_{\mathrm{n}} + \mu_{\mathrm{p}})n_{\mathrm{i}}, \tag{5.54}$$

where μ_{n} and μ_{p} denote the electron and hole mobilities, respectively, and n_{i} is the intrinsic carrier density.

Since both the electrical conductivity and Hall coefficient are measurable quantities, the product of these two parameters, known as the Hall mobility, can also be obtained experimentally. Using (5.49), (5.50), (5.52), and (5.53) the Hall mobilities for an n-type and a p-type semiconductor are given, respectively, by

$$\mu_{\mathrm{Hn}} = R_{\mathrm{Hn}}\sigma_{\mathrm{n}} = \gamma_{\mathrm{n}}\mu_{\mathrm{n}} \quad \text{for n-type}, \tag{5.55}$$

$$\mu_{\mathrm{Hp}} = R_{\mathrm{Hp}}\sigma_{\mathrm{p}} = \gamma_{\mathrm{p}}\mu_{\mathrm{p}} \quad \text{for p-type}, \tag{5.56}$$

where γ_{n} and γ_{p} denote the Hall factor for n- and p-type semiconductors, respectively. The ratio of Hall mobility to conductivity mobility is equal to the Hall factor, which depends only on the scattering mechanisms.

5.7. Heavy Doping Effects in a Degenerate Semiconductor

As discussed earlier, the electrical conductivity of a semiconductor may be changed by many orders of magnitude by simply doping the semiconductor with shallow-donor or shallow-acceptor impurities. However, when the doping density is greater than 10^{19} cm^{-3} for the cases of silicon and germanium, the materials become degenerate, and hence change in the fundamental physical properties of the semiconductor result. The heavy doping effects in a degenerate semiconductor include the broadening of the shallow impurity level in the forbidden gap from a discrete level into an impurity band, the shrinkage of the energy band gap, the formation of a band tail at the conduction and valence band edges, and distortion of the density-of-states function from its square-root dependence on the energy. All these phenomena are referred to as heavy doping effects in a degenerate semiconductor. In a heavily doped semiconductor, the Fermi–Dirac (F-D) statistics rather than the Maxwell–Boltzmann (M-B) statistics must be employed in calculating the carrier density and other transport coefficients in such a material.

There are two key heavy-doping effects in a degenerate semiconductor that must be considered. The first consideration is that the F-D statistics must be employed to calculate the carrier density in a degenerate semiconductor. The second

heavy-doping effect is related to the band gap narrowing effect. It is noted that the heavy-doping effect is a very complicated physical problem, and existing theories for dealing with the heavy-doping effects are inadequate. Due to the existence of the heavy-doping regime in various silicon devices and ICs, most of the theoretical and experimental studies on heavy-doping effects reported in recent years have been focused on degenerate silicon material. The results of these studies on the band gap narrowing effect and carrier degeneracy for heavily doped silicon are discussed next.

Measurements of band gap narrowing as a function of doping density in heavily doped silicon samples have been widely reported in the literature. A semiempirical formula, based on the stored electrostatic energy of majority–minority carrier pairs, has been derived for the band gap reduction. The band gap narrowing, ΔE_g, for an n-type silicon is given by[2]

$$\Delta E_g = \left(\frac{3q^2}{16\pi\varepsilon_0\varepsilon_s}\right)\left(\frac{q^2 N_D}{\varepsilon_s\varepsilon_0 k_B T}\right)^{1/2}. \tag{5.57}$$

At room temperature, the band gap narrowing versus donor density for n-type silicon given by (5.57) becomes

$$\Delta E_g = 22.5(N_D/10^{18})^{1/2}\ \text{meV}, \tag{5.58}$$

where N_D is the donor density. Using (5.58), a band gap reduction of 225 meV is obtained at a doping density of 10^{20} cm^{-3} for n-type silicon. This value appears to be larger than the measured value reported for silicon. Figure 5.12 shows a plot of band gap narrowing versus donor density for silicon at 300 K, and a comparison of the calculated values of ΔE_g from (5.58) with experimental data.

Another important physical parameter to be considered here is the $n_0 p_0$ product, which is equal to the square of the effective intrinsic carrier density, n_{ie}^2. The $n_0 p_0$ product is an important parameter in a heavily doped p$^+$-n or an n$^+$-p junction diode and in the emitter region of a p$^+$-n-p or an n$^+$-p-n bipolar junction transistor. It relates the band gap narrowing effect to the saturation current density in the heavily doped emitter region of a bipolar junction transistor. To explain this, consider the square of the intrinsic carrier density in a heavily doped n-type semiconductor, which is given by

$$n_{ie}^2 = n_0 p_0 = N_c N_v e^{-E_g'/k_B T} F_{1/2}(\eta) e^{-\eta} = n_i^2 e^{\Delta E_g/k_B T} F_{1/2}(\eta) e^{-\eta}, \tag{5.59}$$

where $E_g' = E_g - \Delta E_g$ is the effective band gap of a heavily doped n-type semiconductor. Equation (5.59) is obtained using (5.5) for n_0 and (5.12) for p_0. Thus, when the band gap narrowing effect is considered, the $n_0 p_0$ product (or n_{ie}^2) is found to be much larger for a degenerate semiconductor than for a nondegenerate semiconductor. The increase of n_{ie}^2 in the heavily doped emitter region of a bipolar junction transistor (BJT) will lead to higher dark current, higher Auger recombination rate, and hence shorter carrier lifetime and lower current gain in a BJT.

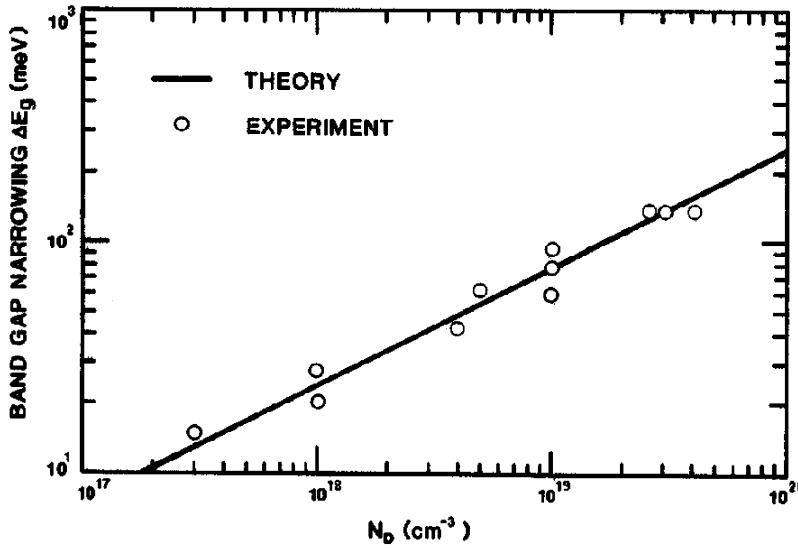


FIGURE 5.12. Calculated and measured values of bandgap narrowing versus donor density in n-type silicon material.[2]

In this chapter the physical and electrical properties of semiconductors under equilibrium conditions have been described. Key physical parameters, such as electron and hole densities in the conduction and valence bands, the ionization energies of the shallow- and deep-level impurities, and the band gap narrowing effect, have been derived and discussed. The importance of these physical parameters on the transport properties of semiconductors and device performance will be discussed further in Chapter 7.

Problems

5.1. Consider an n-type silicon doped with phosphorus impurities. The resistivity of this sample is 10 Ω cm at 300 K. Assuming that the electron mobility is equal to 1350 $cm^2/(Vs)$ and the density-of-states effective mass for electrons is $m^*_{dn} = 1.065\ m_0$:
   (a) Calculate the density of phosphorus impurities, assuming full ionization of phosphorus ions at 300 K.
   (b) Determine the location of the Fermi level relative to the conduction band edge, assuming that $N_A = 0$ and $T = 300$ K.
   (c) Determine the location of the Fermi level if $N_A = 0.5N_D$ and $T = 300$ K.
   (d) Find the electron density, n_0, at $T = 20$ K, assuming that $N_A = 0$, $g_D = 2$, and $E_c - E_D = 0.044$ eV.
   (e) Repeat (d) for $T = 77$ K.

5.2. If the temperature dependence of the energy band gap for InAs material is given by

$$E_g = 0.426 - 3.16 \times 10^{-4} T^2 (93 + T)^{-1} \text{ eV}$$

and the density-of-states effective masses for electrons and holes are given, respectively, by $m_n^* = 0.002\, m_0$ and $m_p^* = 0.4\, m_0$, plot the intrinsic carrier density, n_i, as a function of temperature for $200 < T < 700$ K. Assume that effective masses of electrons and holes do not change with temperature.

5.3. If the dielectric constant for GaAs is equal to 12, and the electron effective mass is $m_n^* = 0.086\, m_0$, calculate the ionization energy and the radius of the first Bohr orbit using Bohr's model given in the text. Repeat for an InP material.

5.4. Plot the Fermi level as a function of temperature for a silicon specimen with $N_D = 1 \times 10^{16}$ cm^{-3} and compensation ratios of $N_D/N_A = 0.1, 0.5, 2, 10$.

5.5. Show that the expressions given by (5.10) and (5.12) for electron and hole densities can be written in terms of the intrinsic carrier density, n_i, as follows:

$$n_0 = n_i e^{(E_f - E_i)/k_B T} \quad \text{and} \quad p_0 = n_i e^{(E_i - E_f)/k_B T},$$

where E_f is the Fermi level, and E_i is the intrinsic Fermi level.

5.6. Consider a semiconductor specimen. If it contains a small density of shallow-donor impurity such that $k_B T \ll (E_c - E_D) \ll (E_D - E_v)$, show that at $T = 0$ K the Fermi level is located halfway between E_c and E_D, assuming that E_D is completely filled at $T = 0$ K.

5.7. Derive an expression for the Hall coefficient of an intrinsic semiconductor in which conduction is due to both electrons and holes. Find the condition under which the Hall coefficient vanishes.

5.8. (a) When a current of 1 mA and a magnetic field intensity of 10^3 gauss are applied to an n-type semiconductor bar 1 cm wide and 1 mm thick, a Hall voltage of 1 mV is developed across the sample. Calculate the Hall coefficient and the electron density in this sample.

(b) If the electrical conductivity of this sample is equal to $2.5\, \Omega^{-1}\text{cm}^{-1}$, what is the Hall mobility? If the Hall factor is equal to 1.18, what is the conductivity mobility of an electron?

5.9. (a) Using the charge neutrality equation given by (5.24), derive a general expression for the hole density versus temperature, and discuss both the lightly compensated and highly compensated cases at low temperatures (i.e., in the deionization regime). Assume that the degeneracy factor for the acceptor level is $g_A = 4$. Plot p_0 versus T for the case $N_A = 5 \times 10^{16}$ cm^{-3}, $N_D = 10^{14}$ cm^{-3}, and $E_A = 0.044$ eV.

(b) Repeat for the case $N_D = 0.5 N_A$; $N_A = 10^{16}$ cm^{-3}.

(c) Plot the Fermi level versus temperature.

5.10. Using Fermi statistics and taking into account the band gap narrowing effects, show that the product $n_0 p_0$ is given by (5.66) for a heavily doped n-type

semiconductor:

$$n_{ie}^2 = n_0 p_0 = n_i^2 \exp(\Delta E_g / kT) F_{1/2}(\eta) \exp(-\eta),$$

where n_i is the intrinsic carrier concentration for the nondegenerate case, ΔE_g is the band gap narrowing, $F_{1/2}(\eta)$ is the Fermi integral of order one-half, and $\eta = -(E_c - E_f)/k_B T$ is the reduced Fermi energy. Using the above equation, calculate the values of n_{ie}^2 and ΔE_g for an n-type degenerate silicon for $\eta = 1$ and 4 (assuming fully ionization of the donor ions) and $T = 300$ K. Given:

$$\Delta E_g = 22.5(N_D/10^{18})^{12} \text{ meV},$$
$$F_{1/2}(\eta) = (4/3\sqrt{x})(\eta^2 + \pi^2/6)^{3/4},$$
$$N_c = 2.75 \times 10^{19} \text{ cm}^{-3},$$
$$N_v = 1.28 \times 10^{19} \text{ cm}^{-3}.$$

5.11. Using (5.66), plot n_{ie}^2 versus N_D for n-type silicon with N_D varying from 10^{17} to 10^{20} cm^{-3}.

5.12. Plot the ratios $n_\Gamma/n_{0,}/n_L/n_0, /n_x/n_0$ versus temperature (T) for a GaAs crystal for $0 < T < 1000$ K. The electron effective masses for the Γ-, X-, and L-conduction band minima are given, respectively, by Γ-band, $m_\Gamma = 0.0632\ m_0$; L-band, $m_l \approx 1.9\ m_0; m_t \approx 0.075\ m_0; m_L = (16\ m_l m_t^2)^{1/3} = 0.56\ m_0$; X-band: $m_l \approx 1.9\ m_0; m_t \approx 0.19\ m_0; m_X = (9\ m_l m_t^2)^{1/3} = 0.85\ m_0; n_0 = n_\Gamma + n_X + n_L; m_L$ and m_X are the density-of-states effective masses for the L- and X-bands, and

$$n_\Gamma = N_c^\Gamma \exp[(E_F - E_c)/k_B T] = 2(2\pi m_\Gamma k_B T/h^2)^{3/2} \exp[\eta/k_B T],$$
$$n_x = 2(2\pi m_x k_B T/h^2)^{3/2} \exp[(\eta - \Delta_x)/k_B T],$$
$$n_L = 2(2\pi m_L k_B T/h^2)^{3/2} \exp[(\eta - \Delta_{\Gamma L})/k_B T],$$

where
$\eta = E_F - E_c \Delta_{\Gamma X} = E_X - E_\Gamma = 0.50$ eV, and $\Delta_{\Gamma L} = E_L - E_\Gamma = 0.33$ eV.

5.13. (a) If the energy band gaps for InP, InAs, and InSb are given by $E_g = 1.34$ (InP), 0.36 (InAs), and 0.17 eV (InSb), respectively, at 300 K, calculate the intrinsic carrier densities in these materials at 300, 400, and 500 K. (b) Plot ln n_i versus 1/T for these three materials for $200 < T < 600$ K. (Given: $n_i = 1.2 \times 10^8$, 1.3×10^{15}, and 2.0×10^{16} cm^{-3}, for InP, InAs, and InSb at 300 K, respectively.)

5.14. Using (5.23) and Table 5.2 plot the energy band gap versus temperature for Si, GaAs, Ge, and InP for $200 \leq T \leq 600$ K.

References

1. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1981).
2. H. P. D. Lanyon and R. A. Tuft, *Band Gap Narrowing in Heavily-Doped Silicon*, International Electron Device Meeting, IEEE Tech. Dig. (1978) p. 316.

3. S. M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd ed., Wiley, New York (2002).
4. R. F. Pierret, *Advanced Semiconductor Fundamentals,* 2nd ed., Prentice Hall, New Jersey (2003).
5. D. A. Neamen, *Semiconductor Physics and Devices*, 3rd ed., McGraw Hill, New York (2003).

Bibliography

J. S. Blakemore, *Semiconductor Statistics*, Pergamon Press, New York (1962).
F. J. Blatt, *Physics of Electronic Conduction in Solids*, McGraw-Hill, New York (1968).
R. H. Bube, *Electronic Properties of Crystalline Solids*, Academic Press, New York (1973).
V. I. Fistul', *Heavily Doped Semiconductors*, Plenum Press, New York (1969).
A. F. Gibson and R. E. Burgess, *Progress in Semiconductors*, Wiley, New York (1964).
N. B. Hannay, *Semiconductors*, Reinhold, New York (1959).
D. C. Look, *Electrical Characterization of GaAs Materials and Devices*, Wiley, New York (1989).
J. P. McKelvey, *Solid-State and Semiconductor Physics*, 2nd ed., Harper & Row, New York (1966).
A. G. Milnes, *Deep Impurities in Semiconductors*, Wiley, New York (1973).
K. Seeger, *Semiconductor Physics*, 3rd ed., Springer-Verlag, New York (1973).
W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, New York (1950).
M. Shur, *Physics of Semiconductor Devices*, Prentice-Hall, New York (1990).
R. A. Smith, *Semiconductors*, 2nd ed., Cambridge University Press, Cambridge (1956).
H. F. Wolf, *Semiconductors*, Wiley, New York (1971).

6
Excess Carrier Phenomenon in Semiconductors

6.1. Introduction

The generation of excess carriers in a semiconductor may be accomplished by either electrical or optical means. For example, electron–hole pairs are created in a semiconductor when photons with energies exceeding the band gap energy of the semiconductor are absorbed. Similarly, minority carrier injection can be achieved by applying a forward bias voltage across a p-n junction diode or a bipolar junction transistor. The inverse process to the generation of excess carriers in a semiconductor is that of recombination. The annihilation of excess carriers generated by optical or electrical means in a semiconductor may take place via different recombination mechanisms. Depending on the ways in which the energy of an excess carrier is removed during a recombination process, there are three basic recombination mechanisms that are responsible for carrier annihilation in a semiconductor. They are (1) nonradiative recombination (i.e., the multiphonon process), (2) band-to-band radiative recombination, and (3) Auger band-to-band recombination. The first recombination mechanism, known as the nonradiative or multiphonon recombination process, is usually the predominant recombination process for indirect band gap semiconductors such as silicon and germanium. In this process, recombination is accomplished via a deep-level recombination center in the forbidden gap, and the energy of the excess carriers is released via phonon emission. The second recombination mechanism, band-to-band radiative recombination, is usually the predominant process occurring in direct band gap semiconductors such as GaAs and InP. In this case, the band-to-band recombination of electron-hole pairs is accompanied by the emission of a photon. Auger band-to-band recombination is usually the predominant recombination process occurring in degenerate semiconductors and small-band-gap semiconductors such as InSb and HgCdTe materials. The Auger recombination process can also become the predominant recombination mechanism under high-injection conditions. Unlike the nonradiative and radiative recombination processes, which are two-particle processes, Auger band-to-band recombination is a three-particle process, which involves two electrons and one hole for n-type semiconductors, or one electron and two holes for p-type semiconductors. For an n-type semiconductor, Auger

recombination is accomplished first via electron–electron collisions in the conduction band, and followed by electron–hole recombination in the valence band. On the basis of the principle of detailed balance, the rate of recombination is equal to the rate of generation of excess carriers under thermal equilibrium conditions, and hence a charge-neutrality condition prevails throughout the semiconductor specimen.

Equations governing the recombination lifetimes for the three basic recombination mechanisms described above are derived in Sections 6.2, 6.3, and 6.4, respectively. The continuity equations for the excess carrier transport in a semiconductor are presented in Section 6.5, and the charge-neutrality equation is discussed in Section 6.6. The Haynes–Shockley experiment and the drift mobility for minority carriers are presented in Secton 6.7. Section 6.8 presents methods of determining the minority carrier lifetimes in a semiconductor. The surface states and surface recombination mechanisms in a semiconductor are discussed in Section 6.9. Finally, the deep-level transient spectroscopy (DLTS) technique for characterizing deep-level defects in a semiconductor is described in Section 6.10.

6.2. Nonradiative Recombination: The Shockley–Read–Hall Model

In the nonradiative recombination process, the recombination of electron-hole pairs may take place at the localized trap states in the forbidden gap of a semiconductor. This process involves the capture of electrons (or holes) by the trap states, followed by the recombination with holes in the valence band (or electrons in the conduction band). When electron–hole pairs recombine, energy is released via phonon emission. The localized trap states may be created by deep-level impurities (e.g., transition metals or normal metals such as Fe, Ni, Co, W, Au), or by radiation- and process-induced defects such as vacancies, interstitials, antisite defects and their complexes, dislocations, and grain boundaries. The nonradiative recombination process in a semiconductor can be best described by the Shockley–Read–Hall (SRH) model,[1,2] which is discussed next.

Figure 6.1 illustrates the energy band diagram for the SRH model. In this figure, the four transition processes for the capture and emission of electrons and holes via a localized recombination center are shown. A localized deep-level trap state may be in one of the two charge states differing by one electronic charge. Therefore, the trap could be in either a neutral or a negatively charged state or in a neutral or a positively charged state. If the trap state is neutral, then it can capture an electron from the conduction band. This capture is illustrated in Figure 6.1a. In this case, the capture of electrons by an empty neutral trap state is accomplished through the simultaneous emission of phonons during the capture process. Figure 6.1b shows the emission of an electron from a filled trap state. In this illustration, the electron gains its kinetic energy from the thermal energy of the host lattice. Figure 6.1c shows the capture of a hole from the valence band by a filled trap state, and Figure 6.1d shows the emission of a hole from the empty trap state to the valence band.

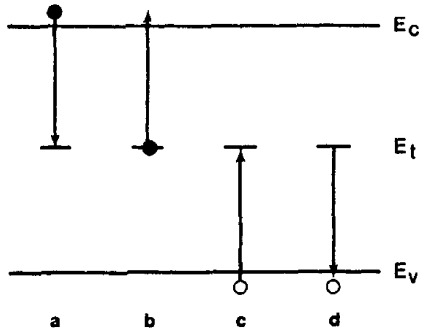


FIGURE 6.1. Capture and emission of an electron and a hole via a deep-level trap. The Shockley–Read–Hall model: (a) electron capture, (b) electron emission, (c) hole capture, and (d) hole emission.

The rate equations that describe the SRH model can be derived from the four emission and capture processes shown in Figure 6.1. In deriving the SRH model, it is assumed that the semiconductor is nondegenerate and that the density of trap states is small compared to the majority carrier density. When the specimen is in thermal equilibrium, f_t denotes the probability that a trap state located at E_t in the forbidden gap is occupied by an electron. Using the Fermi–Dirac (F-D) statisics described in Chapter 3, the distribution function f_t of a carrier at the trap state is given by

$$f_t = \frac{1}{(1 + e^{(E_t - E_f)/k_B T})} . \tag{6.1}$$

The physical parameters used in the SRH model are defined as follows:

U_{cn} is the electron capture probability per unit time per unit volume (cm^{-3}/s).
U_{en} is the electron emission probability per unit time per unit volume.
U_{cp} is the hole capture probability per unit time per unit volume.
U_{ep} is the hole emission probability per unit time per unit volume.
c_n and c_p are the electron and hole capture coefficients (cm^3/s).
e_n and e_p are the electron and hole emission rates (s^{-1}).
N_t is the trap density (cm^{-3}).

In general, the rate of electron capture probability is a function of the density of electrons in the conduction band, capture cross section, and density of the empty traps. However, the rate of electron emission probability depends only on the electron emission rate and the density of traps being filled by the electrons. Thus, the expressions for U_{cn} and U_{en} can be written as

$$U_{cn} = c_n n N_t (1 - f_t), \tag{6.2}$$

$$U_{en} = e_n N_t f_t . \tag{6.3}$$

Similarly, U_{cp}, the rate of hole capture probability, and U_{ep}, the rate of hole emission probability, are given by

$$U_{cp} = c_p p N_t f_t, \tag{6.4}$$

$$U_{ep} = e_p N_t (1 - f_t) . \tag{6.5}$$

According to the principle of detailed balance, the rates of emission and capture at a trap level are equal in thermal equilibrium. Thus, one can write

$$U_{\mathrm{cn}} = U_{\mathrm{en}} \quad \text{for electrons,} \tag{6.6}$$

$$U_{\mathrm{cp}} = U_{\mathrm{ep}} \quad \text{for holes.} \tag{6.7}$$

Solving (6.2) through (6.7) yields

$$e_{\mathrm{n}} = c_{\mathrm{n}} n_0 (1 - f_{\mathrm{t}})/f_{\mathrm{t}}, \tag{6.8}$$

$$e_{\mathrm{p}} = c_{\mathrm{p}} p_0 f_{\mathrm{t}}/(1 - f_{\mathrm{t}}). \tag{6.9}$$

From (6.8) and (6.9) one obtains

$$e_{\mathrm{n}} e_{\mathrm{p}} = c_{\mathrm{n}} c_{\mathrm{p}} n_0 p_0 = c_{\mathrm{n}} c_{\mathrm{p}} n_i^2. \tag{6.10}$$

From (6.1) one can write

$$(1 - f_{\mathrm{t}})/f_{\mathrm{t}} = \mathrm{e}^{(E_{\mathrm{t}} - E_{\mathrm{f}})/k_{\mathrm{B}}T}. \tag{6.11}$$

Now solving (6.8), (6.9), and (6.11), one obtains

$$e_{\mathrm{n}} = c_{\mathrm{n}} n_1, \tag{6.12}$$

$$e_{\mathrm{p}} = c_{\mathrm{p}} p_1, \tag{6.13}$$

where n_1 and p_1 denote the electron and hole densities, respectively, when the Fermi level E_{f} coincides with the trap level E_{t}. Expressions for n_1 and p_1 are given respectively by

$$n_1 = n_0 \, \mathrm{e}^{(E_{\mathrm{t}} - E_{\mathrm{f}})/k_{\mathrm{B}}T}, \tag{6.14}$$

$$p_1 = p_0 \, \mathrm{e}^{(E_{\mathrm{f}} - E_{\mathrm{t}})/k_{\mathrm{B}}T}. \tag{6.15}$$

Solving (6.14) and (6.15) yields

$$n_1 p_1 = n_0 p_0 = n_i^2. \tag{6.16}$$

Under steady-state conditions, the net rate of electron capture per unit volume may be found by solving (6.2) through (6.16), which yields

$$U_{\mathrm{n}} = U_{\mathrm{cn}} - U_{\mathrm{en}} = c_{\mathrm{n}} N_{\mathrm{t}} [n(1 - f_{\mathrm{t}}) - n_1 f_{\mathrm{t}}]. \tag{6.17}$$

Similarly, the net rate of hole capture per unit volume may be written as

$$U_{\mathrm{p}} = U_{\mathrm{cp}} - U_{\mathrm{ep}} = c_{\mathrm{p}} N_{\mathrm{t}} [p f_{\mathrm{t}} - p_1 (1 - f_{\mathrm{t}})]. \tag{6.18}$$

The excess carrier lifetimes under steady-state conditions are defined by the ratio of the excess carrier density and the net capture rate for electrons and holes, and are given respectively by

$$\tau_{\mathrm{n}} = \frac{\Delta n}{U_{\mathrm{n}}} \quad \text{for electrons,} \tag{6.19}$$

$$\tau_{\mathrm{p}} = \frac{\Delta p}{U_{\mathrm{p}}} \quad \text{for holes.} \tag{6.20}$$

For the small-injection case (i.e., $\Delta n \ll n_0$ and $\Delta p \ll p_0$), the charge-neutrality condition requires that

$$\Delta n = \Delta p. \tag{6.21}$$

Under steady-state conditions, if one assumes that the net rates of electron and hole capture via a recombination center are equal, then one can write

$$U = U_{\mathrm{n}} = U_{\mathrm{p}}. \tag{6.22}$$

Substituting (6.21) and (6.22) into (6.19) and (6.20) one finds that the electron and hole lifetimes are equal (i.e., $\tau_{\mathrm{n}} = \tau_{\mathrm{p}}$) for the small-injection case. The electron distribution function f_{t} at the trap level can be expressed in terms of the electron and hole capture coefficients as well as the electron and hole densities. Now solving (6.17), (6.18), and (6.21), one obtains

$$f_{\mathrm{t}} = \frac{(c_{\mathrm{n}} n + c_{\mathrm{p}} p_1)}{c_{\mathrm{n}}(n + n_1) + c_{\mathrm{p}}(p + p_1)}. \tag{6.23}$$

A general expression for the net recombination rate can be obtained by substituting (6.23) into (6.17) or (6.18), and the result is

$$U = U_{\mathrm{n}} = U_{\mathrm{p}} = \frac{(np - n_i^2)}{\tau_{p0}(n + n_1) + \tau_{n0}(p + p_1)}, \tag{6.24}$$

where τ_{p0} and τ_{n0} are given respectively by

$$\tau_{p0} = \frac{1}{c_{\mathrm{p}} N_{\mathrm{t}}}, \tag{6.25}$$

$$\tau_{n0} = \frac{1}{c_{\mathrm{n}} N_{\mathrm{t}}}, \tag{6.26}$$

where $c_{\mathrm{p}} = \sigma_{\mathrm{p}} \langle \nu_{\mathrm{th}} \rangle$ and $c_{\mathrm{n}} = \sigma_{\mathrm{n}} \langle \nu_{\mathrm{th}} \rangle$ denote the hole- and electron-capture coefficients; σ_{p} and σ_{n} are the hole- and electron-capture cross-sections, respectively, and $\langle \nu_{\mathrm{th}} \rangle = (3k_{\mathrm{B}}T/m^*)^{1/2}$ is the average thermal velocity of electrons or holes; τ_{p0} is the minority hole lifetime for an n-type semiconductor, and τ_{n0} is the minority electron lifetime for a p-type semiconductor. Now solving (6.17) through (6.26), one obtains a general expression for the excess carrier lifetime, which is given by

$$\tau_0 = \frac{\Delta n}{U_{\mathrm{n}}} = \frac{\Delta p}{U_{\mathrm{p}}} = \frac{\tau_{p0}(n_0 + n_1 + \Delta n)}{(n_0 + p_0 + \Delta n)} + \frac{\tau_{n0}(p_0 + p_1 + \Delta p)}{(n_0 + p_0 + \Delta p)}, \tag{6.27}$$

where $n = n_0 + \Delta n$ and $p = p_0 + \Delta p$ denote the nonequilibrium electron and hole densities, n_0 and p_0 are the equilibrium electron and hole densities, and Δn and Δp denote the excess electron and hole densities, respectively.

For the small-injection case (i.e., $\Delta n \ll n_0$ and $\Delta p \ll p_0$), the excess carrier lifetime given by (6.27) reduces to

$$\tau_0 = \frac{\tau_{p0}(n_0 + n_1)}{(n_0 + p_0)} + \frac{\tau_{n0}(p_0 + p_1)}{(n_0 + p_0)}, \tag{6.28}$$

which shows that under small-injection conditions, τ_0 is independent of the excess carrier density or injection. It is interesting to note that for an n-type semiconductor with $n_0 \gg p_0, n_1$, and p_1, (6.28) reduces to

$$\tau_0 = \tau_{p0}. \tag{6.29}$$

Similarly, for a p-type semiconductor with $p_0 \gg n_0$, p_1, and n_1, (6.28) becomes

$$\tau_0 = \tau_{n0}. \tag{6.30}$$

Equations (6.29) and (6.30) show that the excess carrier lifetime in an extrinsic semiconductor is dominated by the minority carrier lifetime. Therefore, the minority carrier lifetime is a key physical parameter for determining the excess carrier recombination in an extrinsic semiconductor under low-injection conditions.

For the high-injection case with $\Delta n = \Delta p \gg n_0, p_0$, (6.27) becomes

$$\tau_{\mathrm{h}} = \tau_{p0} + \tau_{n0}, \tag{6.31}$$

which shows that in the high-injection limit the excess carrier lifetime τ_{h} reaches a maximum value and becomes independent of the injection. In general, it is found that in the intermediate-injection ranges the excess carrier lifetime may depend on the injected carrrier density. Furthermore, it is found that the excess carrier lifetime also depends on n_1 and p_1, which in turn depend on the Fermi level and the dopant density. This is clearly illustrated in Figure 6.2.

On the basis of the above discussions, the minority carrier lifetime is an important physical parameter that is directly related to the recombination mechanisms in a semiconductor. A high-quality semiconductor with few defects generally has long minority carrier lifetime, while a poor-quality semiconductor usually has short minority carrier lifetime and large defect density. The minority carrier lifetime plays an important role in the performance of semiconductor devices. For example, the switching speed of a bipolar junction transistor and the conversion efficiency

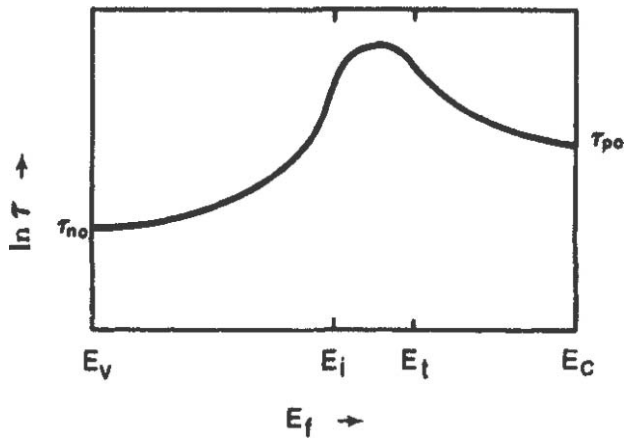


FIGURE 6.2. Dependence of the excess carrier lifetime on the Fermi level.

of a p-n junction solar cell depend strongly on the minority carrier lifetimes of a semiconductor.

It is noted that the SRH model presented in this section is applicable for describing the nonradiative recombination process via a single deep-level recombination center in the forbidden gap of a semiconductor. Treatment of the nonradiative recombination process via multiple deep-level centers in the forbidden gap of the semiconductor can be found in a classical paper by Sah and Shockley.[3]

6.3. Band-to-Band Radiative Recombination

Band-to-band radiative recombination in a semiconductor is the inverse process of optical absorption. Emission of photons as a result of band-to-band radiative recombination is a common phenomenon observed in a direct band gap semiconductor such as GaAs, GaN, or ZnSe. In a nondegenerate semiconductor, the rate at which electrons and holes are annihilated via band-to-band radiative recombination is proportional to the product of electron and hole densities in the conduction and valence bands, respectively. In thermal equilibrium, the rate of band-to-band recombination is equal to the rate of thermal generation, which can be expressed by

$$R_0 = G_0 = B_{\mathrm{r}} n_0 p_0 = B_r n_i^2, \tag{6.32}$$

where B_{r} is the rate of radiative capture probability, which can be derived from the optical absorption process using the principle of detailed balance. Under steady-state conditions, the rate of band-to-band radiative recombination is given by

$$r = B_{\mathrm{r}} np, \tag{6.33}$$

where $n = n_0 + \Delta n$ and $p = p_0 + \Delta p$. The net recombination rate is obtained by solving (6.32) and (6.33), resulting in

$$U_{\mathrm{r}} = r - G_0 = B_{\mathrm{r}}(np - n_i^2). \tag{6.34}$$

The radiative lifetime τ_{r} as a result of band-to-band recombination is obtained by solving (6.34) and (6.21), which yields

$$\tau_{\mathrm{r}} = \frac{\Delta n}{U_{\mathrm{r}}} = \frac{1}{B_{\mathrm{r}}(n_0 + p_0 + \Delta n)}. \tag{6.35}$$

From (6.35), it is noted that τ_{r} is inversely proportional to the majority carrier density n_0. Under small-injection conditions, (6.35) can be simplified to

$$\tau_{\mathrm{r0}} = \frac{1}{B_{\mathrm{r}}(n_0 + p_0)}, \tag{6.36}$$

which shows that for the small-injection case, the band-to-band radiative lifetime is inversely proportional to the majority carrier density. For the intrinsic case

(i.e., $n_0 = p_0 = n_i$), the radiative lifetime τ_{ri} due to the band-to-band recombination is given by $\tau_{ri} = 1/(2B_r n_i)$.

In the high-injection limit, $\Delta n = \Delta p \gg n_0$, p_0, and (6.35) becomes

$$\tau_{rh} = \frac{1}{B_r \Delta n} \,. \tag{6.37}$$

Equation (6.37) shows that the band-to-band radiative lifetime under high-injection conditions is inversely proportional to the excess carrier density, and is independent of the majority carrier density in the semiconductor.

Since band-to-band radiative recombination is the inverse process of optical absorption, an analytical expression for the radiative recombination capture rate B_r can be derived from the fundamental optical absorption process using the principle of detailed balance.

In a direct band gap semiconductor, the fundamental absorption process is usually dominated by the vertical transition. As will be shown in Chapter 9, the energy dependence of the fundamental optical absorption coefficient for a direct band gap semiconductor can be expressed by

$$\alpha_d = \left(\frac{2^{2/3} q^2}{3nm_0 ch^2}\right)(m_r^{3/2} + m_0 m_r^{1/2})(h\nu - E_g)^{1/2} \,, \tag{6.38}$$

where n is the index of refraction, E_g is the energy band gap, $m_r^{-1} = (m_e + m_h)/m_e m_h$ is the reduced electron and hole effective mass, and m_0 is the free-electron mass. Equation (6.38) shows that for $h\nu \geq E_g$, the optical absorption coefficient for a direct band gap semiconductor is proportional to the square root of the photon energy.

In order to correlate the rate of capture probability coefficient B_r to the optical absorption coefficient α_d, one can treat the semiconductor as a blackbody radiation source and use the principle of detailed balance under thermal equilibrium conditions. From (6.32), B_r may be determined by setting the rate of radiative recombination equal to the rate of total blackbody radiation absorbed by the semiconductor due to band-to-band radiative recombination, which can be expressed by

$$B_r n_i^2 = \int \frac{n^2 \alpha E^2 \, dE}{(\pi^2 q^2 h^3)(e^{E/k_B T} - 1)} \,, \tag{6.39}$$

where $E = h\nu$ is the photon energy. The right-hand side of (6.39) is obtained from the Planck blackbody radiation formula. Solving (6.38) and (6.39), one obtains the rate of capture probability B_r for the direct transition, which reads

$$B_r = \left(\frac{E_g}{n_i}\right)^2 (2\pi)^{3/2} \left(\frac{hq^2}{3m_0^2 c^2}\right) \eta (1 + m_0/m_r) \left(\frac{m_0}{m_e + m_h}\right)^{3/2} (k_B T)^{-3/2} (m_0 c^2)^{-1/2}. \tag{6.40}$$

It is important to note from (6.40) that B_r is inversely proportional to the square of the intrinsic carrier density, which shows an exponential dependence of B_r on temperature. This implies that the band-to-band radiative recombination

TABLE 6.1. Band-to-band radiative recombination parameters for some elemental and compound semiconductors at 300 K.

Semiconductors	E_g (eV)	n_i (cm^{-3})	B_r or B_d (cm^3/s)	τ_i	$\tau_0{}^a$(μs)
Si	1.12	1.5×10^{10}	2.0×10^{-15}	4.6 h	2500
Ge	0.67	2.4×10^{13}	3.4×10^{-15}	0.61s	150
GaSb	0.71	4.3×10^{12}	1.3×10^{-11}	9 ms	0.37
InAs	0.31	1.6×10^{15}	2.1×10^{-11}	15μs	0.24
InSb	0.18	2×10^{16}	4×10^{-11}	0.62 μs	0.12
PbTe	0.32	4×10^{15}	5.2×10^{-11}	2.4 μs	0.19

[a] Calculated, assuming n_0 or $p_0 = 10^{17}$ cm^{-3}. τ_i: lifetime due to indirect transition, τ_0: lifetime due to direct transition.

lifetime is a strong function of temperature. Table 6.1 lists the values of B_r calculated from (6.40) for GaSb, InAs, and InSb. The results are found to be in reasonable agreement with the published data for these materials.

A similar calculation of the capture probablity for the indirect transition involving the absorption and emission of phonons in an indirect band gap semiconductor yields the capture probability coefficient B_i, which is given by

$$B_i = \left(\frac{4\pi h^3}{m_0^3 c^3}\right)(A\mu^2)\left(\frac{m_0^2}{m_e m_h}\right)^{3/2} E_g^2 \coth(\theta/2T)\,, \tag{6.41}$$

where A and μ are adjustable parameters used to fit the measured absorption data. Equation (6.41) shows that B_i depends weakly on temperature. For a direct band gap semiconductor in which recombination is via band-to-band radiative transition, values of B_r can be quite high (i.e., 3×10^{-11} cm^3/s). On the other hand, for indirect transitions, values of B_i are found to be 3 to 4 orders of magnitude smaller than B_r for direct transitions. Table 6.1 lists the calculated values of B_r and B_i and the radiative lifetimes for some direct and indirect band gap semiconductors at $T = 300$ K.

6.4. Band-to-Band Auger Recombination

As discussed in Section 6.3, band-to-band radiative recombination is the inverse process of fundamental optical absorption in a semiconductor. In a similar manner, Auger recombination is the inverse process of impact ionization. Band-to-band Auger recombination is a three-particle process, which involves either electron–electron collisions in the conduction band followed by recombination with holes in the valence band, or hole–hole collisions in the valence band followed by recombination with electrons in the conduction band. These two recombination processes and their inverse processes are shown schematically in Figure 6.3.

For small-band-gap semiconductors such as InSb, the minority carrier lifetime is usually controlled by band-to-band Auger recombination, and energy loss is carried

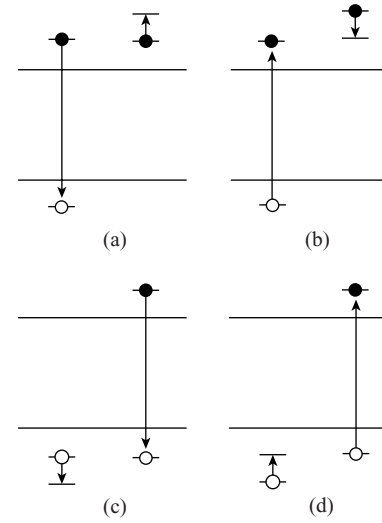


FIGURE 6.3. Auger recombination and its inverse process, which shows the annihilation and creation of an electron–hole pair: (a) annihilation of an electron–hole pair by electron–electron collisions, (b) creation of an electron–hole pair by electron impact ionization, (c) destruction of an electron–hole pair by hole–hole collisions, and (d) creation of an electron–hole pair by hole impact ionization.

out either by electron–electron collisions or hole–hole collisions and subsequent Auger recombination.

To derive the band-to-band Auger recombination lifetime, the rate of Auger recombination in equilibrium conditions can be written as

$$R_{\mathrm{a}} = G_0 = C_{\mathrm{n}} n_0^2 p_0 + C_{\mathrm{p}} p_0^2 n_0. \tag{6.42}$$

Under nonequilibrium conditions, the Auger recombination rate is given by

$$r_{\mathrm{A}} = C_{\mathrm{n}} n^2 p + C_{\mathrm{p}} p^2 n. \tag{6.43}$$

Therefore, the net Auger recombination rate under steady-state conditions can be obtained from (6.42) and (6.43), which yields

$$U_{\mathrm{A}} = r_{\mathrm{A}} - G_0 = C_{\mathrm{n}}(n^2 p - n_0^2 p_0) + C_{\mathrm{p}}(p^2 n - p_0^2 n_0), \tag{6.44}$$

where C_{n} and C_{p} are the capture probability coefficients when the third carrier is either an electron or a hole. Both C_{n} and C_{p} can be calculated from their inverse process, namely, impact ionization. In thermal equilibrium, the rate at which carriers are annihilated via Auger recombination is equal to the generation rate averaged over the Boltzmann distribution function in which the electron–hole pairs are generated by impact ionization. Thus, one obtains

$$C_{\mathrm{n}} n_0^2 p_0 = \int_0^\infty P(E)(\mathrm{d}n/\mathrm{d}E)\,\mathrm{d}E, \tag{6.45}$$

where $P(E)$ is the probability per unit time that an electron with energy E makes an ionizing collision. It can be described by

$$P(E) = (mq^4/2h^3)G(E/E_{\mathrm{t}} - 1)^s, \tag{6.46}$$

where $G < 1$ is a parameter that is a complicated function of the band structure of the semiconductor. The exponent s is an integer that is determined by the symmetry of the crystal in momentum space at a threshold energy E_t. The value of E_t for impact ionization is roughly equal to $1.5E_g$, where E_g is the energy band gap of the semiconductor. By substituting (6.46) into (6.45), one obtains

$$n_i^2 C_n = \left(\frac{s}{\sqrt{\pi}}\right)\left(\frac{mq^4}{h^3}\right) G \left(\frac{k_B T}{E_t}\right)^{(s-1/2)} e^{-E_t/k_B T}. \tag{6.47}$$

Equation (6.47) shows that the Auger capture coefficient C_n for electrons depends exponentially on both the temperature and energy band gap of the semiconductor. The Auger lifetime may be derived from (6.44), with the result

$$\tau_A = \frac{\Delta n}{U_A} = \frac{1}{n^2 C_n + 2n_i^2(C_n + C_p) + p^2 C_p}. \tag{6.48}$$

If one assumes that $C_n = C_p$ and $n = p = n_i$, then (6.48) shows that τ_A has a maximum value of $\tau_i = 1/6n_i^2 C_n$ for an intrinsic semiconductor. For an extrinsic semiconductor, τ_A is inversely proportional to the square of the majority carrier density. For the intrinsic case, the Auger lifetime can be obtained from (6.47) and (6.48) with $s = 2$ and $C_n \neq C_p$, which yields

$$\tau_{Ai} = \frac{1}{3n_i^2(C_n + C_p)} = 3.6 \times 10^{-17}(E_t/k_B T)^{3/2} e^{E_t/k_B T}, \tag{6.49}$$

which shows that the intrinsic Auger lifetime is an exponential function of temperature and energy band gap ($E_t \approx 1.5E_g$). Note that the temperature dependence of the Auger lifetime in an extrinsic semiconductor is not as strong as in an intrinsic semiconductor. However, because of the strong temperature dependence of the Auger lifetime, it is possible to identify the Auger recombination process by analyzing the measured lifetime as a function of temperature in a semiconductor. For a heavily doped semiconductor, (6.48) predicts that the Auger lifetime is inversely proportional to the square of the majority carrier density. The Auger recombination has been found to be the dominant recombination process for degenerate semiconductors and small-band-gap semiconductors. Values of Auger recombination coefficients for silicon and germanium are $C_n = 2.8 \times 10^{-31}$ and $C_p = 10^{-31}$ cm^6/s for silicon, $C_n = 8 \times 10^{-32}$ and $C_p = 2.8 \times 10^{-31}$ cm^6/s for germanium. Using these values, the intrinsic Auger lifetime for silicon is equal to 4.48×10^9 s at 300 K, and is equal to 1.61×10^3 s for germanium. Thus, the Auger recombination is a very unlikely recombination process for intrinsic semiconductors (with the exception of small-band-gap semiconductors such as InSb). It is noted that the Auger recombination lifetime for n-type silicon reduces to about 10^{-8} s at a doping density of 10^{19} cm^{-3}.

Under high-injection conditions (i.e., $n_0, p_0 \ll \Delta n = \Delta p$), Auger recombination may become the predominant recombination process. In this case, the Auger

lifetime is given by

$$\tau_{\mathrm{Ah}} = \frac{1}{\Delta n^2 (C_{\mathrm{n}} + C_{\mathrm{p}})} = \left(\frac{3n_i^2}{\Delta n^2} \right) \tau_{\mathrm{Ai}}, \tag{6.50}$$

where τ_{Ai} is the intrinsic Auger lifetime given by (6.49).

As an example, consider a germanium specimen. If the injected carrier density is $\Delta n = 10^{18}$ cm^{-3} and $E_{\mathrm{t}} = 1.0$ eV, then the Auger lifetime τ_{Ai}, as calculated from (6.50), was found equal to 1 μs. For small-band-gap semiconductors, one expects the Auger recombination to be the predominant recombination process even at smaller injection level. Additional discussion on the Auger recombination and the band-to-band radiative recombination mechanisms in semiconductors can be found in a special issue of *Solid State Electronics* edited by Landsberg and Willoughby.[4]

In order to obtain an overall picture of the various recombination processes taking place in a semiconductor, Figures 6.4 and 6.5 show the qualitative plots of the excess carrier lifetimes due to different recombination mechanisms as a function of the majority carrier density for a Ge and GaSb crystal respectively.

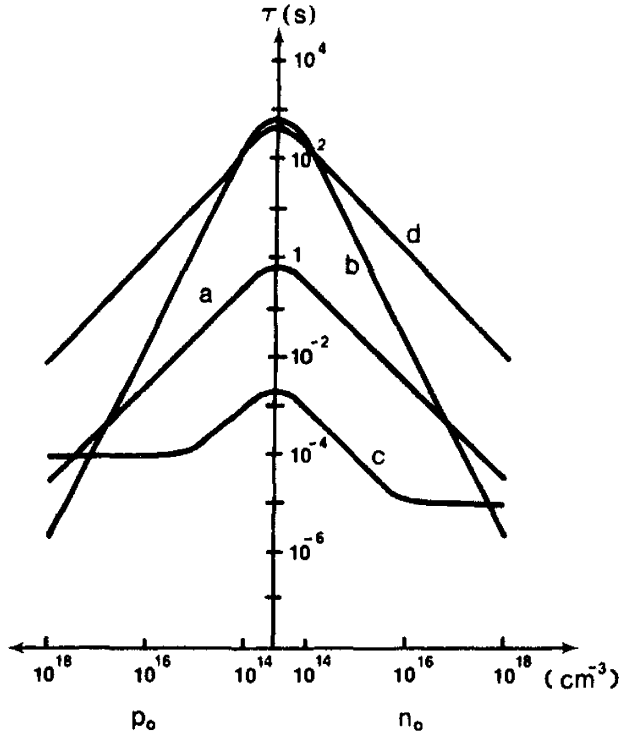


FIGURE 6.4. A comparison of the recombination lifetimes at $T = 300$ K for a germanium crystal for the cases in which recombination is dominated by (a) band-to-band radiative recombination, (b) Auger band-to-band recombination, (c) multiphonon process (Shockley–Read–Hall model), and (d) impurity-to-band Auger recombination.

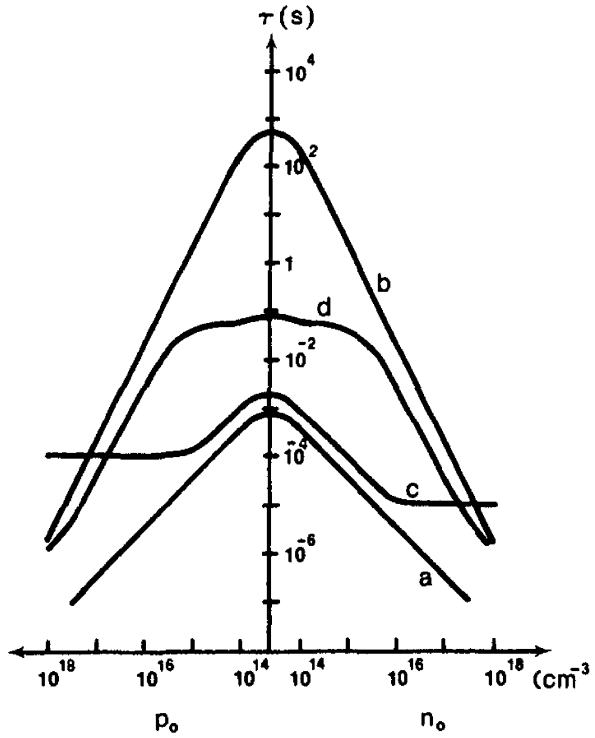


FIGURE 6.5. A comparison of the recombination lifetimes at 300 K for a direct-gap semiconductor such as GaSb when the recombination is dominated by (a) band-to-band radiative recombination, (b) Auger band-to-band recombination, (c) multiphonon process, and (d) Auger impurity-to-band recombination.

From these two figures, a significant difference in the dominant recombination process was observed between these two materials. The difference in the dominant recombination process in Ge and GaSb can be attributed to the fact that Ge is an indirect band gap semiconductor, while GaSb is a direct band gap semiconductor. For Ge, the SRH recombination process is expected to be the predominant process over a wide range of doping densities (except in the very high doping densities), while for GaSb the band-to-band radiative recombination is expected to be the predominant process for the low to medium doping density ranges.

6.5. Basic Semiconductor Equations

The spatial and time-varying function of the excess carrier phenomena in a semiconductor under nonequilibrium conditions may be analyzed by using the basic semiconductor equations. These equations contain the drift and diffusion

components (for both electrons and holes) as well as the recombination and generation terms. There are two continuity equations for the excess carriers in a semiconductor: one for electrons and one for holes. As will be discussed later, both the steady-state and transient effects can, in principle, be solved from these two continuity equations.

In a semiconductor, the electron–hole pairs can be created by either thermal or optical means and annihilated by different recombination processes. In thermal equilibrium, the rate of generation must be equal to the rate of recombination. Otherwise, space charge will be built up within the semiconductor specimen. The nonequilibrium condition is established when an external excitation is applied to the semiconductor specimen. For example, excess electron–hole pairs can be generated in a semiconductor by the absorption of photons with energies greater than the band gap energy (i.e., $h\nu \geq E_g$) of the semiconductor. The continuity equations for both electrons and holes under nonequilibrium conditions are given respectively by

$$\frac{dn}{dt} = \frac{1}{q}\nabla \cdot J_n - \frac{n}{\tau_n} + g_T, \tag{6.51}$$

$$\frac{dp}{dt} = \frac{-1}{q}\nabla \cdot J_p - \frac{p}{\tau_p} + g_T, \tag{6.52}$$

where $n = \Delta n + n_0$ and $p = \Delta p + p_0$ are the nonequilibrium electron and hole densities, respectively; g_T is the total generation rate; n/τ_n and p/τ_p are the rates of recombination for electrons and holes; τ_n and τ_p are the electron and hole lifetimes; and J_n and J_p denote the electron and hole current densities, respectively. In general, in addition to the thermal generation rate, the excess electron–hole pairs can be created by external excitation. Thus, the total generation rate can be written as

$$g_T = G_{th} + g_E, \tag{6.53}$$

where G_{th} is the thermal generation rate and g_E is the external generation rate. According to the principle of detailed balance, under thermal equilibrium, the rate of generation must be equal to the rate of recombination. Thus, in thermal equilibrium, one can write

$$G_{th} = R_0 = \frac{n_0}{\tau_n} = \frac{p_0}{\tau_p}. \tag{6.54}$$

The continuity equations for the excess electron and hole densities can be obtained by substituting (6.53) and (6.54) into (6.51) and (6.52), yielding

$$\frac{\partial \Delta n}{\partial t} = \frac{1}{q}\nabla \cdot J_n - \frac{\Delta n}{\tau_n} + g_E, \tag{6.55}$$

$$\frac{\partial \Delta p}{\partial t} = \frac{-1}{q}\nabla \cdot J_p - \frac{\Delta p}{\tau_p} + g_E. \tag{6.56}$$

The electron and hole current densities in a semiconductor consist of two components, namely, the drift and diffusion currents. These two current components are given respectively by

$$J_n = q\mu_n n\mathcal{E} + qD_n\nabla n, \tag{6.57}$$

$$J_p = q\mu_p p\mathcal{E} - qD_p\nabla p, \tag{6.58}$$

where $\mathcal{E}$ is the electric field; μ_n and μ_p are the electron and hole mobilities; and D_n and D_p are the electron and hole diffusivities, respectively. The first term on the right-hand side of (6.57) and (6.58) is called the drift current component, while the second term is the diffusion current component. The total current density is equal to the sum of electron and hole current densities, which is given by

$$J_{\mathrm{T}} = J_n + J_p. \tag{6.59}$$

In thermal equilibrium, both the electron and hole current densities are equal to zero. Now, letting $J_n = 0$ and $J_p = 0$ in (6.57) and (6.58), one obtains

$$D_n = -\left(\frac{n_0}{|\nabla n_0|}\right)\mu_n\mathcal{E}, \tag{6.60}$$

$$D_p = +\left(\frac{p_0}{|\nabla p_0|}\right)\mu_p\mathcal{E}. \tag{6.61}$$

The electric field $\mathcal{E}$ in a bulk semiconductor can be related to the electrostatic potential ϕ by

$$\mathcal{E} = -\nabla\phi. \tag{6.62}$$

If a concentration gradient due to the nonuniform impurity profile exists in a semiconductor, then a chemical potential term must be added to the electrostatic potential term given in (6.62). This is usually referred to as the electrochemical potential or the Fermi potential. The equilibrium carrier density for both electrons and holes can also be expressed in terms of the intrinsic carrier density and the electrostatic potential using M-B statistics, which are given by

$$n_0 = n_i\, \mathrm{e}^{q\phi/k_{\mathrm{B}}T}, \tag{6.63}$$

$$p_0 = n_i\, \mathrm{e}^{-q\phi/k_{\mathrm{B}}T}, \tag{6.64}$$

where ϕ $(= (E_{\mathrm{f}} - E_i)/k_{\mathrm{B}}T)$ is the electrostatic potential measured relative to the intrinsic Fermi level E_i, and n_i is the intrinsic carrier density. Solving (6.60) through (6.64) yields the relationships between μ_n and D_n, and μ_p and D_p in thermal equilibrium, which are given respectively by

$$D_n = \left(\frac{k_{\mathrm{B}}T}{q}\right)\mu_n, \tag{6.65}$$

$$D_p = \left(\frac{k_{\mathrm{B}}T}{q}\right)\mu_p. \tag{6.66}$$

Equations (6.65) and (6.66) are the well-known Einstein relations. The Einstein relation shows that under thermal equilibrium, the ratio of diffusivity

and mobility of electrons and holes (i.e., D_n/μ_n and D_p/μ_p) in a semiconductor is equal to $k_B T/q$. This relation is valid for the nondegenerate semiconductors. For heavily doped semiconductors, Fermi statistics should be used instead, and (6.65) and (6.66) must be modified to account for the degeneracy effect (see Problem 6.5).

In addition to the five basic semiconductor equations described above, Poisson's equation should also be included. This equation, which relates the divergence of the electric field to the charge density in a semiconductor, is given by

$$\nabla \cdot \mathcal{E} = -\nabla \phi^2 = \frac{\rho}{\varepsilon_0 \varepsilon_s} = \left(\frac{q}{\varepsilon_0 \varepsilon_s}\right)\left(N_D^+ - N_A^- + p - n\right), \tag{6.67}$$

where N_D^+ and N_A^- denote the ionized donor and acceptor impurity densities, respectively, and ε_s is the dielectric constant of the semiconductor. Equations (6.55) through (6.59) plus (6.67) are known as the six basic semiconductor equations, which are commonly used in solving a wide variety of spatial and time-dependent problems related to the steady-state and transient behavior of the excess carriers in a semiconductor. Examples of using these basic semiconductor equations to solve the excess carrier phenomena in a semiconductor are given in Sections 6.7 and 6.8.

6.6. The Charge-Neutrality Equation

In a homogeneous semiconductor, charge neutrality is maintained under thermal equilibrium conditions, and (6.67) is equal to zero. However, a departure from the charge neutrality condition may arise from one of the following two sources: (1) a nonuniformly doped semiconductor with fully ionized impurities in thermal equilibrium conditions, and (2) unequal densities of electrons and holes arising from carrier trapping under nonequilibrium conditions. In both situations, an electrochemical potential (i.e., the quasi-Fermi potential) and a built-in electric field may be established within the semiconductor. In this section, a nonuniformly doped semiconductor is considered.

From (6.63), the electrostatic potential for an n-type semiconductor can be expressed by

$$\phi = \left(\frac{k_B T}{q}\right) \ln\left(\frac{N}{n_i}\right), \tag{6.68}$$

where $N(x) = N_D - N_A$ is the net dopant density, which could be a function of position in a nonuniformly doped semiconductor. Now, substituting (6.62), (6.63), and (6.64) into (6.67), the Poisson equation becomes

$$\nabla^2 \phi = \left(\frac{2qn_i}{\varepsilon_0 \varepsilon_s}\right)\left[\sinh\left(q\phi/k_B T\right) - (N/2n_i)\right]. \tag{6.69}$$

Equation (6.69) can be rewritten as

$$\nabla^2\varphi = \left(\frac{2q^2 n_i}{k_B T \varepsilon_0 \varepsilon_s}\right)(\sinh(\varphi) - \sinh(\varphi_0))$$
$$= \left(\frac{4q^2 n_i}{k_B T \varepsilon_0 \varepsilon_s}\right)\left[\cosh\left(\frac{\varphi + \varphi_0}{2}\right)\sinh\left(\frac{\varphi - \varphi_0}{2}\right)\right]. \tag{6.70}$$

In (6.70), the normalized electrostatic potential φ is defined by

$$\varphi = q\phi / k_B T \tag{6.71}$$

and

$$\sinh(\varphi_0) = \frac{N}{2n_i}. \tag{6.72}$$

The physical significance of (6.70) can be best described by considering the one-dimensional (1-D) case in which the impurity density N is a function only of x in the semiconductor. If $(\varphi - \varphi_0) \ll 1$ (i.e., a small inhomogeneity in the semiconductor) in (6.70), one obtains

$$\cosh\left(\frac{\varphi + \varphi_0}{2}\right) \approx \cosh(\varphi_0) = \left[1 + \sinh(\varphi_0)^2\right]^{1/2} \approx \frac{N}{2n_i} \tag{6.73}$$

and

$$\sinh\left(\frac{\varphi - \varphi_0}{2}\right) \approx \frac{(\varphi - \varphi_0)}{2}. \tag{6.74}$$

Now substituting (6.73) and (6.74) in (6.70), the 1-D Poisson equation can be written as

$$\frac{\partial^2\varphi}{\partial x^2} \approx \frac{\partial^2(\varphi - \varphi_0)}{\partial x^2} \simeq \left(\frac{q^2 N}{k_B T \varepsilon_0 \varepsilon_s}\right)(\varphi - \varphi_0) = \frac{1}{L_D^2}(\varphi - \varphi_0), \tag{6.75}$$

which has a solution given by

$$\varphi - \varphi_0 \approx e^{-x/L_D}, \tag{6.76}$$

where

$$L_D = \sqrt{\frac{k_B T \varepsilon_0 \varepsilon_s}{q^2 N}} \tag{6.77}$$

is known as the extrinsic Debye length. The physical meaning of L_D is that it is a characteristic length used to determine the distance in which a small variation of the potential can smooth itself out in a homogeneous semiconductor.

Equation (6.76) predicts that in an extrinsic semiconductor under thermal equilibrium, no significant departure from the charge-neutrality condition is expected over a distance greater than a few Debye lengths. It can be shown that L_D in (6.77) for an n-type semiconductor can also be expressed as

$$L_{D_n} = \sqrt{D_n \tau_d}, \tag{6.78}$$

where $\tau_d = \varepsilon_0 \varepsilon_s / \sigma$ is the dielectric relaxation time and σ is the electrical conductivity of the semiconductor.

6.7. The Haynes–Shockley Experiment

In this section an example is given to illustrate how the basic semiconductor equations described in Section 6.6 can be applied to solve the space- and time-dependent excess carrier phenomena in a semiconductor. First consider a uniformly doped n-type semiconductor bar in which N electron–hole pairs are generated instantaneously at $x = 0$ and $t = 0$. If one assumes that the semiconductor bar is infinitely long in the x-direction (see Figure 6.7), then the continuity equation given by (6.55) for the excess holes under a constant applied electric field can be reduced to a 1-D equation, which is given by

$$\frac{\partial \Delta p}{\partial t} = D_p \frac{\partial^2 \Delta p}{\partial x^2} - \mu_p \mathcal{E} \frac{\partial \Delta p}{\partial x} - \frac{\Delta p}{\tau_p}. \tag{6.79}$$

Equation (6.79) is obtained by substituting J_p, given by (6.58), into (6.56) and assuming that the external generation rate g_E is zero. The solution of (6.79) is given by

$$\Delta p(x,t) = \left[\frac{N e^{-t/\tau_p}}{(4\pi D_p t)^{1/2}} \right] \exp\left[-(x - \mu_p \mathcal{E} t)^2 / 4 D_p t \right]. \tag{6.80}$$

From (6.80), it is seen that the initial value of $\Delta p(x, 0)$ is zero except at $x = 0$, where $\Delta p(x, 0)$ approaches infinity. Thus, the initial hole concentration distribution corresponds to a Dirac delta function. For $t > 0$, the distribution of $\Delta p(x, t)$ has a Gaussian shape. The half-width of $\Delta p(x, t)$ will increase with time, and its maximum amplitude will decrease with distance along the direction of the applied electric field, with a drift velocity $v_d = \mu_p \mathcal{E}$. The total excess carrier density injected at time t into the semiconductor is obtained by integrating (6.80) with respect to x from $-\infty$ to $+\infty$, which yields

$$\Delta p(t) = \int_{-\infty}^{+\infty} \Delta p(x,t)\, dx = N e^{-t/\tau_p}. \tag{6.81}$$

Equation (6.81) shows that $\Delta p(t)$ decays exponentially with time and with a time constant equal to the hole lifetime τ_p. Figure 6.6 shows the space and time dependence of the excess carrier density in an n-type extrinsic semiconductor under a constant applied electric field. As shown in this figure, in order to maintain the original injection hole density profile, a large hole lifetime τ_p is needed. This implies that the semiconductor specimen should be of high quality with a very low defect density. Figure 6.7 shows the schematic diagram of the Haynes–Shockley experiment for measuring both the diffusivity and drift mobility of minority carriers in a semiconductor. In this experiment, P_1 and P_2 denote the injection and collector contacts for the minority carriers (i.e., holes in the present case), and

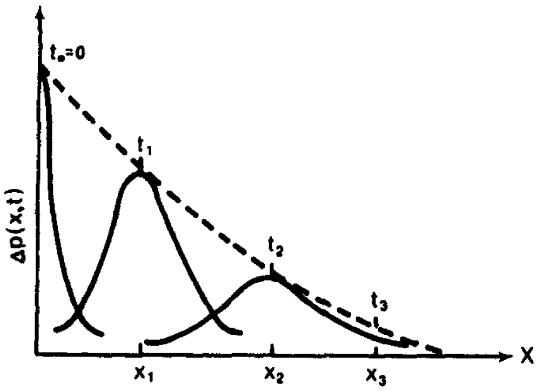


FIGURE 6.6. Space and time dependence of the excess hole density in an n-type extrinsic semiconductor bar under a constant applied electric field.

V_1 and V_2 are the voltages applied to the respective contacts in order to create a uniformed electric field along the specimen and to provide a reverse bias voltage to the collector contact. The injection of minority carriers at contact P_1 can be achieved by using either an electric pulse generator or a pulsed laser. An oscilloscope is used to display the pulse shape at contacts P_1 and P_2 and to measure the time delay of minority carriers traveling between the injecting and collecting contacts.

The Haynes–Shockley experiment is described as follows. At $t = 0$, holes are injected at point P_1 of the sample in the form of a pulse of very short duration (on the order of a few microseconds or less). After this initial hole injection, the excess holes will move along the direction of the applied electric field (i.e., the x-direction) and are collected at contact P_2. This collection results in a current flow and a voltage drop across the load resistor R. The time elapsed between the initial injection pulse at P_1 and the arrival of the collection pulse at P_2 is a measure

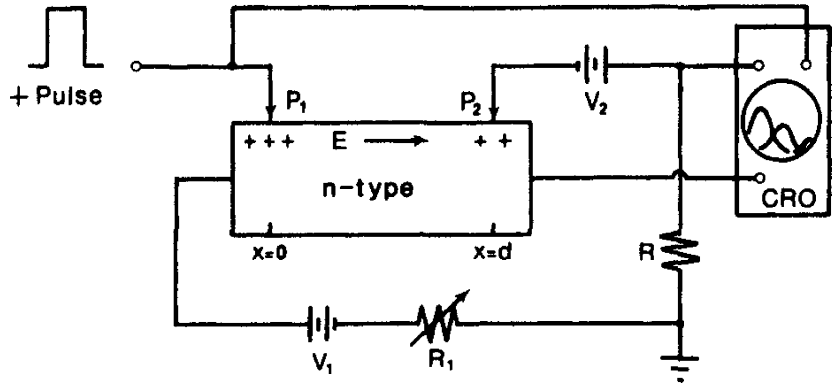


FIGURE 6.7. Schematic diagram for the Haynes–Shockley experiment.

of the drift velocity of holes in the n-type semiconductor bar. In addition to the drift motion along the direction of the applied electric field, the hole density is also dispersed and broadened because of the diffusion effect. This explains why the pulse shown on the right-hand side of Figure 6.6 is not as sharp as the initial injection pulse shown at $x = 0$.

The procedures involved in determining the values of μ_p and D_p from the Haynes–Shockley experiment and (6.80) are discussed as follows: If t_0 is the time required for the peak of the hole pulse to move from contact P_1 to contact P_2 when an electric field is applied to the specimen, then the distance that the hole pulse traveled is given by

$$d = v_d t_0 = \mu_p t_0 \left(\frac{V_a}{l} \right), \tag{6.82}$$

where d is the distance between the injection and collecting contacts of the specimen, and V_a is the applied voltage across the sample of length l. If values of d and t_0 are known, then the hole drift mobility μ_p can be easily calculated from (6.82).

The hole diffusion constant D_p can be determined from the width of the Gaussian distribution function $\Delta p(x, t)$. The output voltage V_R of the hole pulse will drop to 0.367 of its peak value when the second exponential factor on the right-hand side of (6.80) is equal to unity. Thus, one obtains

$$(d - \mu_p \mathcal{E} \Delta t)^2 = 4 D_p \Delta t. \tag{6.83}$$

If t_1 and t_2 denote the two delay time constants that satisfy (6.83) and $\Delta t = t_2 - t_1$, then D_p can be determined from the expression given by

$$D_p = (\mu_p \mathcal{E})^2 (\Delta t)^2 / 16 t_0. \tag{6.84}$$

The approximation given above is valid as long as the exponential factor $-t/\tau_p$ given by (6.80) does not change appreciably over the measured time interval Δt. In practice, the diffusion constants for electrons and holes are determined from the electron and hole mobilities using the Einstein relations given by (6.65) and (6.66). Values of the electron and hole drift mobilities for silicon and germanium determined by the Haynes–Shockley experiment at room temperature are listed in Table 6.2.

TABLE 6.2. Drift mobilities ($cm^2/V \cdot s$) for Si and Ge measured at 300 K using the Haynes–Shockley experiment.

Silicon	Germanium
$\mu_n = 1350 \pm 100$	$\mu_p = 3900 \pm 100$
$\mu_p = 480 \pm 15$	$\mu_p = 1900 \pm 50$

6.8. The Photoconductivity Decay Experiment

In this section, measurement of the minority carrier lifetime in a semiconductor by the transient photoconductivity decay method is depicted. The theoretical and experimental aspects of the transient photoconductivity effect in a semiconductor are discussed. As shown in Figure 6.8, if the semiconductor bar is illuminated by a light pulse, which contains photons with energies greater than the band gap energy of the semiconductor, then electron–hole pairs will be generated in the specimen. The creation of excess carriers by the absorbed photons will result in a change of the electrical conductivity in the semiconductor bar. This phenomenon is known as the photoconductivity effect in a semiconductor. If the light pulse is abruptly turned off at $t = 0$, then the photoconductivity of the specimen will decay exponentially with time and gradually return to its equilibrium value under dark conditions. The time constant of photoconductivity decay is controlled by the lifetimes of minority carriers. By measuring the photoconductivity decay time constant, one can determine the minority carrier lifetime in a semiconductor specimen.

The problem of the transient photoconductivity decay experiment for the excess hole density in an n-type semiconductor can be solved using (6.56). As shown in Figure 6.8, assuming that the light pulse is impinging along the y-direction of the sample, the spatial and time-dependent excess hole density for $t \geq 0$ can be written as

$$\frac{\partial \Delta p}{\partial t} = D_{\mathrm{p}} \frac{\partial^2 \Delta p}{\partial y^2} - \frac{\Delta p}{\tau_{\mathrm{p}}}. \tag{6.85}$$

Equation (6.85) is obtained from (6.56) by assuming that the light pulse is uniformly illuminated in the x–z plane of the specimen so that its diffusion components $\partial^2 \Delta p / \partial x^2$ and $\partial^2 \Delta p / \partial z^2$ are negligible compared to the diffusion component in the y-direction. The electric field is also assumed to be small, so that the drift term in (6.56) can be neglected. As shown in Figure 6.8, the boundary conditions at the

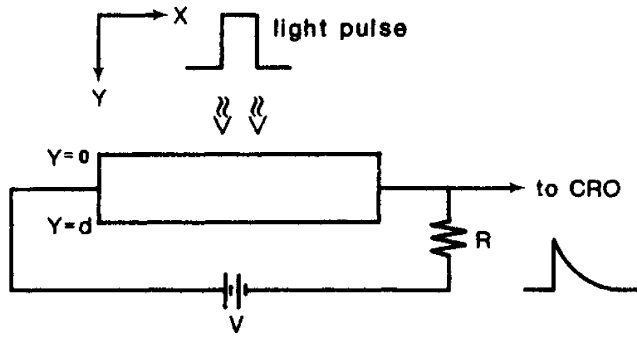


FIGURE 6.8. Photoconductivity-decay experiment for the minority carrier lifetime measurement in a semiconductor.

top and bottom surfaces are given respectively by

$$D_{\mathrm{p}}\frac{\partial \Delta p}{\partial y} = -s_{\mathrm{b}}\Delta p \quad \text{at } y = d, \tag{6.86}$$

$$D_{\mathrm{p}}\frac{\partial \Delta p}{\partial y} = s_{\mathrm{f}}\Delta p \qquad \text{at } y = 0, \tag{6.87}$$

where s_{f} and s_{b} denote the surface recombination velocities at the top and bottom surfaces of the specimen, respectively. The values of s_{b} and s_{f} depend strongly on the surface treatment. In addition to the boundary conditions given by (6.86) and (6.87), the initial and final conditions are assumed by

$$\Delta p(y, t = 0) = \Delta p_0 = \text{constant}, \tag{6.88}$$

$$\Delta p(y, t \to \infty) = 0. \tag{6.89}$$

Since (6.85) is a homogeneous linear partial differential equation for $\Delta p(y, t)$, its solution can be written as the product of two independent functions of t and y:

$$\Delta p(y, t) = A\,\mathrm{e}^{-(b^2 D_{\mathrm{p}} + 1/\tau_{\mathrm{p}})t}\cos(by). \tag{6.90}$$

It is noted that (6.90) does not satisfy the boundary conditions imposed by (6.88) and (6.89). Therefore, the most general solution for (6.85) corresponding to an arbitrary initial condition at $t = 0$ can be expressed in terms of a series sum of the solution given by (6.90) such that

$$\Delta p(y, t) = \sum_{n=0}^{\infty} A_{\mathrm{n}}\mathrm{e}^{-(b_{\mathrm{n}}^2 D_{\mathrm{p}} + 1/\tau_{\mathrm{p}})t}\cos(b_{\mathrm{n}} y). \tag{6.91}$$

Substituting (6.91) into (6.86) for $y = d$ yields the boundary condition

$$\sin(b_{\mathrm{n}} d) = s/D_{\mathrm{p}} b_{\mathrm{n}}. \tag{6.92}$$

Solutions for the surface recombination velocity s can be obtained graphically for different values of b_{n} (i.e., for $n = 0, 1, 2, \ldots$). The coefficient A_{n} in (6.91) can be determined from the initial condition given by (6.88). Furthermore, one can assume that at $t = 0$,

$$\Delta p(y, 0) = \sum_{n=0}^{\infty} A_{\mathrm{n}}\cos(b_{\mathrm{n}} y) = \Delta p_0 = \text{constant}. \tag{6.93}$$

Multiplying (6.93) by $\cos(b_m y)$ and integrating both sides of the equation from $y = 0$ to $y = d$ yields

$$\int_0^d \Delta p_0 \cos(b_m y)\,\mathrm{d}y = \int_0^d \sum_{n=0}^{\infty} A_{\mathrm{n}}\cos(b_{\mathrm{n}} y)\cos(b_m y)\,\mathrm{d}y. \tag{6.94}$$

If a set of functions of $\cos(b_m y)$ and $\cos(b_n y)$ is orthogonal for $0 < y < d$, then the integration on the right-hand side of (6.94) will vanish, except for the term with $n = m$. Thus, one obtains

$$A_{\mathrm{n}} = \frac{4\Delta p_0 \sin(b_{\mathrm{n}} d)}{2b_{\mathrm{n}} d + \sin(2b_{\mathrm{n}} d)}\,. \tag{6.95}$$

From (6.91) and (6.95) one can derive a general solution for (6.85) that satisfies the boundary and initial conditions given by (6.86) through (6.89). Therefore, the general solution for $\Delta p(y, t)$ is

$$\Delta p(y, t) = 4\Delta p_0 \, \mathrm{e}^{-t/\tau_\mathrm{p}} \sum_{n=0}^{\infty} \left[\frac{\sin(b_\mathrm{n} d) \cos(b_\mathrm{n} y)}{[2b_\mathrm{n} d + \sin(2b_\mathrm{n} d)]} \right] \mathrm{e}^{-b_\mathrm{n}^2 D_\mathrm{p} t}. \tag{6.96}$$

Using (6.96), the transient photoconductivity can be expressed by

$$\begin{aligned} \Delta\sigma(t) &= q\mu_p(b+1) \int_0^d \Delta p(y, t) \, \mathrm{d}y \\ &= 4q\mu_p(b+1)\Delta p_0 \, \mathrm{e}^{-t/\tau_\mathrm{p}} \sum_{n=0}^{\infty} \left[\frac{\sin^2(b_\mathrm{n} d)}{b_\mathrm{n}[2b_\mathrm{n} d + \sin(2b_\mathrm{n} d)]} \right] \mathrm{e}^{-b_\mathrm{n}^2 D_\mathrm{p} t}, \end{aligned} \tag{6.97}$$

or

$$\Delta\sigma(t) = \sum_{m}^{\infty} C_m \, \mathrm{e}^{-t/\tau_m}, \tag{6.98}$$

where

$$C_m = \frac{4q\mu_p(b+1)\Delta p_0 \, \sin^2(b_m d)}{b_m[2b_m d + \sin(2b_m d)]} \tag{6.99}$$

and

$$\tau_m^{-1} = \tau_\mathrm{p}^{-1} + b_m^2 D_\mathrm{p}. \tag{6.100}$$

Equation (6.98) shows that the transient photoconductivity is represented by a summation of infinite terms, each of which has a characteristic amplitude C_m and decay time constant τ_m, where $m = 0, 1, 2, \ldots$.

Since b_0 (i.e., $m = 0$) is the zeroth-order mode and the smallest member of the set b_m, the time constant $\tau_0^{-1} = (\tau_\mathrm{p}^{-1} + b_0^2 D_\mathrm{p})$ must be larger than any other higher-order modes. The fact that the higher-order modes will die out much more quickly than the fundamental mode (i.e., $m = 0$) after the initial transient (i.e., for $t > 0$) implies that the decay time constant will be dominated by the zeroth-order mode. Therefore, the minority carrier lifetime can be determined from the photoconductivity-decay experiment using (6.98) for $m = 0$. Figure 6.9 shows a plot of $\Delta\sigma(t)$ versus t for a semiconductor specimen. From the slope of this photoconductivity-decay curve, one obtains the zeroth-order decay mode time constant, which is

$$\tau_0^{-1} = \tau_\mathrm{p}^{-1} + b_0^2 D_\mathrm{p}. \tag{6.101}$$

The first term on the right-hand side of (6.101) denotes the inverse bulk hole lifetime, while the second term represents the inverse surface recombination lifetime (to account for the effect of surface recombination). If the surface recombination velocity is small, then the photoconductivity-decay time constant is equal to the bulk lifetime. However, if the surface recombination term in (6.101) is much larger than the bulk lifetime term, then one can determine the surface recombination velocity from (6.101) by measuring the effective lifetimes of two samples with different thicknesses and similar surface treatment.

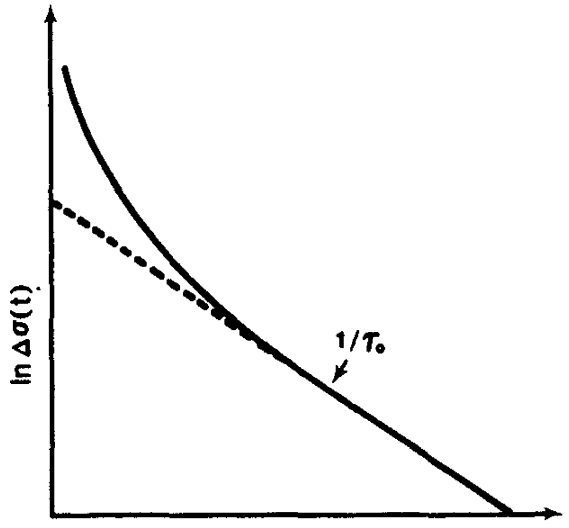


FIGURE 6.9. A typical photoconductivity-decay curve in a semiconductor.

Since the minority carrier lifetime is an important physical parameter for modeling the silicon devices and integrated circuits, it is important to determine the minority carrier lifetimes versus doping concentrations in silicon materials. Figures 6.10 and 6.11 show the measured minority carrier lifetimes as a function of doping concentrations in both n- and p-type silicon, as reported recently by Law et al.[5] The effective carrier lifetime is modeled using a concentration-dependent SRH lifetime τ_{srh} and a band-to-band Auger recombination lifetime τ_A to calculate the total effective lifetime by Mathiessen's rule, which is given by

$$\tau^{-1} = \tau_{srh}^{-1} + \tau_A^{-1}, \tag{6.102}$$

where

$$\tau_{srh} = \frac{\tau_0}{1 + N_1/N_{ref}} \tag{6.103}$$

and

$$\tau_A = \frac{1}{C_A N_I^2} \, . \tag{6.104}$$

Figure 6.10 shows the measured hole lifetimes as a function of the donor density for n-type silicon. The solid line is the best-fit curve using (6.102) through (6.104). The values of parameters used in fitting this curve are given by $\tau_0 = 10\,\mu s$, $N_{ref} = 10^{17}\,cm^{-3}$, and $C_A = 1.8 \times 10^{-31}\,cm^6/s$. Figure 6.11 shows the measured electron lifetimes as a function of the acceptor density for p-type silicon. The solid line is the best-fit curve using values of $\tau_0 = 30\,\mu s$, $N_{ref} = 10^{17}\,cm^{-3}$, and $C_A = 8.3 \times 10^{-32}$ cm^6/s.

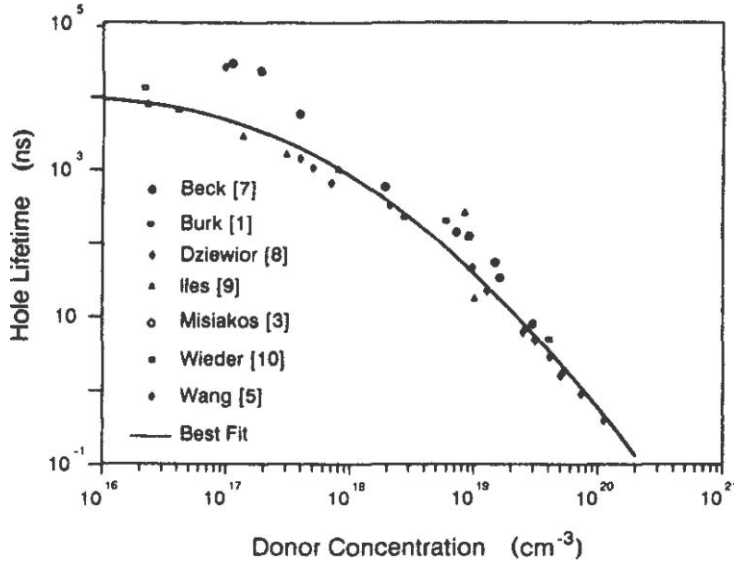


FIGURE 6.10. Measured and best-fit hole lifetimes versus donor concentrations in n-type silicon. After Law et al.,[5] by permission, © IEEE–1990.

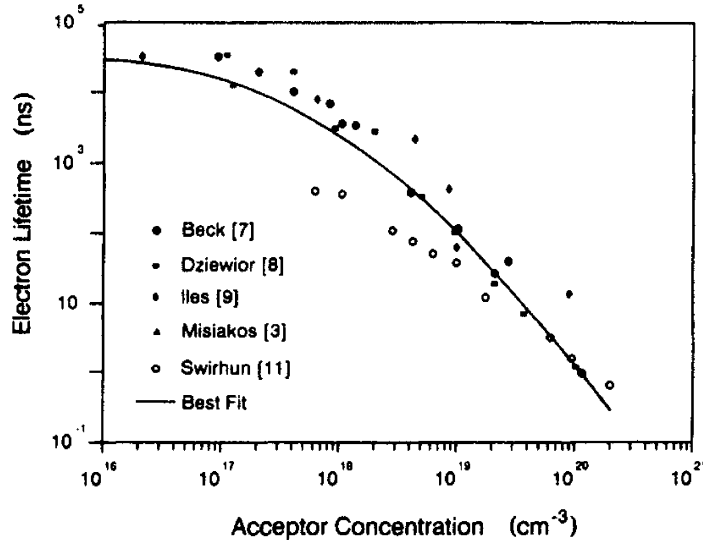


FIGURE 6.11. Measured and best-fit electron lifetimes versus acceptor concentrations in p-type silicon. After Law et al.,[5] by permission, ©IEEE–1990.

6.9. Surface States and Surface Recombination Velocity

It is well known that a thin natural oxide layer can be easily formed on a freshly cleaved or chemically polished semiconductor surface when it is exposed to air. As a result, an oxide–semiconductor interface usually exists at an unpassivated semiconductor surface. In general, as a result of a sudden termination of the periodic structure at the semiconductor surface and the lattice mismatch in the crystallographic structure at the semiconductor–oxide interface, defects are likely to form at the interface, which will create discrete or continuous energy states within the forbidden gap of the semiconductor. Figure 6.14 illustrates the energy band diagram for an oxide–semiconductor interface having surface states in the forbidden gap of the semiconductor.

In general, there are two types of surface states that are commonly observed in a semiconductor surface, namely, slow surface states and fast surface states. In a semiconductor surface, the density of slow states is usually much higher than the density of fast states. Furthermore, these surface states can be either positively or negatively charged. To maintain surface charge neutrality, the bulk semiconductor near the surface must supply an equal amount of opposite electric charges. As a result, the carrier density near the surface is different from that of the bulk semiconductor. Because of the slow surface states, the carrier densities at the semiconductor surface not only can change, but may vary so drastically that the surface conductivity type may convert to the opposite type of the bulk. In other words, if the bulk semiconductor is n-type with $n_0 \gg p_0$, then the hole density p_s at the semiconductor–oxide interface may become much larger than the electron density (i.e., $p_s \gg n_s$), so that the surface is inverted to p-type conduction. This is illustrated in Figure. 6.12a, in which an inversion layer is formed at the semiconductor surface. On the other hand, if the surface electron

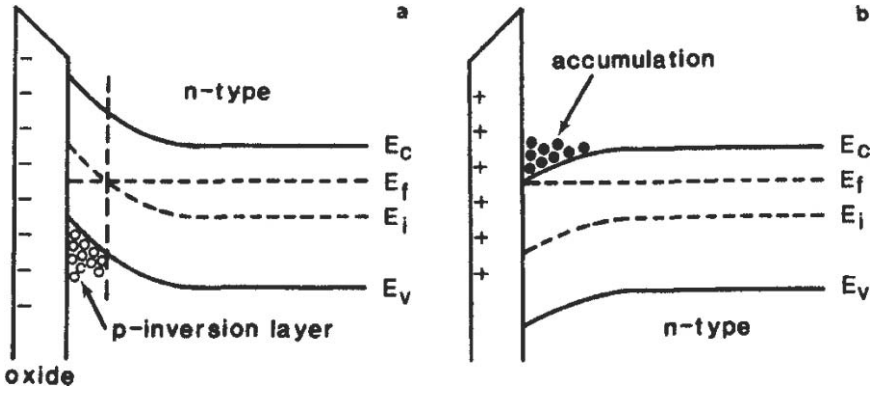


FIGURE 6.12. Potential barrier created at n-type semiconductor surface: (a) negatively charged slow states and the inversion layer, and (b) positively charged slow states and the accumulation layer.

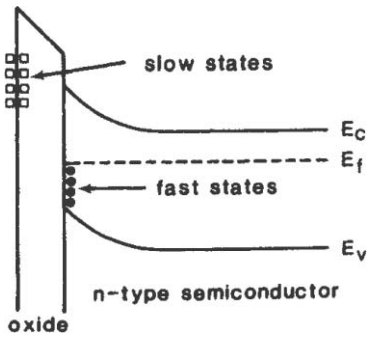


FIGURE 6.13. Energy band diagram of a semiconductor surface in the presence of a thin natural oxide layer and two types of surface states. □ denotes the slow surface states and ● denotes the fast surface states.

density is much greater than the surface hole density and bulk electron density (i.e., $n_s > p_s$ and $n_s > n_0$), then an accumulation layer is formed at the semiconductor surface, as shown in Figure 6.12b. Therefore, the slow surface states at the oxide–semiconductor interface play an important role in controlling the conductivity type of the semiconductor surface.

Figure 6.13 shows both the slow and fast surface states commonly observed in a semiconductor surface. The fast surface states are created either by termination of the periodic lattice structure in the bulk (i.e., creation of dangling bonds at the semiconductor surface) or by lattice mismatch and defects at the oxide–semiconductor interface. These surface states are in intimate electrical contact with the bulk semiconductor, and can reach a state of equilibrium with the bulk within a relatively short period of time (on the order of microseconds or less), and thus are referred to as fast surface states.

Another type of surface state, usually referred to as the slow state, exists inside the thin oxide layer near the oxide–semiconductor interface. This type of surface state may be formed by either chemisorbed ambient ions or defects in the oxide region (e.g., sodium ions or pinholes in the SiO_2 layer). Carriers transporting from such a state to the bulk semiconductor either have to overcome the potential barrier because of the large energy gap of the oxide or tunnel through the thin oxide layer. Such a charge transport process involves a large time constant, typically on the order of seconds or more, and hence these states are usually called slow states.

The concept of surface recombination velocity is discussed next. The SRH model derived earlier for dealing with nonradiative recombination in the bulk semiconductor may also be used to explain the recombination in a semiconductor surface. It is noted that a mechanically roughened surface, such as a sand-blasted surface, will have a very high surface recombination velocity, while a chemically etched surface will have a much lower surface recombination velocity. Undoubtedly, the fast surface states play an important role in controlling the recombination of excess carriers at the semiconductor surface. For example, GaAs has a very large surface state density and hence a very high surface recombination velocity, while an etched silicon surface has a much lower surface recombination velocity than that of GaAs.

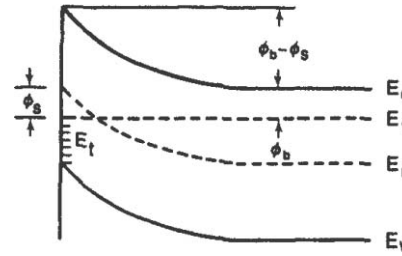


FIGURE 6.14. Energy band diagram of a semiconductor surface showing the fast surface states in the forbidden gap of the vacuum–semiconductor interface; ϕ_s denotes the surface potential and ϕ_b is the bulk potential.

Figure 6.14 shows the energy band diagram for an n-type semiconductor with fast states present at the surface. The energy level introduced by the fast surface states is designated as E_t, while ϕ_s and ϕ_b are the surface and bulk electrostatic potentials, respectively. The equilibrium electron and hole densities (n_s and p_s) at the surface can be expressed in terms of the bulk carrier densities, which are given by

$$n_s = n\, e^{-q(\phi_b-\phi_s)/k_B T}, \tag{6.105}$$

$$p_s = p\, e^{q(\phi_b-\phi_s)/k_B T}, \tag{6.106}$$

$$n_s p_s = np = (n_0 + \Delta n)(p_0 + \Delta p), \tag{6.107}$$

where n and p are the nonequilibrium electron and hole densities, and Δn and Δp are the excess electron and hole densities, respectively.

In (6.24), if n is replaced by the surface electron density n_s and p by the surface hole density p_s, the SRH model for the surface recombination rate is given by

$$U_s = \frac{N_{ts} c_n c_p (n_s p_s - n_i^2)}{c_p(p_s + p_1) + c_n(n_s + n_1)}, \tag{6.108}$$

where

$$n_1 = n_i\, e^{(E_t - E_i)/k_B T}, \tag{6.109}$$

$$p_1 = n_i\, e^{-(E_t - E_i)/k_B T}, \tag{6.110}$$

E_t is the energy level of the fast surface states, and N_{ts} the surface state density per unit area (cm^2). Thus, the surface recombination rate U_s has the dimensions of cm^{-2}/s. The surface recombination velocity s can be defined by

$$s = \frac{U_s}{\Delta n} = \frac{U_s}{\Delta p}, \tag{6.111}$$

where U_s is given by (6.108). For the small-injection case (i.e., $\Delta n \ll n_0$), the surface carrier densities n_s and p_s can be approximated by their respective equilibrium carrier densities p_{s0} and n_{s0}. Solving (6.105) and (6.106)

yields

$$n_s \approx n_{s0} = n_i \, \mathrm{e}^{q\phi_s/k_\mathrm{B}T}, \tag{6.112}$$

$$p_s \approx p_{s0} = n_i \, \mathrm{e}^{-q\phi_s/k_\mathrm{B}T}, \tag{6.113}$$

where $\phi_\mathrm{s} = (E_\mathrm{f} - E_\mathrm{is})/q$ is the surface potential, E_f is the Fermi energy, and E_is is the intrinsic Fermi level at the semiconductor surface. Solving (6.108) through (6.113) yields an expression for the surface recombination velocity:

$$s = U_\mathrm{s}/\Delta n = \frac{N_\mathrm{ts} c (p_0 + n_0)/2n_i}{\cosh[(E_\mathrm{t} - E_\mathrm{i} - q\phi_0)/k_\mathrm{B}T] + \cosh[q(\phi_\mathrm{s} - \phi_0)/k_\mathrm{B}T]}, \tag{6.114}$$

where

$$\phi_0 = (k_\mathrm{B}T/2q) \ln(c_\mathrm{p}/c_\mathrm{n}), \tag{6.115}$$

and

$$c = (c_\mathrm{p} c_\mathrm{n})^{1/2} \tag{6.116}$$

is the average rate of capture coefficient.

Equation (6.114) shows that the surface recombination velocity v_s is directly proportional to the surface state density N_ts and the rate of capture coefficient c. It also depends on the surface potential ϕ_s. As the ambient conditions at the surface change, the values of ϕ_s also change accordingly. This fact explains why a stable surface is essential for the operation of a semiconductor device. The surface recombination velocity is closely related to the surface state density. For example, a high surface state density (e.g., $N_\mathrm{ts} > 10^{13}$ cm^{-2}) in a GaAs crystal also leads to a high surface recombination velocity ($v_s > 10^6$ cm/s) in this material. For silicon crystal, the surface state density along the (100) surface can be smaller than 10^{10} cm^{-2} and higher than 10^{11} cm^{-2} along the (111) surface; as a result, the surface recombination velocity for a chemically polished silicon surface can be less than 10^3 cm/s. Therefore, careful preparation of the semiconductor surface is essential for achieving a stable and high-performance device.

6.10. Deep-Level Transient Spectroscopy Technique

As discussed earlier, deep-level defects play an important role in determining the minority carrier lifetimes in a semiconductor. Therefore, it is essential to develop a sensitive experimental tool for characterizing the deep-level defects in a semiconductor. The deep-level transient spectroscopy (DLTS) experiment, a high-frequency (1 MHz) transient capacitance technique, is the most sensitive technique for defect characterization in a semiconductor. For example, by performing the DLTS thermal scan from 77 K to around 450 K one can obtain the emission spectrum of all the deep-level traps (both majority and minority carrier traps) in the forbidden gap of a semiconductor as positive or negative peaks on a flat baseline. The DLTS technique offers advantages such as high sensitivity, ease of analysis,

and the capability of measuring traps over a wide range of depths in the forbidden gap. By properly changing the experimental conditions, we can measure defect parameters that include (1) minority and majority carrier traps, (2) activation energy of deep-level traps, (3) trap density and trap density profile, (4) electron and hole capture cross sections, and (5) type of potential well associated with each trap level. In addition, the electron and hole lifetimes can also be calculated from these measured defect parameters. Therefore, by carefully analyzing the DLTS data, all the defect parameters associated with the deep-level defects in a semiconductor can be determined. We shall next discuss the theoretical and experimental aspects of the DLTS technique.

The DLTS measurements can be performed using a variety of device structures such as Schottky barrier, p-n junction, and MOS structures. The DLTS technique is based on the transient capacitance change associated with the thermal emission of charge carriers from a trap level to thermal equilibrium after an initial nonequilibrium condition in the space-charge region (SCR) of a Schottky barrier diode or a p-n junction diode. The polarity of the DLTS peak depends on the capacitance change after trapping of the minority or majority carriers. For example, an increase in the trapped minority carriers in the junction SCR of a p-n diode would result in an increase in the junction capacitance of the diode. In general, a minority carrier trap will produce a positive DLTS peak, while a majority carrier trap would display a negative DLTS peak. For a p^+-n junction diode, the SCR extends mainly into the n-region, and the local charges are due to positively charged ionized donors. If a forward bias is applied, the minority carriers (i.e., holes) will be injected into this SRC region. Once the minority holes are trapped in a defect level, the net positive charges in the SCR will increase. This in turn will reduce the width of SCR and cause a positive capacitance change. Thus, the DLTS signal will have a positive peak. Similarly, if electrons are injected into the SCR and captured by the majority carrier traps, then the local charge density in the SCR is reduced and the depletion layer width is widened, which results in a decrease in the junction capacitance. Thus, the majority carrier trapping will result in a negative DLTS peak. The same argument can be applied to an n^+-p junction diode.

The peak height of a DLTS signal is directly related to the density of a trap level, which in turn is proportional to the change of junction capacitance $\Delta C(0)$ as a result of carrier emission from the trap level. Therefore, the defect density N_t can be calculated from the capacitance change $\Delta C(0)$ (or the DLTS peak height). If $C(t)$ denotes the transient capacitance across the depletion layer of a Schottky barrier diode or a p-n junction diode, then using abrupt junction approximation, one can write

$$C(t) = A\left[\frac{q\varepsilon_0\varepsilon_s\left(N_d - N_t e^{-t/\tau}\right)}{2(V_{bi} + V_R + k_B T/q)}\right]^{1/2} = C_0\left[1 - \left(\frac{N_t}{N_d}\right) e^{-t/\tau}\right], \tag{6.117}$$

where τ is the thermal emission time constant and $C_0 = C(V_R)$ is the junction capacitance measured at a quiescent reverse bias voltage V_R. If we use the binomial expansion in (6.117) and assume that $N_t/N_d \ll 1$, then $C(t)$ can be simplified to

$$C(t) \approx C_0 \left[1 - \left(\frac{N_t}{2N_d}\right) e^{-t/\tau}\right]. \tag{6.118}$$

At $t = 0$, one obtains

$$N_t \approx (2\Delta C(0)/C_0)N_d, \tag{6.119}$$

where $\Delta C(0) = C_0 - C(0)$ is the net capacitance change due to thermal emission of electrons from the trap level, and $C(0)$ is the capacitance measured at $t = 0$; $\Delta C(0)$ can be determined from the DLTS measurement. It is seen that both the junction capacitance C_0 and the background dopant density N_d are determined from the high-frequency C–V measurements. Therefore, the defect concentration N_t can be determined from (6.119) using DLTS and high-frequency (1 MHz) C–V measurements.

The decay time constant of the capacitance transient in the DLTS thermal scan is associated with a specific time constant, which is equal to the reciprocal of the emission rate. For a given electron trap, the emission rate e_n is related to the capture cross-section and the activation energy of the electron trap by

$$e_n = (\sigma_n \langle v_{th} \rangle N_c/g)\, e^{(E_c - E_t)/k_B T}, \tag{6.120}$$

where E_t is the activation energy of the electron trap, $\langle v_{th} \rangle$ is the average thermal velocity, N_c is the effective density of conduction band states, and g is the degeneracy factor. The electron capture cross-section σ_n, which depends on temperature, can be expressed by

$$\sigma_n = \sigma_0\, e^{-\Delta E_b/k_B T}, \tag{6.121}$$

where σ_0 is the capture cross-section when temperature approaches infinity, and ΔE_b is the activation energy of the capture cross-section. Now substituting σ_n given by (6.121) into (6.120) and using the fact that N_c is proportional to $T^{-3/2}$ and $\langle v_{th} \rangle$ is proportional to $T^{-1/2}$, the electron emission rate e_n given in Eq. (6.120) can be expressed by

$$e_n = BT^2\, e^{(E_c - E_t - \Delta E_b)/k_B T} = BT^2\, e^{(E_c - E_m)/k_B T}, \tag{6.122}$$

where B is a constant that is independent of temperature. From (6.122), it is seen that the electron thermal emission rate e_n is an exponential function of the temperature. The change of capacitance transient can be derived from (6.118), which yields

$$\Delta C(t) = C_0 - C(t) \approx C_0(N_t/2N_d)\, e^{-t/\tau} = \Delta C(0)\, e^{-t/\tau}, \tag{6.123}$$

where $\tau = e_n^{-1}$ is the reciprocal emission time constant.

The experimental procedures for determining the activation energy of a deep-level trap in a semiconductor are described as follows. The first step of the DLTS experiment is to choose the rate windows t_1 and t_2 in a dual-gated integrator of a boxcar averager, which is used in the DLTS system, and measure the capacitance change at a preset t_1 and t_2 rate window. This can be written as

$$\Delta C(t_1) = \Delta C(0)\,\mathrm{e}^{-t_1/\tau}, \tag{6.124}$$

$$\Delta C(t_2) = \Delta C(0)\,\mathrm{e}^{-t_2/\tau}. \tag{6.125}$$

The DLTS scan along the temperature axis is obtained by taking the difference of (6.124) and (6.125), which produces a DLTS spectrum given by

$$S(\tau) = \Delta C(0)(\mathrm{e}^{-t_1/\tau} - \mathrm{e}^{-t_2/\tau}). \tag{6.126}$$

The maximum emission rate $\tau_{\max}^{-1}$ can be obtained by differentiating $S(\tau)$ with respect to τ and setting $\mathrm{d}S(\tau)/\mathrm{d}\tau = 0$, which yields

$$\tau_{\max} = \frac{(t_1 - t_2)}{\ln(t_1/t_2)}. \tag{6.127}$$

Note that $S(\tau)$ reaches its maximum value at a characteristic temperature T_m corresponding to the maximum emission time constant $\tau_{\max}$. The emission rate is related to this $\tau_{\max}$ value by $e_{\mathrm{n}} = 1/\tau_{\max}$ for each t_1 and t_2 rate window setting. By changing the values of the rate window t_1 and t_2 in the boxcar-gated integrator, a series of DLTS scans with different values of e_{n} and T_m can be obtained. From these DLTS thermal scans we can obtain an Arrhenius plot of $e_{\mathrm{n}}T^2$ versus $1/T$ for a specific trap level, as shown in Figure 6.15.[6] The activation energy of the trap level can be calculated from the slope of this Arrhenius plot. Figure 6.16 shows the DLTS scans of electron and hole traps observed in a 290-keV proton-irradiated GaAs p-n junction diode.[6] Three electron traps and three hole traps were observed in this sample. Figure 6.17 shows the DLTS scans of a hole trap versus annealing time for a thermally annealed (170°C) Sn-doped InP grown by the liquid-encapsulated Czochralski (LEC) technique and the trap density versus annealing time for this sample.[7]

From the above description it is clearly shown that the DLTS technique is indeed a powerful tool for characterizing the deep-level defects in a semiconductor. It allows a quick inventory of all deep-level defects in a semiconductor and is widely used for defect characterization in semiconductors.

6.11. Surface Photovoltage Technique

Another characterization method, known as the surface photovoltage (SPV) technique, can be employed to measure the minority carrier diffusion length in a semiconductor wafer. The SPV method is a nondestructive technique since it is a steady-state, contactless optical technique. No junction preparation or high-temperature processing is needed for this method. The minority carrier lifetime can also be determined from the SPV measurements using the relation $\tau = L^2/D$,

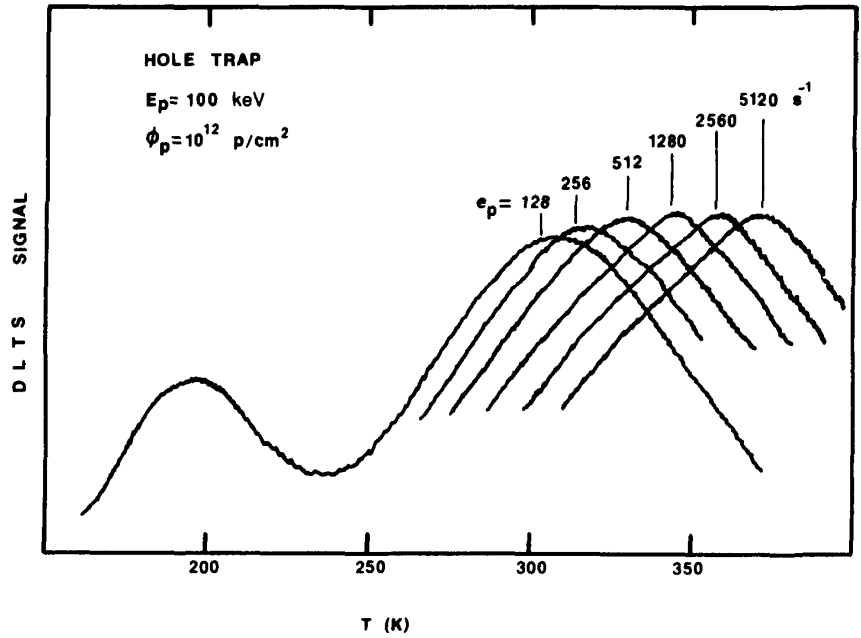


FIGURE 6.15. DLTS scans of the holes traps observed in a 100-keV proton-irradiated GaAs solar cell with a proton fluency of 10^{12} cm^{-2}. Six different DLTS scans were performed for the second hole trap observed at a higher temperature. After Li et al.,[6] by permission, © IEEE–1980.

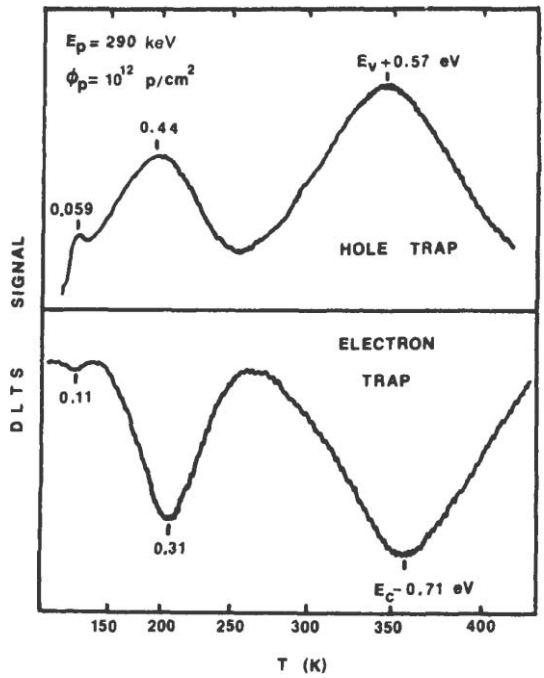


FIGURE 6.16. DLTS scans of electron and hole traps for a 290-keV and 10^{12} p/cm^2 proton irradiation AlGaAs/GaAs p-n junction solar cell. After Li et al.,[6] by permission, © IEEE–1980.

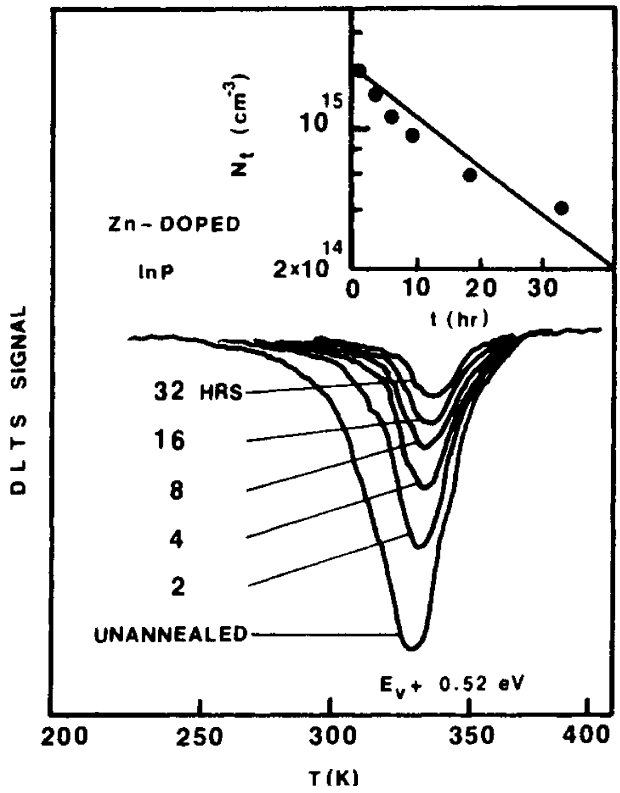


FIGURE 6.17. DLTS scans of a hole trap versus annealing time for a Zn-doped InP specimen annealed at 200°C. After Li et al.,[7] by permission.

where L is the minority carrier diffusion length and D is the diffusivity. The SPV technique has been widely used in determining the minority carrier diffusion length in silicon, GaAs, and InP materials. The basic theory and experimental details of the SPV method are depicted next.

When a semiconductor specimen is illuminated by chopped monochromatic light with its photon energy greater than the band gap energy of the semiconductor, an SPV is induced at the semiconductor surface as the photogenerated electron–hole pairs diffuse into the specimen along the direction of incident light. The SPV signal is capacitively coupled into a lock-in amplifier for amplification and measurement. The light intensity is adjusted to produce a constant SPV signal at different wavelengths of the incident monochromatic light. The light intensity required to produce a constant SPV signal is plotted as a function of the reciprocal absorption coefficient for each wavelength near the absorption edge. The resultant linear plot is extrapolated to zero light intensity and intercepts the horizontal axis at $-1/\alpha$, which is equal to the minority carrier diffusion length.

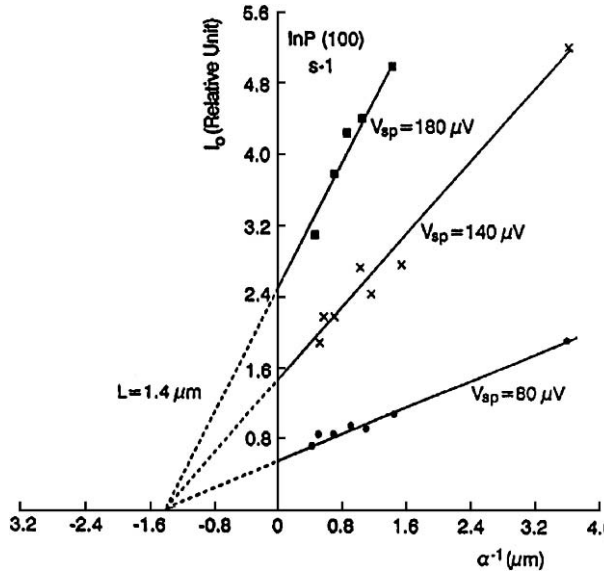


FIGURE 6.18. Relative light intensity I_0 versus inverse optical absorption coefficient α^{-1} for an InP specimen. After Li.[8]

The SPV signal developed at the illuminated surface of a semiconductor specimen is a function of the excess minority carrier density injected into the surface SCR. The excess carrier density is in turn dependent on the incident light intensity, the optical absorption coefficient, and the minority carrier diffusion length. Thus, an accurate knowledge of the absorption coefficient versus wavelength is required for the SPV method. In general, the SPV signal for an n-type semiconductor may be written as

$$V_{\text{SPV}} = f(\Delta p), \tag{6.128}$$

where

$$\Delta p = \frac{\eta I_0(1-R)}{(D_p/L_p + s_1)} \frac{\alpha L_p}{(1+\alpha L_p)} \tag{6.129}$$

is the excess hole density, η is the quantum efficiency, I_0 is the light intensity, R is the reflection coefficient, D_p is the hole diffusion coefficient, s_1 is the front surface recombination velocity, α is the optical absorption coefficient, and L_p is the hole diffusion length. Equation (6.128) holds if $\alpha^{-1} \gg L_p$, $n \gg \Delta p$, and $\alpha d > 1$ (where d is the thickness of the specimen).

If η and R are assumed constant over the measured wavelength range, the incident light intensity I_0 required to produce a constant SPV signal is directly proportional to the reciprocal absorption coefficient α^{-1} and can be written as

$$I_0 = C(\alpha^{-1} + L_p), \tag{6.130}$$

where C is a constant, independent of the photon wavelength. The linear plot of I_0 versus α^{-1} is extrapolated to zero light intensity and the negative intercept value is the effective hole diffusion length.

Figure 6.18 shows the relative photon intensity I_0 versus the inverse absorption coefficient α^{-1} for an n-type InP specimen.[8] The negative intercept yields $L_p = 1.4$ μm. The SPV measurements have been widely used in determining the minority carrier diffusion lengths in silicon wafers, with measured minority carrier diffusion lengths in the undoped silicon wafers greater than 100 μm.

Problems

6.1. Consider an n-type silicon sample with a dopant density of 2×10^{15} cm^{-3}. If the sample is illuminated by a mercury lamp with variable intensity, plot the excess carrier lifetimes as a function of the excess carrier density for Δn varying from 2×10^{13} to 5×10^{16} cm^{-3}. It is assumed that the recombination of excess carriers is dominated by the SRH process, $\tau_{n0} = \tau_{p0} = 1 \times 10^{-8}$ s, $n_0 \gg p_0$, and $n_0 \gg n_1$, p_1.

6.2. Consider a gold-doped silicon sample. There are two energy levels for the gold impurity in silicon. The gold acceptor level is located at 0.55 eV below the conduction band edge, and the gold donor level is 0.35 eV above the valence band edge. If the electron capture rate C_n for the gold acceptor center is assumed equal to 5×10^{-8} cm^3/s, the hole capture rate C_p is 2×10^{-8} cm^3/s, and the density of the gold acceptor center, N_Au, is equal to 5×10^{15} cm^{-3}:

(a) Compute the electron and hole lifetimes in this sample.

(b) If the temperature dependence of the electron emission rate is given by

$$e_\mathrm{n} = A_m(T/300)^m \exp[-(E_c - E_\mathrm{Au}^-)/k_\mathrm{B}T],$$

find a solution for e_n when $m = 0$ and 2.

(c) Calculate e_p from (a) and (b).

6.3. The kinetics of recombination, generation, and trapping at a single energy level inside the forbidden band gap of a semiconductor have been considered in detail by Shockley and Read.[1] From the appendix of this paper derive an expression for the excess carrier lifetime for the case that a large trap density is present in the semiconductor [see also reference[9]].

6.4. Plot the radiative lifetime for a GaAs sample as a function of excess carrier density Δn at $T = 300$ K, for $\Delta n/n_i = 0, 1, 3, 10, 30$ and $n_0/n_i = 10^{-2}, 10^{-1}, 1, 10, 10^2$. Here $n_i = 1 \times 10^7$ cm^{-3} is the intrinsic carrier density, and the generation rate G_r is assumed equal to 10^7 cm^{-3}/s.

6.5. Show that the Einstein relation (i.e., D_n/μ_n) for an n-type degenerate semiconductor is equal to $(k_\mathrm{B}T/q)F_{1/2}(\eta)/F_{-1/2}(\eta)$, where $F_{1/2}(\eta)$ is the Fermi integral of order one-half. Plot the D_n/μ_n versus dopant density N_D for n-type silicon at 300 K.

6.6. Derive an expression for the extrinsic Debye length for nondegenerate and degenerate semiconductors, and calculate the Debye lengths L_Dn for an n-type silicon sample with $N_\mathrm{D} = 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}$, and 10^{19} cm^{-3}.

6.7. Plot the energy band diagram for a p-type semiconductor surface under (a) inversion, (b) accumulation, and (c) depletion conditions.
6.8. Calculate the surface recombination velocity versus surface state density for an n-type silicon with $N_{ts} = 10^9$, 10^{10}, 10^{11}, and 10^{12} cm^{-2}. Assume that $E_T = E_c - 0.5$ eV, $c_n = c_p = 10^{-8}$ cm^3/s, $n_0 = 10^{16}$ cm^{-3}, $n_0 \gg p_0$, and $T = 300$ K.
6.9. From the paper "Fast Capacitance Transient Apparatus: Application to Zn- and O-centers in GaP p–n Junctions," in reference (10), describe the electron emission and capture processes in a Zn–O doped GaP p-n diode and their correlation to the DLTS thermal scan. Explain under what conditions the DLTS theory described in this paper fails.
6.10. Using the Arrhenius plot (i.e., e_p/T^2 versus $1/T$) find the activation energy of the second hole trap (located at a higher temperature) shown in Figure 6.15.

References

1. W. Shockley and W. T. Read, *Phys. Rev.* **87**, 835 (1952).
2. R. N. Hall, *Phys. Rev.* **87**, 387 (1952).
3. C. T. Sah and W. Shockley, *Phys. Rev.* **109**, 1103 (1958).
4. P. T. Landsberg and A. F. W. Willoughby (eds.), *Solid State Electron.* **21**, 1273 (1978).
5. M. E. Law, E. Solley, M. Liang, and D. E. Burk, *IEEE Elec. Dev. Lett.* **12**, 40 (1991).
6. S. S. Li, W. L. Wang, P. W. Lai, and R. Y. Loo, *IEEE Trans. Electron Devices*, **ED-27**, 857 (1980).
7. S. S. Li, W. L. Wang, and E. H. Shaban, *Solid State Commun*. **51**, 595 (1984).
8. S. S. Li, *Appl. Phys. Lett.* **29**, 126 (1976).
9. R. N. Hall, *Proc. IEEE* **106B**, 923 (1959).
10. D. V. Lang, *J. Appl. Phys.* **45**, 3014–3022 (1974).

Bibliography

J. S. Blakemore, *Semiconductor Statistics*, Pergamon Press, New York (1962).
R. H. Bube, *Photoconductivity of Solids*, Wiley, New York (1960).
P. T. Landsberg, *Solid State Physics in Electronics and Telecommunications*, Academic Press, London (1960).
A. Many and R. Bray, in: *Progress in Semiconductors*, Vol. 3, Heywood and Co., London (1958), pp. 117–151.
J. P. McKelvey, *Solid State and Semiconductor Physics*, 2nd ed., Chapter 10, Harper & Row, New York (1982).
D. A. Neamen, *Semiconductor Physics and Devices*, 3rd ed., McGraw-Hill, New York (2003).
R. F. Pierret, *Advanced Semiconductor Fundamentals*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ (2003).
W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, New York (1950).
R. A. Smith, *Semiconductors*, Cambridge University Press, London (1961).
M. Shur, *Physics of Semiconductor Devices*, Prentice Hall, New York (1990).
S. M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd ed. Wiley, New York (2002).

7
Transport Properties of Semiconductors

7.1. Introduction

In this chapter the carrier transport phenomena in a semiconductor under the influence of applied external fields are presented. Different galvanomagnetic, thermoelectric, and thermomagnetic effects created by the applied electric and magnetic fields as well as the temperature gradient in a semiconductor are discussed in this chapter. The transport coefficients associated with the galvanomagnetic, thermoelectric, and thermomagnetic effects in a semiconductor are derived from the Boltzmann transport equation using the relaxation time approximation. In the event that the relaxation time approximation fails, the solutions for the Boltzmann transport equation could be obtained using variational principles.

The effect of an applied electric field, magnetic field, or temperature gradient on the electrons in a semiconductor is to change the distribution function of electrons from its equilibrium condition. As discussed in Chapter 5, in the absence of external fields, the distribution of electrons in a semiconductor or a metal under equilibrium conditions may be described by the Fermi–Dirac distribution function, which is given by

$$f_0(E) = \frac{1}{1 + \mathrm{e}^{(E-E_\mathrm{f})/k_\mathrm{B}T}} \,. \tag{7.1}$$

Equation (7.1) shows that in thermal equilibrium the electron distribution function $f_0(E)$ depends not only on the electron energy but also on the Fermi energy E_f, a many-body parameter, and the temperature T. However, under the influence of external fields, $f_0(E)$ given in (7.1) may change from its equilibrium distribution function in a semiconductor. This can best be explained by considering the case in which an electric field or a magnetic field is applied to the semiconductor specimen. When an electric field or a magnetic field is applied to the semiconductor, the Lorentz force will tend to change the wave vector of electrons (i.e., $\boldsymbol{F} = -q(\mathcal{E} + \boldsymbol{v} \times \boldsymbol{B}) = \hbar\, \mathrm{d}\boldsymbol{k}/\mathrm{d}t$) along the direction of the applied fields. As a result, the distribution function is modified by the changing wave vector of electrons under the influence of Lorentz force. Furthermore, since f_0 depends on both the energy and temperature as well as the electron concentration, one expects that

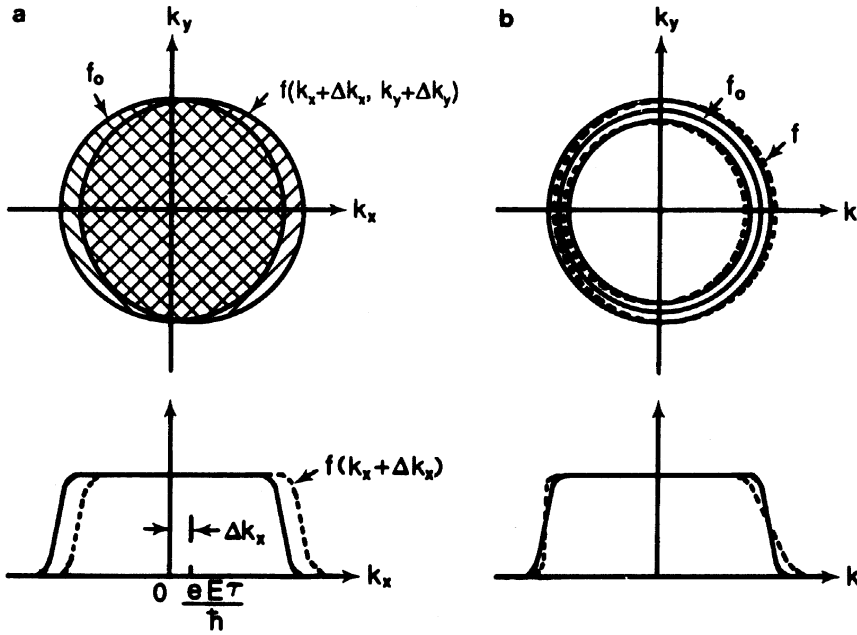


FIGURE 7.1. The effect of an applied electric field and a temperature gradient on the electron distribution function in a semiconductor.

the nonequilibrium distribution function of electrons will also be a function of the position in space when a temperature gradient or a concentration gradient is presented across the semiconductor specimen.

To illustrate the effect of external forces on the electron distribution function, Figures 7.1a and 7.1b show the two-dimensional (2-D) electron distribution functions in the presence of an applied electric field and a temperature gradient, respectively. As shown in Figure 7.1a, when an electric field is applied along the x-direction, the change of electron wave vector in the x-direction is given by $\Delta k_x = -q\mathcal{E}_x\tau/\hbar$, where τ is the mean relaxation time of electrons, and $\mathcal{E}_x$ is the applied electric field in the x-direction. In this case, the electron distribution function as a whole moves to the right by Δk_x from its equilibrium position. It is noted that the shape of the nonequilibrium distribution function remains unchanged from its equilibrium condition. The fact that the shape of the electron distribution function in k-space does not change because of the electric field can be explained by the force acting on each quantum state k. Since the Lorentz force ($F_x = -q\mathcal{E}_x$) due to the electric field is equal to $\hbar\dot{k}_x$, the rate of change of k_x is the same for all electrons. Consequently, if there is no relaxation mechanism to restore the distribution function to equilibrium, an applied electric field can cause the distribution function to drift, unaltered in shape, only along the k_x-direction at a constant velocity ($v_x = \hbar k_x/m^*$), and the change in crystal momentum as a result of this drift is given by $\Delta k_x = q\mathcal{E}_x\tau/\hbar$. On the other hand, the change in electron distribution function

is quite different when a temperature gradient is applied to a semiconductor. In this case, the nonequilibrium distribution function is shifted to the right by an amount equal to Δk_x for those electrons with energies greater than E_f, and shifted to the left by the same amount Δk_x for those electrons with energies less than E_f, as illustrated in Figure 7.1b. The physical mechanisms causing this shift can be explained using a Taylor series expansion of $(E - E_f)$ about the Fermi energy $(E_f = \hbar^2 k_f^2 / 2m^*)$. Assuming that $|E - E_f| \ll E_f$, one can replace $(E - E_f)$ by $(\hbar^2 k_f / m^*)(k - k_f)$ and obtain

$$\Delta k_x = \frac{\tau \hbar k_f}{m^* T}(k - k_f)\frac{\partial T}{\partial x} \ . \tag{7.2}$$

This result shows that the distribution of quantum states at the Fermi surface (i.e., at $E = E_f$) is not affected by the temperature gradient (i.e., $\Delta k_x = 0$). From (7.2) it is noted that for those quantum states with energies greater than E_f (i.e., $E > E_f$) their centers shift in the same direction as the temperature gradient (Δk_x is positive), whereas for those quantum states with energies smaller than E_f (i.e., $E < E_f$), their centers move in the opposite direction to the temperature gradient (Δk_x is negative).

Section 7.2 describes various galvanomagnetic, thermoelectric, and thermomagnetic effects in a semiconductor. These include electrical conductivity, the Hall effect, the Seebeck and Pelter effects, the Nernst and Ettinghousen effects, and the magnetoresistance effect. In Section 7.3, the Boltzmann transport equation for the steady-state case is derived. Expressions for the electrical conductivity, electron mobility, Hall coefficient, magnetoresistance, and Nernst and Seebeck coefficients for n-type semiconductors are derived in Section 7.4. Transport coefficients for the mixed conduction case are considered in Section 7.5. Section 7.6 presents some experimental results on the transport coefficients for germanium, silicon, and III-V compound semiconductors.

7.2. Galvanomagnetic, Thermoelectric, and Thermomagnetic Effects

In this section, the galvanomagnetic, thermoelectric, and thermomagnetic effects in a semiconductor are discussed. These effects are created by the transport of electrons (for n-type) or holes (for p-type) in a semiconductor when an external electric field, a magnetic field, or a temperature gradient is applied separately or simultaneously to a semiconductor specimen. The transport coefficients to be described here include electrical conductivity, thermal conductivity, Hall coefficient, Seebeck coefficient, Nernst coefficient, and the magnetoresistance of an n-type semiconductor. Transport coefficients derived for an n-type semiconductor can also be applied to a p-type semiconductor, provided that the positive charge and positive effective mass of holes are used instead. It is noted that for nondegenerate semiconductors, Maxwell–Boltzmann (M-B) statistics are used in the derivation of

transport coefficients, while Fermi–Dirac (F-D) statistics are used in the derivation of transport coefficients for degenerate semiconductors and metals.

7.2.1. Electrical Conductivity

In this section the current conduction due to electrons in an n-type semiconductor is described. When a small electric field is applied to the specimen, the electrical current density can be related to the electric field using Ohm's law, which reads

$$J_{\mathrm{n}} = \sigma_{\mathrm{n}}\mathcal{E} = q\mu_{\mathrm{n}}n\mathcal{E} \ , \tag{7.3}$$

where $\sigma_{\mathrm{n}} = q\mu_{\mathrm{n}}n$ is the electrical conductivity, μ_{n} is the electron mobility, and n denotes the electron density.

The electrical current density can also be expressed in terms of the electron density and electron drift velocity v_{d} along the direction of the applied electric field by

$$J_{\mathrm{n}} = qnv_{\mathrm{d}} \ , \tag{7.4}$$

where q is the electronic charge. Comparing (7.3) and (7.4), one finds that the electron drift velocity is related to the electric field by

$$v_{\mathrm{d}} = \mu_{\mathrm{n}}\mathcal{E} \ , \tag{7.5}$$

where μ_{n} is the low-field electron drift mobility, which is defined as the electron drift velocity per unit electric field strength. For metals, μ_{n} can be expressed in terms of the mean collision time τ and the electron effective mass m^*:

$$\mu_{\mathrm{n}} = \frac{q\tau}{m^*} \ . \tag{7.6}$$

From (7.3) and (7.6), the electrical conductivity σ_{n} for a metal can be expressed in terms of the mean collision time and the electron effective mass as

$$\sigma_{\mathrm{n}} = \frac{q^2n\tau}{m^*} \ . \tag{7.7}$$

In the collision processes, the transition probability for electron collision is directly related to the density of collision centers, and the collision rate is inversely proportional to the collision time constant. For example, in the case of electron–phonon scattering, the number of scattering centers is equal to the phonon population in thermal equilibrium. At high temperatures the average phonon density is proportional to temperature. Consequently, at high temperatures, the collision time τ varies as $1/T$, and hence the electrical conductivity σ_{n} varies inversely with temperature T. This prediction is consistent with the observed temperature dependence of electrical conductivity in a metal.

Equations (7.6) and (7.7) can also be applied to n-type semiconductors, provided that the free-electron mass is replaced by the conductivity effective mass of electrons in the conduction bands m_{n}^*, and τ is replaced by the average relaxation time $\langle\tau\rangle$. In general, the electron density in a semiconductor is a strong function

of temperature, and the relaxation time may depend on both the energy and temperature. A general expression for the current density in an n-type semiconductor can be derived as follows. From (5.3), the density of quantum states $g_n(E)$ for a single-valley semiconductor with a parabolic conduction band can be written as

$$g_n(E) = \left(\frac{4\pi}{h^3}\right)(2m_n^*)^{3/2}E^{1/2}. \tag{7.8}$$

Using (7.4) and (7.8), a general expression for the electron current density can be expressed by

$$J_n = -qnv_x = -q\int_0^\infty v_x f(E)g_n(E)\,dE, \tag{7.9}$$

where $f(E)$ is the nonequilibrium electron distribution function, which can be obtained by solving the Boltzmann transport equation to be described in Section 7.4. The integration of (7.9) is carried out over the entire conduction band. The minus sign in (7.9) stands for electron conduction in an n-type semiconductor. For hole conduction in a p-type semiconductor, a plus sign should be used instead. Figure 7.2a shows the applied electric field and the current flow in an n-type

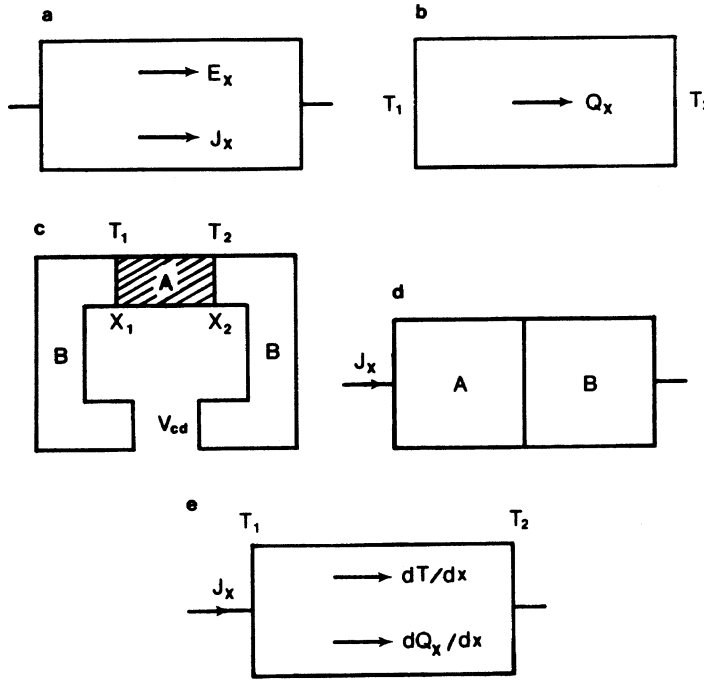


FIGURE 7.2. Longitudinal transport effects in the presence of an electric field or a temperature gradient: (a) electrical conductivity; (b) electronic thermal conductivity; (c) Seebeck effect, $S_{ab} = V_{dc}/(T_2 - T_1)$; (d) Peltier effect, $\Pi_{ab} = S_{ab}T$; (e) Thomson effect.

semiconductor specimen. The electrical conductivity for an n-type semiconductor is derived in Section 7.4.

7.2.2. Electronic Thermal Conductivity

Electronic thermal conductivity is due to the flow of thermal energy carried by electrons when a temperature gradient is applied across a semiconductor specimen. As shown in Figure 7.2b, when a temperature gradient is created in a semiconductor specimen, a heat flux flow will appear across the specimen. The electronic thermal conductivity K_n is defined as the thermal flux density per unit temperature gradient, and can be expressed by

$$K_n = -\left.\frac{Q_x}{(\partial T/\partial x)}\right|_{J_x=0}, \tag{7.10}$$

where Q_x is the thermal flux density given by

$$Q_x = nv_x E = \int_0^\infty v_x E f(E) g_n(E)\,dE. \tag{7.11}$$

Note that the integration on the right-hand side of (7.11) is carried out over the entire conduction band. Equations (7.9) and (7.11) are the two basic equations that describe the flow of electric current density and heat flux density in an n-type semiconductor, respectively. All the transport coefficients described in this section can be derived from (7.9) and (7.11), provided that the nonequilibrium distribution function $f(k, r)$ is known. The steady-state nonequilibrium distribution $f(k, r)$ can be derived by solving the Boltzmann transport equation. It is noted that in thermal equilibrium, both J_n and Q_x, given by (7.9) and (7.11), are equal to zero, and $f(k, r)$ reduces to the equilibrium Fermi distribution function $f_0(E)$. Figures 7.2 and 7.3 show the plots of various galvanomagnetic, thermoelectric, and thermomagnetic effects in a semiconductor in the presence of the electric field, current density, heat flux, and temperature gradient. Various galvanomagnetic, thermoelectric, and thermomagnetic effects as well as the transport coefficients associated with the applied electric field, magnetic field, and the temperature gradient in a semiconductor are discussed next.

7.2.3. Thermoelectric Coefficients

When a temperature gradient, an electric field, or both are applied across a semiconductor or a metal specimen, three different kinds of thermoelectric effects can be observed. They are the Seebeck, Peltier, and Thomson effects. The thermoelectric coefficients associated with each of these effects can be defined according to Figures 7.2c–e, in which two pieces of conductors (A and B) are joined at junctions x_1 and x_2. If a temperature difference ΔT is established between junctions x_1 and x_2, then an open-circuit voltage V_{cd} is developed between terminals "c" and "d." This is known as the Seebeck effect. In this case, the differential Seebeck

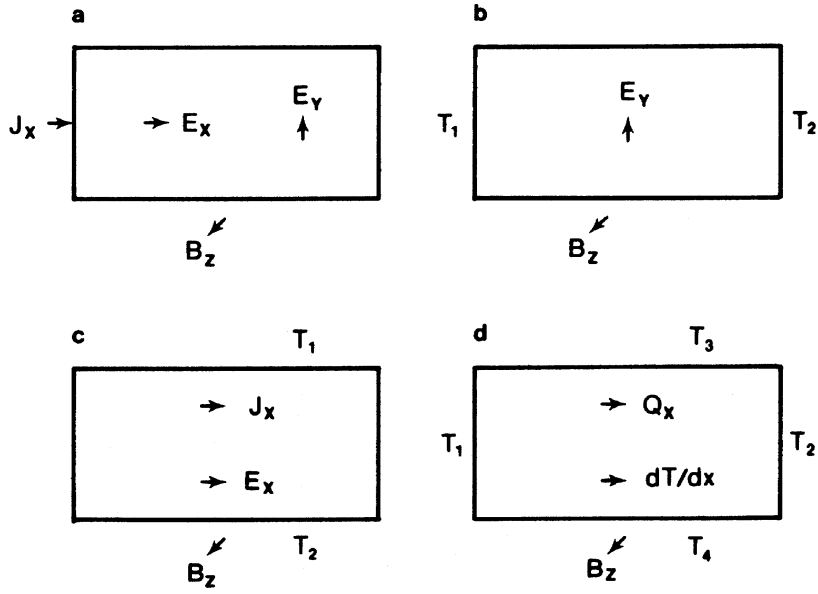


FIGURE 7.3. Galvanomagnetic and thermomagnetic effects in a semiconductor under the influence of an applied electric field, a magnetic field, and a temperature gradient. The polarity shown is for hole conduction: (a) Hall effect, (b) Nernst effect, (c) Ettingshausen effect, (d) Righi–Leduc effect.

coefficient, or the thermoelectric power, can be defined by

$$S_{\text{ab}} = \frac{V}{\Delta T} = \frac{V_{\text{cd}}}{(T_2 - T_1)} . \tag{7.12}$$

If junctions x_1 and x_2 are initially maintained at the same temperature, then by applying a voltage across terminals "c" and "d" one observes an electrical current flow through these two conductors. If the result is a rate of heating at junction x_1, then there will be a cooling at the same rate at junction x_2. This is the well-known Peltier effect, the basic principle of thermoelectric cooling. The differential Peltier coefficient is defined by

$$\Pi_{ab} = \frac{Q_x}{I_x} . \tag{7.13}$$

The Thomson effect occurs when an electric current and a temperature gradient are applied simultaneously in the same direction on a semiconductor specimen. In this case, the simultaneous presence of the current flow I_x and the temperature gradient $\partial T/\partial x$ in the x-direction will produce a rate of heating or cooling $(\partial Q_x/\partial x)$ per unit length. Thus, the Thomson coefficient can be expressed by

$$\tau = \frac{(\partial Q_x/\partial x)}{I_x(\partial T/\partial x)} . \tag{7.14}$$

Applying thermodynamic principles to the thermoelectric effects shown in Figures 7.2c–e, Thomson derived two important equations, later known as the Kelvin relations, that relate the three thermoelectric coefficients. The Kelvin relations can be expressed by

$$\Pi_{ab} = S_{\mathrm{ab}} T, \tag{7.15}$$

$$\tau_{\mathrm{a}} - \tau_{\mathrm{b}} = T \frac{\mathrm{d}S_{\mathrm{ab}}}{\mathrm{d}T}. \tag{7.16}$$

Equations (7.15) and (7.16) not only have a sound theoretical basis, but also have been verified experimentally by the measured thermoelectric figure of merit. Equation (7.15) is particularly useful for thermoelectric refrigeration applications. This is due to the fact that the rate of cooling by means of the Peltier effect can be expressed in terms of the Seebeck coefficient, which is a much easier quantity to measure. Equation (7.16) allows one to account for the influence of the Thomson effect on the cooling power of a thermoelectric refrigerator via the variation of Seebeck coefficient with temperature.

Since the Thomson coefficient is defined for a single conductor, it is appropriate to introduce here the absolute Seebeck and Peltier coefficients for a single conductor. The differential Seebeck and Peltier coefficients between the two conductors are given by $(S_{\mathrm{a}} - S_{\mathrm{b}})$ and $(\Pi_{\mathrm{a}} - \Pi_{\mathrm{b}})$, respectively, where S_{a} and S_{b} are the absolute Seebeck coefficients for conductors A and B, and Π_{a} and Π_{b} denote the absolute Peltier coefficients for conductors A and B, respectively. The Kelvin relations for a single conductor given by (7.15) and (7.16) can be expressed as

$$\Pi = ST \ , \tag{7.17}$$

$$\tau_{\mathrm{s}} = T \frac{\mathrm{d}S}{\mathrm{d}T} \ , \tag{7.18}$$

where Π, S, and τ_{s} denote the absolute Peltier, Seebeck, and Thomson coefficients for a single conductor or a semiconductor, respectively. Therefore, using the definitions of thermoelectric coefficients described in this section, the general expressions of thermoelectric coefficients can be derived from the Boltzmann transport equation to be discussed in Section 7.4.

7.2.4. Galvanomagnetic and Thermomagnetic Coefficients

When a magnetic field is applied to a semiconductor specimen in addition to an electric field and/or a temperature gradient, the transport phenomena become much more complicated than those without a magnetic field. Fortunately, most of the important galvanomagnetic effects in a semiconductor associated with the applied magnetic fields are focused on the cases in which a magnetic field is applied in a direction perpendicular to the electric field or the temperature gradient. These are usually referred to as the transverse galvanomagnetic effects, which include the Hall, Nernst, and magnetoresistance effects. In this section, the transverse galvanomagnetic effects in an n-type semiconductor are described.

The Hall effect is the best-known galvanomagnetic effect found in a semiconductor. As shown in Figure 7.3a, when a magnetic field is applied in the z-direction and an electric field is applied in the x-direction, an electric field (known as the Hall field) will be developed in the y-direction of the specimen. The Hall coefficient R_{H}, under isothermal conditions, can be defined by

$$R_{\mathrm{H}} = \left.\frac{\mathcal{E}_y}{J_x B_z}\right|_{J_y=0}, \tag{7.19}$$

where $\mathcal{E}_y$ is the Hall field induced in the y-direction, J_x is the electric current density flow in the x-direction, and B_z is the applied magnetic field in the z-direction.

Figure 7.3a shows a schematic diagram of the Hall effect across a semiconductor specimen under isothermal conditions. Note that both the electric current density in the direction of the Hall field (i.e., the y-direction) and the temperature gradient across the specimen are assumed equal to zero. The polarity of the Hall voltage depends on the type of charge carriers (i.e., electrons or holes) in the specimen. This is due to the fact that electrons and holes in a semiconductor will experience an opposite Lorentz force when they are moving in the same direction of the specimen. Therefore, the polarity of the Hall voltage will be different for an n-type and a p-type semiconductor, and the Hall effect measurement is often used to determine the conduction types (i.e., n- or p-type) and the majority carrier concentration in a semiconductor.

If a temperature gradient is applied in the x-direction and a magnetic field in the z-direction, then a transverse electric field will be developed in the y-direction of the specimen, as illustrated in Figure 7.3b. This effect is known as the Nernst effect. The Nernst coefficient, which is thermodynamically related to the Ettingshausen coefficient in the same way as the Seebeck coefficient is related to the Peltier coefficient, is defined by

$$Q_{\mathrm{n}} = \left.\frac{\mathcal{E}_y}{B_z(\partial T/\partial x)}\right|_{J_x=J_y=0}, \tag{7.20}$$

where $\mathcal{E}_y$ is the Nernst field developed in the y-direction when a temperature gradient is applied in the x-direction and a magnetic field is in the z-direction. Note that the electric current density along the x- and y-directions and the temperature gradient in the y-direction (i.e., $\partial T/\partial y = 0$) are assumed equal to zero.

As is shown in Figure 7.3c, if an electric field is applied in the x-direction and a magnetic field in the z-direction, then a temperature gradient will be developed in the y-direction of the specimen. This is known as the Ettingshausen effect. It is this effect that forms the basis of thermomagnetic cooling as a counterpart to thermoelectric cooling by the Peltier effect discussed earlier. The Ettingshausen coefficient P_{E} is defined by

$$P_{\mathrm{E}} = \left.\frac{(\partial T/\partial y)}{J_x B_z}\right|_{J_y=\partial T/\partial x=0}, \tag{7.21}$$

where $\partial T/\partial y$ is the temperature gradient developed in the y-direction of the specimen. Note that in defining the Ettingshausen effect, the current density in the y-direction and the temperature gradient in the x-direction are assumed equal to zero.

The Ettingshausen coefficient P_{E} and the Nernst coefficient Q_{n} are related by the Bridgeman equation, which is given by

$$P_{\mathrm{E}} K_{\mathrm{n}} = Q_{\mathrm{n}} T \ , \tag{7.22}$$

where K_{n} is the electronic thermal conductivity defined by (7.11).

The Righi–Leduc effect refers to the creation of a transverse temperature gradient $\partial T/\partial y$ when a temperature gradient $\partial T/\partial x$ in the x-direction and a magnetic field B_z in the z-direction are applied simultaneously to a semiconductor specimen. The Righi–Leduc coefficient R_{L} can be expressed by

$$R_{\mathrm{L}} = \left. \frac{(\partial T/\partial y)}{(\partial T/\partial x) B_z} \right|_{J_x = J_y = 0} \ . \tag{7.23}$$

Note that the current densities J_x and J_y in the x- and y-directions are assumed equal to zero in (7.23).

In order to derive general expressions for the different transport coefficients described above, the nonequilibrium distribution function $f(E)$ in (7.9) and (7.11) will first be solved from the Boltzmann transport equation using relaxation time approximation, which will be discussed next.

7.3. Boltzmann Transport Equation

An analytical expression for the Boltzmann transport equation can be derived for an n-type semiconductor using the relaxation time approximation. The relaxation time approximation assumes that all the collision processes are elastic and can be treated in terms of a unique relaxation time. Elastic scattering requires that the change of electron energy during the scattering process must be small compared to the energy of electrons, and the relaxation time is a scalar quantity. Typical examples of elastic scattering processes include the scattering of electrons by the longitudinal acoustical phonons, ionized impurities, and neutral impurities in a semiconductor. It is noted that the relaxation time τ may be a function of temperature and energy, depending on the types of scattering mechanisms involved, as will be discussed later, in Chapter 8. The transport coefficients for an n-type semiconductor to be derived in this section include electrical conductivity, the Hall coefficient, the Seebeck coefficient, the Nernst coefficient, and magnetoresistance.

According to Liouville's theorem, if $f(k, r, t)$ denotes the nonequilibrium distribution function of electrons at time t, in a volume element of $\mathrm{d}^3 r\, \mathrm{d}^3 k$, and located at (r, k) in r- and k-space, then $f(k + \dot{k}\, \mathrm{d}t, r + \dot{r}\, \mathrm{d}t, t + \mathrm{d}t)$ represents the distribution function at time $(t + \mathrm{d}t)$ within the same volume element. The difference between $f(k, r, t)$ and $f(k + \dot{k}\, \mathrm{d}t, r + \dot{r}\, \mathrm{d}t, t + \mathrm{d}t)$ must be balanced by the collision processes that occur inside a semiconductor or a metal. Therefore, the total rate of change of the distribution function with respect to time in the presence of

a Lorentz force or a temperature gradient can be written as

$$\frac{\mathrm{d}f}{\mathrm{d}t} = (-\dot{k} \cdot \nabla_k f - \dot{r} \cdot \nabla_r f) + \left.\frac{\partial f}{\partial t}\right|_{\mathrm{c}} + \frac{\partial f}{\partial t}\ , \tag{7.24}$$

where $\dot{k} = \mathrm{d}k/\mathrm{d}t$ and $\dot{r} = \mathrm{d}r/\mathrm{d}t = v$ denote the change of crystal momentum and the electron velocity, respectively. The two terms inside the parentheses on the right-hand side of (7.24) represent the external force terms due to the Lorenz force and temperature gradient; the next term is the internal collision term, which tends to offset the external force terms; and the last term is the time-dependent term that exists only for the transient case. Equation (7.24) is the generalized Boltzmann transport equation.

In this section, only the steady-state case will be considered. The transport coefficients are derived when the time-independent external forces are applied to a semiconductor specimen. In this case, the third term on the right-hand side of (7.24) is set equal to zero, and the Boltzmann equation given by (7.24) becomes

$$\dot{k} \cdot \nabla_k f + \dot{r} \cdot \nabla_r f = \left.\frac{\partial f}{\partial t}\right|_{\mathrm{c}}\ . \tag{7.25}$$

In general, the nonequilibrium distribution function $f(k, r)$ can be obtained from solving (7.25), and the transport coefficients of a semiconductor or a metal can be derived once $f(k, r)$ is found from (7.25).

In order to obtain an analytical expression for $f(k, r)$ from (7.25), it is necessary to assume that the scattering of charge carriers in a semiconductor is elastic so that the relaxation time approximation can be applied to the Boltzmann equation. According to the classical model, electron velocity is accelerated by the applied electric field over a period of time inside the crystal, while its drift velocity drops to zero through the internal collision process. It is, however, more appropriate to consider the way in which the electron system is relaxed toward its equilibrium distribution once the external perturbation is removed. Therefore, if $f(k, r)$ represents the distribution function of electrons under the influence of an applied electric field and $f_0(E)$ is the thermal equilibrium distribution function, then the collision term given by (7.25) can be expressed in terms of the relaxation time τ as

$$\left.\frac{\partial f}{\partial t}\right|_{\mathrm{c}} = -\frac{f - f_0}{\tau}. \tag{7.26}$$

Equation (7.26) is the basis of the relaxation time approximation in which the collision term on the right-hand side of (7.25) is replaced by the difference in the nonequilibrium and equilibrium distribution functions divided by the relaxation time constant τ. The relaxation time constant is usually dependent on the types of scattering mechanisms in a semiconductor. If the external forces are removed, then the nonequilibrium distribution function will decay exponentially to its equilibrium value with a time constant τ governed by the internal scattering processes.

For n-type semiconductors, the relaxation time τ depends on the energy of electrons according to the simple power law

$$\tau = aE^s \ , \tag{7.27}$$

where s is a constant whose value depends on the types of scattering mechanisms involved. The constant a may or may not be a function of temperature, depending on the types of scattering mechanisms. For example, in a semiconductor in which the ionized impurity scattering is dominated, s is equal to 3/2 and a is independent of temperature, while for acoustical phonon scattering, s is equal to $-1/2$ and a varies inversely with temperature. For neutral impurity scattering, τ is independent of the electron energy, and $s = 0$.

7.4. Derivation of Transport Coefficients for n-Type Semiconductors

In this section, the transport coefficients for n-type semiconductors are derived for the cases in which an electric field, a magnetic field, or a temperature gradient is applied to the specimen. Transport coefficients such as electrical conductivity, Hall coefficient, Seebeck and Nernst coefficients, and magnetoresistance can be derived from (7.24) using the relaxation time approximation.

Derivation of the nonequilibrium distribution function from the Boltzmann equation for an n-type semiconductor is described first. The Lorentz forces acting on the electrons because of the presence of an electric field and a magnetic field can be expressed by

$$\boldsymbol{F} = -q(\mathcal{E} + \boldsymbol{v} \times \boldsymbol{B}) = \hbar \dot{\boldsymbol{k}} \quad \text{or} \quad \dot{\boldsymbol{k}} = -\frac{q}{\hbar}(\mathcal{E} + \boldsymbol{v} \times \boldsymbol{B}) \ . \tag{7.28}$$

Now substituting $\dot{k}$ given by (7.28) into (7.25), the first term on the left-hand side of (7.25) becomes

$$-\dot{\boldsymbol{k}} \cdot \nabla_k f = \left(\frac{q}{\hbar}\right)(\mathcal{E} + \boldsymbol{v} \times \boldsymbol{B}) \cdot \nabla_k f \ . \tag{7.29}$$

The second term on the left-hand side of (7.25) is due to the presence of a temperature gradient or a concentration gradient in a semiconductor. Using the relaxation time approximation, the collision term is given by (7.26). Now, substituting (7.29) and (7.26) into (7.25), one obtains

$$\left(\frac{q}{\hbar}\right)(\mathcal{E} + \boldsymbol{v} \times \boldsymbol{B}) \cdot \nabla_k f - \boldsymbol{v} \cdot \nabla_r f = \frac{f - f_0}{\tau}, \tag{7.30}$$

or

$$\left(\frac{q}{m_{\mathrm{n}}^*}\right)(\mathcal{E} + \boldsymbol{v} \times \boldsymbol{B}) \cdot \nabla_v f - \boldsymbol{v} \cdot \nabla_r f = \frac{f - f_0}{\tau}. \tag{7.31}$$

Equation (7.31) is the generalized steady-state Boltzmann equation, which is obtained using de Broglie's wave–particle duality relation ($\hbar k = m_{\mathrm{n}}^* v$) in (7.30). To obtain an analytical solution for $f(k, r)$ from (7.31), certain approximations must

be used. Since the equilibrium distribution function f_0 depends only on energy and temperature, the nonequilibrium distribution function $f(k, r)$ must contain terms that depend only on the electron velocity and energy. Therefore, it is appropriate to write a generalized trial solution for (7.31) in terms of the equilibrium distribution function and a first-order correction term, which contains both the energy and velocity components. This is given by

$$f = f_0 - \boldsymbol{v} \cdot \boldsymbol{P}(E)\frac{\partial f_0}{\partial E} , \tag{7.32}$$

where $\boldsymbol{P}(E)$ is an unknown vector quantity, which depends only on the electron energy. For the small-perturbation case [i.e., $(f - f_0) \ll f_0$], each term in (7.31) can be approximated by

$$\boldsymbol{v} \cdot \nabla_r f \approx \boldsymbol{v}.\nabla_r f_0 = \boldsymbol{v} \cdot \nabla_r T \left[\frac{(E_\mathrm{f} - E)}{T}\frac{\partial f_0}{\partial E}\right] , \tag{7.33}$$

$$\mathcal{E} \cdot \nabla_v f \approx \mathcal{E} \cdot \nabla_v f_0 = \mathcal{E} \cdot (\nabla_v E)\frac{\partial f_0}{\partial E} = \mathcal{E} \cdot (m^* \boldsymbol{v})\frac{\partial f_0}{\partial E} , \tag{7.34}$$

$$(\boldsymbol{v} \times \boldsymbol{B}) \cdot \nabla_v f \approx -\boldsymbol{v} \cdot [\boldsymbol{B} \times \boldsymbol{P}(E)]\frac{\partial f_0}{\partial E} . \tag{7.35}$$

Now, substituting (7.33), (7.34), and (7.35) into (7.31), one obtains

$$-q\tau\boldsymbol{v} \cdot \mathcal{E} + \left(\frac{q\tau}{m_\mathrm{n}^*}\right)[\boldsymbol{B} \times \boldsymbol{P}(E)] \cdot \boldsymbol{v} + \tau\frac{(E_\mathrm{f} - E)}{T}\boldsymbol{v} \cdot \nabla_r T - \boldsymbol{v} \cdot \boldsymbol{P}(E) = 0 . \tag{7.36}$$

Equation (7.36) is a generalized steady-state Boltzmann equation in the presence of an applied electric field, a magnetic field, and a temperature gradient. Note that (7.36) can be further simplified by factoring out the velocity component, and the result is

$$\boldsymbol{P}(E) - \left(\frac{q\tau}{m_\mathrm{n}^*}\right)[\boldsymbol{B} \times \boldsymbol{P}(E)] = -q\tau\mathcal{E} + \tau\frac{(E_\mathrm{f} - E)}{T}\nabla_r T . \tag{7.37}$$

In order to obtain a solution for the unknown vector function $\boldsymbol{P}(E)$ in (7.37), it is assumed that the applied electric fields and temperature gradients are in the x–y plane of the semiconductor specimen, and the magnetic field is in the z-direction. Under this assumption, the components for $\boldsymbol{P}(E)$ in (7.37) in the x- and y-directions of the specimen are given respectively by

$$P_x(E) + (q\tau/m_\mathrm{n}^*)B_z P_y(E) = -q\tau\mathcal{E}_x + \tau\frac{(E_\mathrm{f} - E)}{T}\frac{\partial T}{\partial x} , \tag{7.38}$$

$$P_y(E) - (q\tau/m_\mathrm{n}^*)B_z P_x(E) = -q\tau\mathcal{E}_y + \tau\frac{(E_\mathrm{f} - E)}{T}\frac{\partial T}{\partial y} . \tag{7.39}$$

Solving (7.38) and (7.39) for $P_x(E)$ and $P_y(E)$, one obtains

$$P_x(E) = \frac{(\beta - \delta\gamma)}{(1 + \delta^2)} , \tag{7.40}$$

$$P_y(E) = \frac{(\gamma - \delta\beta)}{(1 + \delta^2)} , \tag{7.41}$$

where

$$\delta = \frac{q\tau B_z}{m_n^*} = \omega\tau \ , \tag{7.42}$$

$$\beta = \tau\left[-q\mathcal{E}_x + \frac{(E_f - E)}{T}\frac{\partial T}{\partial x}\right] , \tag{7.43}$$

$$\gamma = \tau\left[-q\mathcal{E}_y + \frac{(E_f - E)}{T}\frac{\partial T}{\partial y}\right] . \tag{7.44}$$

Thus, the expressions of transport coefficients for an n-type semiconductor described in Section 7.2 can be derived using (7.31) through (7.44). This is discussed next.

7.4.1. Electrical Conductivity

Consider the cases in which the applied electric field and the current flow are in the x- and y-directions of the specimen. By substituting (7.32) for $f(E)$ into (7.9), the electric current density components due to electron conduction along the x- and y-directions are given respectively by

$$J_x = -nqv_x = -\int_0^\infty q v_x f(E)g(E)\,\mathrm{d}E \tag{7.45}$$

$$= q\int_0^\infty v_x^2 P_x(E)g(E)\frac{\partial f_0}{\partial E}\,\mathrm{d}E \ ,$$

$$J_y = q\int_0^\infty v_y^2 P_y(E)g(E)\frac{\partial f_0}{\partial E}\,\mathrm{d}E \ , \tag{7.46}$$

where $P_x(E)$ and $P_y(E)$ are obtained from (7.40) through (7.44) by setting δ, $\partial T/\partial x$, and $\partial T/\partial y$ equal to zero, results yielding

$$P_x(E) = -q\tau\mathcal{E}_x \ , \tag{7.47}$$

$$P_y(E) = -q\tau\mathcal{E}_y \ . \tag{7.48}$$

From (7.45) through (7.48), it is noted that J_x and J_y vanish if $P(E)$ is equal to zero. To derive the electrical conductivity, it is assumed that the electron velocity is isotropic within the specimen, and hence the square of the velocity components along the x-, y-, and z-directions can be expressed in terms of the kinetic energy of electrons by

$$v_x^2 = v_y^2 = v_z^2 = \frac{2E}{3m_n^*} \ , \tag{7.49}$$

where E is the total kinetic energy of electrons. Equation (7.49) is obtained by using the fact that the electron kinetic energy is equal to $(1/2)m_n^* v^2$, where $v^2 = v_x^2 + v_y^2 + v_z^2$ and it is assumed that $v_x = v_y = v_z$. Now, substituting (7.47) and (7.49) into (7.45), one obtains

$$\sigma_n = \frac{J_x}{\mathcal{E}_x} = \left(\frac{-2q^2}{3m_n^*}\right)\int_0^\infty \tau E g(E)\frac{\partial f_0}{\partial E}\,\mathrm{d}E \ . \tag{7.50}$$

For a nondegenerate semiconductor, the F-D distribution function given by (7.1) is reduced to the classical M-B distribution function, which reads

$$f_0 \approx \exp[(E_\mathrm{f} - E)/k_\mathrm{B}T] \tag{7.51}$$

and

$$\frac{\partial f_0}{\partial E} = -\frac{f_0}{k_\mathrm{B}T} \,. \tag{7.52}$$

Now, substituting (7.52) into (7.50), the electrical conductivity can be expressed as

$$\begin{aligned} \sigma_\mathrm{n} &= \left(\frac{2q^2}{3m_\mathrm{n}^* k_B T}\right) \int_0^\infty \tau E g(E) f_0 \, \mathrm{d}E \\ &= \left(\frac{2n_0 q^2}{3m_\mathrm{n}^* k_B T}\right) \frac{\int_0^\infty \tau E^{3/2} f_0 \, \mathrm{d}E}{\int_0^\infty E^{3/2} f_0 \, \mathrm{d}E} \\ &= \frac{n_0 q^2 \langle \tau \rangle}{m_\mathrm{n}^*} \,, \end{aligned} \tag{7.53}$$

where

$$\langle \tau \rangle = \frac{\int_0^\infty \tau E^{3/2} \, \mathrm{e}^{-E/k_\mathrm{B}T} \, \mathrm{d}E}{\int_0^\infty E^{3/2} \, \mathrm{e}^{-E/k_\mathrm{B}T} \, \mathrm{d}E} \tag{7.54}$$

is the average relaxation time. It is noted that (7.54) is valid only for the nondegenerate semiconductors in which the M-B statistics one applicable. Equation (7.53) is obtained using the expression of electron density given by

$$n_0 = \int_0^\infty f_0 g(E) \, \mathrm{d}E = \left(\frac{4\pi}{h^3}\right) (2m_\mathrm{n}^*)^{3/2} \int_0^\infty E^{1/2} \, \mathrm{e}^{-E/k_\mathrm{B}T} \, \mathrm{d}E \,. \tag{7.55}$$

The average kinetic energy of electrons for a nondegenerate n-type semiconductor can be obtained using the expression

$$\langle E \rangle = \frac{\int_0^\infty E f_0 g(E) \, \mathrm{d}E}{\int_0^\infty f_0 g(E) \, \mathrm{d}E} = \frac{\int_0^\infty E^{3/2} \, \mathrm{e}^{-E/k_\mathrm{B}T} \, \mathrm{d}E}{\int_0^\infty E^{1/2} \, \mathrm{e}^{-E/k_\mathrm{B}T} \, \mathrm{d}E} = \frac{3k_\mathrm{B}T}{2} \,. \tag{7.56}$$

A generalized expression for the average relaxation time to the nth power $\langle \tau^n \rangle$ is given by

$$\langle \tau^n \rangle = \frac{\int_0^\infty \tau^n E g(E) \, \partial f_0 / \partial E \, \mathrm{d}E}{\int_0^\infty E g(E) \, \partial f_0 / \partial E \, \mathrm{d}E} \,, \tag{7.57}$$

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

$$\langle \tau E^n \rangle = \frac{\int_0^\infty (\tau E^n) E g(E) \, \partial f_0 / \partial E \, \mathrm{d}E}{\int_0^\infty E g(E) \, \partial f_0 / \partial E \, \mathrm{d}E} \,. \tag{7.58}$$

It is noted that the electrical conductivity for an n-type semiconductor given by (7.53) is similar to that of (7.7) for a metal. The only difference is that the free-electron mass in (7.7) is replaced by the effective mass of electrons, m_n^*, and the

constant relaxation time τ is replaced by an average relaxation time $\langle\tau\rangle$ defined Eq. (7.54).

Using the M-B distribution function for f_0 in (7.57) and (7.58), the expressions of $\langle\tau^n\rangle$ and $\langle\tau E^n\rangle$ for a nondegenerate semiconductor are given respectively by

$$\langle\tau^n\rangle = \frac{\int_0^\infty \tau^n E^{3/2}\, \mathrm{e}^{-E/k_\mathrm{B}T}\, \mathrm{d}E}{\int_0^\infty E^{3/2}\, \mathrm{e}^{-E/k_\mathrm{B}T}\mathrm{d}E}\ , \tag{7.59}$$

$$\langle\tau E^n\rangle = \frac{\int_0^\infty (\tau E^n) E^{3/2}\, \mathrm{e}^{-E/k_B T}\mathrm{d}E}{\int_0^\infty E^{3/2}\, \mathrm{e}^{-E/k_\mathrm{B}T}\mathrm{d}E}\ . \tag{7.60}$$

Now, solving (7.53) and (7.54) and using $\tau = \tau_0 E^s$, one obtains the expression of electrical conductivity for a nondegenerate n-type semiconductor as

$$\sigma_\mathrm{n} = \left(\frac{n_0 q^2 \tau_0}{m_\mathrm{n}^*}\right)(k_\mathrm{B}T)^s \frac{\Gamma_{(5/2+s)}}{\Gamma_{(5/2)}}\ , \tag{7.61}$$

where

$$\Gamma_n(x) = \int_0^\infty x^{n-1}\, \mathrm{e}^{-x}\, \mathrm{d}x \tag{7.62}$$

is the gamma function of order n, $\Gamma_n = (n-1)!$, and $\Gamma_{1/2} = \sqrt{\pi}$. Since the electrical conductivity is related to the electron mobility by (7.6), an expression of the electron mobility can be derived from (7.6) and (7.61), and the result is

$$\mu_\mathrm{n} = \left(\frac{q\tau_0}{m_n^*}\right)(k_\mathrm{B}T)^s \frac{\Gamma_{(5/2+s)}}{\Gamma_{(5/2)}}\ . \tag{7.63}$$

It is noted that for acoustical phonon scattering, τ_0 varies as T^{-1} and $s = -1/2$, and hence the electron mobility μ_n varies with $T^{-3/2}$. For ionized impurity scattering, $s = +3/2$, and τ_0 independent of temperature, the electron mobility varies as $T^{3/2}$. Detailed scattering mechanisms in a semiconductor will be discussed in Chapter 8.

The electrical conductivity given by (7.53) was derived on the basis of the single-valley model with a spherical constant-energy surface for the conduction band. This applies to most of the III-V compound semiconductors such as GaAs and InP in which the conduction band minimum is assumed to have spherical constant-energy surface (i.e., parabolic band). In this case the conductivity effective mass m_n^* is an isotropic scalar quantity, and n is the total carrier concentration in the single spherical conduction band. For multivalley semiconductors such as silicon and germanium, since their crystal structures possess cubic symmetry, the electrical conductivity remains isotropic. Thus, (7.53) is still applicable for the multivalley semiconductors, provided that the average relaxation time is assumed isotropic

and the conductivity effective mass m_n^* is replaced by

$$m_\sigma^* = \left[\frac{1}{3}\left(\frac{1}{m_\mathrm{l}} + \frac{2}{m_\mathrm{t}}\right)\right]^{-1} = \frac{3m_\mathrm{l}}{(2K+1)}, \tag{7.64}$$

where $K = m_\mathrm{l}/m_\mathrm{t}$ is the ratio of the longitudinal and transverse effective masses of an electron along the two main axes of the ellipsoidal energy surface near the conduction band edge. Values of m_l and m_t can be determined by the cyclotron resonance experiment at 4.2 K. Equation (7.64) is obtained using the geometrical average of the electron mass along the two main axes of the ellipsoidal energy surface.

7.4.2. Hall Coefficients

The general expression of the Hall coefficient for a nondegenerate n-type semiconductor with a single-valley spherical energy band can be derived from (7.45) and (7.46) using the definition given by (7.19). Consider the case of small magnetic field (i.e., $\mu B \ll 1$) in which the δ^2 term in (7.40) and (7.41) is negligible (i.e., $\delta^2 \ll 1$). Thus, by substituting $P_x(E)$ given by (7.40) into (7.45) and setting $\partial T/\partial x$ equal to zero, one obtains

$$\begin{aligned} J_x &= q\int_0^\infty v_x^2(\beta - \gamma\delta)g(E)\frac{\partial f_0}{\partial E}\,\mathrm{d}E \\ &= \left(\frac{2q^2}{3m_n^* k_\mathrm{B} T}\right)\int_0^\infty \tau E\left[\mathcal{E}_x - \left(\frac{q\tau B_z}{m_n^*}\right)\mathcal{E}_y\right] g(E) f_0\,\mathrm{d}E\,. \end{aligned} \tag{7.65}$$

Similarly, (7.46) can be expressed as

$$\begin{aligned} J_y &= q\int_0^\infty v_y^2(\gamma + \delta\beta)g(E)\frac{\partial f_0}{\partial E}\,\mathrm{d}E \\ &= \left(\frac{2q^2}{3m_\mathrm{n}^* k_B T}\right)\int_0^\infty \tau E\left[\mathcal{E}_y + \left(\frac{q\tau B_z}{m_\mathrm{n}^*}\right)\mathcal{E}_x\right] g(E) f_0\,\mathrm{d}E\,. \end{aligned} \tag{7.66}$$

By setting $J_y = 0$ in (7.66), $\mathcal{E}_x$ can be expressed in terms of $\mathcal{E}_y$, which can then be substituted into (7.65) to obtain an expression for the Hall coefficient using the definition of R_Hn given by (7.19), and one obtains

$$R_\mathrm{Hn} = \left.\frac{\mathcal{E}_y}{J_x B_z}\right|_{J_y=0} = -\left(\frac{3k_\mathrm{B}T}{2q}\right)\frac{\int_0^\infty \tau^2 E g(E) f_0\,\mathrm{d}E}{[\int_0^\infty \tau E g(E) f_0\,\mathrm{d}E]^2} = -\frac{1}{qn_0}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}, \tag{7.67}$$

where $\langle\tau\rangle$ is the average relaxation time and $\langle\tau^2\rangle$ is the average of the relaxation time squared, which can be determined using (7.57). The minus sign in (7.67) denotes electron conduction in an n-type semiconductor. For p-type semiconductors,

the Hall coefficient is given by

$$R_{\mathrm{Hp}} = \frac{1}{qp_0}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}, \tag{7.68}$$

which has a positive Hall coefficient due to the hole conduction. From (7.67) and (7.68), the Hall factor γ_{H} can be expressed by

$$\gamma_{\mathrm{H}} = \frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}. \tag{7.69}$$

If the relaxation time is given by $\tau = aE^s$, then the Hall coefficient for a nondegenerate n-type semiconductor is given by

$$R_{\mathrm{Hn}} = -\frac{1}{qn}\frac{\Gamma_{(2s+5/2)}\Gamma_{(5/2)}}{[\Gamma_{(s+5/2)}]^2} = -\frac{\gamma_{\mathrm{Hn}}}{qn}, \tag{7.70}$$

where Γ_{n} is the gamma function defined by (7.62) and γ_{Hn} is the Hall factor for an n-type semiconductor. In general, the Hall factor can be calculated if the scattering mechanisms in the semiconductor are known. The expression for the Hall factor for a p-type semiconductor is identical to that of n-type semiconductors discussed above.

Another important physical parameter, which is usually referred to as Hall mobility, can be obtained from the product of electrical conductivity and the Hall coefficient. Thus, using (7.61) and (7.70) one obtains

$$\mu_{\mathrm{Hn}} = R_{\mathrm{Hn}}\sigma_{\mathrm{n}} = \left(\frac{q\tau_0}{m_n^*}\right)(k_{\mathrm{B}}T)^s\frac{\Gamma_{(2s+5/2)}}{\Gamma_{(s+5/2)}}. \tag{7.71}$$

The Hall factor for a nondegenerate semiconductor, which is defined as the ratio of Hall mobility and conductivity mobility, can be obtained from (7.69) and (7.70)

$$\gamma_{\mathrm{Hn}} = \frac{\mu_{\mathrm{Hn}}}{\mu_{\mathrm{n}}} = \frac{\langle\tau^2\rangle}{\langle\tau\rangle^2} = \frac{\Gamma_{(2s+5/2)}\Gamma_{(5/2)}}{[\Gamma_{(s+5/2)}]^2}. \tag{7.72}$$

Values of γ_{Hn} may vary between 1.18 and 1.93 depending on the types of scattering mechanisms involved. For example, for acoustical phonon scattering with $s = -1/2$, the Hall factor is equal to $3\pi/8$ ($\approx$1.18), and for ionized impurity scattering with $s = 3/2$, the Hall factor was found equal to $315\pi/512$ ($\approx$1.93). For neutral impurity scattering with $s = 0$, the Hall factor is equal to 1. Values of the Hall factor given above are obtained for nondegenerate semiconductors with a single-valley spherical energy surface in the conduction band.

For multivalley semiconductors such as silicon and germanium where the conduction band valley has an ellipsoidal energy surface, the expression for the Hall factor should be modified to include the mass anisotropic effect. In this case, a

"Hall mass factor" a_0 is multiplied by the Hall factor given by (7.69) for the single-valley case to account for the mass anisotropic effect. Thus, the Hall factor for a multivalley semiconductor can be expressed by

$$\gamma_{\mathrm{H}} = \frac{\mu_{\mathrm{Hn}}}{\mu_{\mathrm{n}}} = \frac{\langle \tau^2 \rangle a_0}{\langle \tau \rangle^2}, \tag{7.73}$$

where a_0 is known as the "Hall mass factor," which can be expressed by

$$a_0 = \left(\frac{m_\sigma^*}{m_{\mathrm{H}}^*} \right)^2 = \frac{3K(K+2)}{(2K+1)^2}, \tag{7.74}$$

where

$$m_{\mathrm{H}}^* = m_{\mathrm{l}} \sqrt{3/[K(K+2)]} \tag{7.75}$$

is the Hall effective mass, m_σ^* is the conductivity effective mass defined by (7.64), and $K = m_{\mathrm{l}}/m_{\mathrm{t}}$ is the ratio of the longitudinal and transverse effective masses of electrons along the two major axes of the constant ellipsoidal energy surface of the conduction band. For germanium, $K \approx 20$ and $a_0 = 0.785$, while for silicon, $K = 5.2$ and $a_0 = 0.864$. Thus, the Hall coefficient given by (7.67) for n-type silicon and germanium should be multiplied by a Hall mass factor given by (7.74).

For p-type silicon, the Hall factor may be smaller than unity because of the warped and nonparabolic valence band structures. In general, it is usually difficult to obtain an exact Hall factor from the Hall effect measurements. In fact, it is a common practice to assume that the Hall factor is equal to one so that the majority carrier density in a semiconductor can be readily determined from the Hall effect measurements.

7.4.3. Seebeck Coefficients

The Seebeck coefficient for a single-valley n-type semiconductor with spherical constant-energy surface can be derived from (7.40) and (7.45) by letting $\delta = 0$ in (7.40) and $J_x = 0$ in (7.45). One then has

$$\begin{aligned} S_{\mathrm{n}} &= \left. \frac{\mathcal{E}_x}{(\partial T/\partial x)} \right|_{J_x = 0} = \left(-\frac{1}{qT} \right) \left[\frac{\int_0^\infty \tau E^2 g(E) \partial f_0/\partial E \, \mathrm{d}E}{\int_0^\infty \tau E g(E) \partial f_0/\partial E \, \mathrm{d}E} - E_{\mathrm{f}} \right] \\ &= -\left(\frac{1}{qT} \right) \left[\frac{\langle \tau E \rangle}{\langle \tau \rangle} - E_{\mathrm{f}} \right]. \end{aligned} \tag{7.76}$$

For a nondegenerate semiconductor, the Seebeck coefficient given by (7.76) can be derived using (7.59) and (7.60) to obtain $\langle \tau E \rangle$ and $\langle \tau \rangle$ with $\tau = aE^s$, yielding

$$S_{\mathrm{n}} = -\left(\frac{1}{qT} \right) [(5/2 + s) k_{\mathrm{B}} T - E_{\mathrm{f}}]. \tag{7.77}$$

It is noted that values of the Seebeck coefficient given by (7.77) can be determined if the types of scattering mechanisms and the position of Fermi level are known. For example, if acoustical phonon scattering is dominant (i.e., $s = -1/2$), then the Seebeck coefficient is given by

$$S_{\mathrm{n}} = -\left(\frac{1}{qT}\right)(2k_{\mathrm{B}}T - E_{\mathrm{f}}). \tag{7.78}$$

On the other hand, if ionized impurity scattering (i.e., $s = 3/2$) is dominant, then the Seebeck coefficient becomes

$$S_{\mathrm{n}} = -\left(\frac{1}{qT}\right)(4k_{\mathrm{B}}T - E_{\mathrm{f}}). \tag{7.79}$$

From (7.78) and (7.79), it is seen that the Fermi energy for a nondegenerate semiconductor can be determined from the measured Seebeck coefficient, provided that the dominant scattering mechanism is known. The minus sign in (7.77) for the Seebeck coefficient indicates that the conduction is due to electrons in an n-type semiconductor. Thus, measurement of the Seebeck coefficient can also be used to determine the conduction type of a semiconductor.

For p-type semiconductors, the sign of the Seebeck coefficient given by (7.76) is positive, since the conduction is carried out by holes. Expressions of Seebeck coefficient derived in this section are applicable only for single-valley nondegenerate semiconductors. However, the results can also be applied to multivalley semiconductors such as silicon and germanium, provided that the density-of-states effective mass for electrons is modified to account for the multivalley conduction bands. For degenerate semiconductors, the F-D distribution function should be used in deriving the Seebeck coefficients. This is left as an exercise for the reader in the problems.

7.4.4. Nernst Coefficients

The Nernst effect for a nondegenerate n-type semiconductor with a single-valley spherical energy surface in the conduction band is discussed next. The Nernst coefficient Q_{n}, defined by (7.20), can be derived from (7.40) through (7.48). Consider the case of low magnetic fields (i.e., $\mu B = \delta \ll 1$). Under isothermal conditions, the Nernst coefficient can be derived by letting $J_x = J_y = 0$ and $\partial T/\partial y = 0$ in (7.45) and (7.46), which can be expressed as

$$\begin{aligned} J_x = 0 &= q\int_0^\infty v_x^2(\beta - \delta\gamma)g(E)\frac{\partial f_0}{\partial E}\,\mathrm{d}E \\ &= \left(\frac{2q^2}{3m_{\mathrm{n}}^* k_{\mathrm{B}}T}\right)\int_0^\infty \tau E\left[\mathcal{E}_x - \frac{(E_{\mathrm{f}} - E)}{T}\left(\frac{\partial T}{\partial x}\right) - \omega\tau\mathcal{E}_y\right]g(E)f_0\,\mathrm{d}E \end{aligned} \tag{7.80}$$

and

$$J_y = 0 = \left(\frac{2q^2}{3m_n^* k_B T}\right) \int_0^\infty \tau E \left[\mathcal{E}_y - \frac{(E_f - E)}{T} \omega\tau \left(\frac{\partial T}{\partial x}\right) + \omega\tau \mathcal{E}_x\right] g(E) f_0 \, dE. \tag{7.81}$$

Now solving (7.80), (7.81), and using (7.20) for Q_n, one obtains

$$\begin{aligned} Q_n &= \left.\frac{\mathcal{E}_y}{B_z(\partial T/\partial x)}\right|_{J_x = J_y = 0} \\ &= \left\{\left(\frac{1}{m^* T}\right) \frac{\int_0^\infty \tau^2 E^2 g(E) f_0 \, dE}{\int_0^\infty \tau E g(E) f_0 \, dE} - \frac{\int_0^\infty \tau^2 E g(E) f_0 \, dE \int_0^\infty \tau E^2 g(E) f_0 \, dE}{[\int_0^\infty \tau E g(E) f_0 \, dE]^2}\right\} \\ &= \left(\frac{\mu_n}{qT}\right) \left[\frac{\langle \tau^2 E \rangle}{\langle \tau \rangle^2} - \frac{\langle \tau^2 \rangle \langle \tau E \rangle}{\langle \tau \rangle^3}\right], \end{aligned} \tag{7.82}$$

where $\mu_n = q\langle\tau\rangle/m_n^*$ is the electron conductivity mobility. Now substituting $\tau = aE^s$ into (7.82), one obtains the Nernst coefficient for a nondegenerate n-type semiconductor as

$$Q_n = \left(\frac{k_B}{q}\right) \mu_n s \frac{\Gamma_{(2s+5/2)} \Gamma_{(5/2)}}{[\Gamma_{(s+5/2)}]^2}. \tag{7.83}$$

If the acoustical phonon scattering ($s = -1/2$) or ionized impurity scattering ($s = 3/2$) is the dominant scattering mechanism, then (7.83) is given by

$$\begin{aligned} Q_n &= -\left(\frac{3\pi}{16}\right)\left(\frac{k_B}{q}\right)\mu_n \quad \text{for } s = -1/2, \\ &= \left(\frac{945\pi}{1024}\right)\left(\frac{k_B}{q}\right)\mu_n \quad \text{for } s = 3/2. \end{aligned} \tag{7.84}$$

It is interesting to note that, in contrast to both the Hall and Seebeck coefficients, the sign of the Nernst coefficient given by (7.83) depends only on the types of scattering mechanisms (s) rather than on the types of charge carriers. For example, the Nernst coefficient is negative when the acoustical phonon scattering (i.e., $s = -1/2$) is dominant, and is positive when the ionized impurity scattering (i.e., $s = +3/2$) is dominant, as shown in (7.84). Equation (7.83) can be applied to p-type semiconductors, provided that the hole mobility is used in the expression.

7.4.5. Transverse Magnetoresistance

The transverse magnetoresistance effect describes the change of electrical resistivity when a transverse magnetic field is applied across a semiconductor specimen. For example, if an electric field in the x-direction and a magnetic field in the z-direction are applied simultaneously to a semiconductor specimen, then an increase in resistance with the applied magnetic field may be observed along the direction of current flow. The magnetoresistance in a semiconductor can be derived using (7.40) through (7.46). To derive an expression for transverse magnetoresistance in a single-valley semiconductor with spherical energy band, it is assumed that

the specimen is subject to isothermal conditions with $\partial T/\partial x = \partial T/\partial y = 0$ and $J_y = 0$. Equations (7.40) and (7.41) can be rewritten as

$$P_x(E) = \frac{(\beta - \delta\gamma)}{(1+\delta^2)} = \frac{(\beta - \omega\tau\gamma)}{(1+\omega^2\tau^2)}, \tag{7.85}$$

$$P_y(E) = \frac{(\gamma + \delta\beta)}{(1+\delta^2)} = \frac{(\gamma + \omega\tau\beta)}{(1+\omega^2\tau^2)}, \tag{7.86}$$

where

$$\gamma = -q\tau\mathcal{E}_y, \quad \beta = -q\tau\mathcal{E}_x, \quad \delta = \omega\tau = \frac{qB_z\tau}{m_{\mathrm{n}}^*}. \tag{7.87}$$

Now substituting (7.85) and (7.86) into (7.45) and (7.46) yields

$$J_x = \left(\frac{2q^2}{3m_{\mathrm{n}}^* k_{\mathrm{B}} T}\right) \int_0^\infty \left[\frac{(\tau\mathcal{E}_x + \omega\tau^2\mathcal{E}_y)}{(1+\omega^2\tau^2)}\right] Eg(E) f_0 \, \mathrm{d}E, \tag{7.88}$$

$$J_y = \left(\frac{2q^2}{3m_{\mathrm{n}}^* k_{\mathrm{B}} T}\right) \int_0^\infty \left[\frac{(\tau\mathcal{E}_y - \omega\tau^2\mathcal{E}_x)}{(1+\omega^2\tau^2)}\right] Eg(E) f_0 \, \mathrm{d}E. \tag{7.89}$$

Solving (7.88) and (7.89), one obtains the transverse magnetoresistance coefficients of a nondegenerate n-type semiconductor for two limiting cases, namely, the low and high magnetic fields cases.

*(i) The low magnetic field case (*i.e., $\delta = \omega\tau \ll 1$*).* In this case, the $\omega^2\tau^2$ term in the denominator of (7.88) and (7.89) is retained. From (7.89), $\mathcal{E}_y$ may be expressed in terms of $\mathcal{E}_x$ by setting $J_y = 0$, which yields

$$\mathcal{E}_y = -\mathcal{E}_x \left[\frac{\int_0^\infty \omega\tau^2 Eg(E) f_0 \, \mathrm{d}E}{\int_0^\infty \tau Eg(E) f_0 \, \mathrm{d}E}\right]. \tag{7.90}$$

Now, substituting (7.90) for $\mathcal{E}_y$ into (7.88) and using the binomial expansion for $(1+\omega^2\tau^2)^{-1} \approx (1-\omega^2\tau^2)$ in (7.88) for $\omega\tau \ll 1$, one obtains the electrical conductivity σ_{n} for the low magnetic field case:

$$\begin{aligned}
\sigma_{\mathrm{n}} &= \frac{J_x}{\mathcal{E}_x} \\
&= \left(\frac{2q^2}{3m_{\mathrm{n}}^* k_{\mathrm{B}} T}\right) \left[\int_0^\infty \tau E(1-\omega^2\tau^2) g(E) f_0 \, \mathrm{d}E + \frac{\omega^2 (\int_0^\infty \tau^2 Eg(E) f_0 \, \mathrm{d}E)^2}{\int_0^\infty \tau Eg(E) f_0 \, \mathrm{d}E}\right] \\
&= \sigma_0 \left[1 - \omega^2 \left(\frac{\langle\tau^3\rangle}{\langle\tau\rangle} - \frac{\langle\tau^2\rangle^2}{\langle\tau\rangle^2}\right)\right],
\end{aligned} \tag{7.91}$$

where $\sigma_0 = nq^2 \langle\tau\rangle / m_{\mathrm{n}}^*$ is the electrical conductivity at zero magnetic field.

For the low magnetic field case, the electrical conductivity is given by

$$\sigma_{\mathrm{n}} = \frac{1}{\rho_{\mathrm{n}}} = \sigma_0 - \Delta\sigma = \sigma_0 \left(1 - \frac{\Delta\sigma}{\sigma_0}\right), \tag{7.92}$$

where ρ_n is the resistivity of the semiconductor in the presence of a small magnetic field, which can be expressed by

$$\rho_n = \rho_0 \left(1 + \frac{\Delta\rho}{\rho_0}\right) \approx \sigma_0^{-1} \left(1 + \frac{\Delta\sigma}{\sigma_0}\right). \tag{7.93}$$

Solving (7.91) through (7.93), one obtains

$$\begin{aligned}
\frac{\Delta\rho}{\rho_0} = \frac{\Delta\sigma}{\sigma_0} &= \omega^2 \left[\frac{\langle\tau^3\rangle}{\langle\tau\rangle} - \frac{\langle\tau^2\rangle^2}{\langle\tau\rangle^2}\right] \\
&= (\sigma_0^2 B_z^2) \left[\left(\frac{1}{nq}\right)\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}\right]^2 \left[\frac{\langle\tau^3\rangle\langle\tau\rangle}{\langle\tau^2\rangle^2} - 1\right] \\
&= R_H^2 \sigma_0^2 B_z^2 \left[\frac{\langle\tau^3\rangle\langle\tau\rangle}{\langle\tau^2\rangle^2} - 1\right] \\
&= \mu_H^2 B_z^2 \left[\frac{\langle\tau^3\rangle\langle\tau\rangle}{\langle\tau^2\rangle^2} - 1\right].
\end{aligned} \tag{7.94}$$

The magnetoresistance coefficient can be deduced from (7.94), which yields

$$\xi = \left(\frac{\Delta\rho}{\rho_0 B_z^2}\right)\left(\frac{1}{\mu_H^2}\right) = \frac{\langle\tau^3\rangle\langle\tau\rangle}{\langle\tau^2\rangle^2} - 1. \tag{7.95}$$

For the nondegenerate semiconductor, using $\tau = \tau_0 E^s$, (7.95) becomes

$$\xi = \frac{\Gamma_{(3s+5/2)}\Gamma_{(s+5/2)}}{[\Gamma_{(2s+5/2)}]^2} - 1. \tag{7.96}$$

From (7.96), one finds that ξ is equal to 0.273 for acoustical phonon scattering (i.e., $s = -1/2$) and 0.57 for ionized impurity scattering (i.e., $s = 3/2$).

Under low magnetic field conditions, the above results show that the transverse magnetoresistance in a nondegenerate semiconductor is directly proportional to the square of the magnetic field. The magnetoresistance data obtained for semiconductors with a single spherical energy band were found in good agreement with the theoretical prediction at low magnetic fields. It should be noted that for semiconductors with spherical energy bands, the longitudinal magnetoresistance (i.e., $J_n // B$) should vanish under the small magnetic field condition.

*(ii) The high magnetic field case (*i.e., $\delta = \omega\tau \gg 1$*).* At high magnetic fields, (7.88) and (7.89) become

$$J_x = \left(\frac{2q^2}{3m_n^* k_B T}\right)\int_0^\infty E(-\mathcal{E}_y/\omega) g(E)\, f_0\, dE, \tag{7.97}$$

$$J_y = \left(\frac{nq^2}{m_n^* \omega^2}\right)[\mathcal{E}_y\langle\tau^{-1}\rangle + \omega\mathcal{E}_x]. \tag{7.98}$$

Thus, for $J_y = 0$, one obtains

$$\mathcal{E}_y = -\frac{\omega \mathcal{E}_x}{\langle \tau^{-1} \rangle}. \tag{7.99}$$

Now substituting (7.99) into (7.97) yields

$$\sigma_\infty = \frac{J_x}{\mathcal{E}_x} = \left(\frac{nq^2}{m_{\mathrm{n}}^*}\right)\left(\frac{1}{\langle \tau^{-1} \rangle}\right) = \frac{\sigma_0}{\langle \tau \rangle \langle \tau^{-1} \rangle}, \tag{7.100}$$

or

$$\frac{\sigma_0}{\sigma_\infty} = \frac{\rho_\infty}{\rho_0} = \langle \tau \rangle \langle \tau^{-1} \rangle, \tag{7.101}$$

which shows that the ratio of electrical conductivity at zero and high magnetic fields is equal to a constant whose value depends only on the types of scattering mechanisms involved. In general, the transverse magnetoresistance approaches a constant value at very high magnetic fields. For a nondegenerate semiconductor, (7.101) becomes

$$\frac{\sigma_0}{\sigma_\infty} = \frac{\rho_\infty}{\rho_0} = \frac{\Gamma_{(s+5/2)}\Gamma_{(5/2-s)}}{[\Gamma_{(5/2)}]^2}. \tag{7.102}$$

Using (7.102) one obtains the value of $\rho_\infty/\rho_0 = 1.17$ for $s = -1/2$ and $\rho_\infty/\rho_0 = 3.51$ for $s = 3/2$. The high field magnetoresistance value is three times higher for the ionized impurity scattering than for the acoustical phonon scattering case.

The magnetoresistance coefficients derived above are valid only for the single-valley conduction band with spherical energy surface. For multivalley semiconductors such as silicon and germanium, the situation is more complicated than that presented in this section. The effective mass anisotropy strongly affects the magnetoresistance value. For example, if an electric field and a magnetic field are applied parallel to the x-direction of an n-type germanium in which the conduction valleys are located along the {111} axes, the longitudinal magnetoresistance along the (100) direction is given by

$$\frac{\Delta\rho}{\rho B^2} = \frac{q^2}{m_{\mathrm{l}}^2}\,\frac{2K(K-1)^2}{3(2K+1)}\,\frac{\langle \tau^3 \rangle}{\langle \tau \rangle} = \mu_{\mathrm{Mn}}^2 \frac{2K(K-1)^2}{3(2K+1)}, \tag{7.103}$$

where $\mu_{\mathrm{Mn}}^2 = (q^2/m_{\mathrm{l}}^2)(\langle \tau^3 \rangle/\langle \tau \rangle)$ is the square of the electron mobility associated with the magnetoresistance effect in a semiconductor. For n-type germanium with $K = 20$, the K-dependent factor in (7.103) has a value equal to 118, which reduces to zero for $K = 1$. Thus, the longitudinal magnetoresistance is strongly affected by the effective mass anisotropy for n-type germanium. For n-type silicon, if the electric current and magnetic field are applied simultaneously along the (100) direction, the longitudinal magnetoresistance should vanish since σ_{xx} is independent of the magnetic field B_z.

It is noted that the high magnetic field case discussed above is valid only for the classical limit in that the magnetoresistance coefficient does not exhibit any quantum oscillatory behavior. In the quantum limit, the magnetoresistance exhibits

oscillatory behavior at very high magnetic fields. The oscillatory behavior of the magnetoresistance observed in metals is known as the Shubnikov–de Haas–van Alphen effect. This effect is usually observed in the degenerate electron gas in a metal at very low temperatures (e.g., at 4.2 K) and under very high magnetic fields (e.g., several hundred kilogauss). The Shubnikov–de Haas–van Alphen effect has been widely used in the construction of the energy contour of the Fermi surface in a metal.

For p-type semiconductors in which holes are the majority carriers, the expressions of various transport coefficients derived above are still valid, provided that the positive sign is used for the electronic charge and the electron effective mass is replaced by the hole effective mass.

The transport coefficients derived in the preceding section are valid only for nondegenerate n-type semiconductors in which the classical M-B statistics are used to obtain the average relaxation time contained in the expressions of the transport coefficients. However, for degenerate n-type semiconductors, the F-D statistics should be used in the derivation of these transport coefficients. This will be left as exercises for the reader to derive in the problems section.

7.5. Transport Coefficients for the Mixed Conduction Case

In an intrinsic semiconductor or an extrinsic semiconductor with heavy compensation, both electrons and holes can be participated in the conduction process and hence mixed conduction prevails in carrier transport. In this case, the conduction is a two-carrier process contributed by both electrons and holes. Derivation of transport coefficients for the mixed conduction case is more complicated than that of the single-carrier conduction case discussed in the previous sections. This is discussed next.

7.5.1. Electrical Conductivity

The electrical conductivity for a two-carrier conduction is equal to the sum of the single-carrier conductivities due to electrons and holes. Thus, one can write

$$\sigma = \sigma_{\mathrm{n}} + \sigma_{\mathrm{p}} = q(n_0\mu_{\mathrm{n}} + p_0\mu_{\mathrm{p}}), \tag{7.104}$$

where σ_{n} and σ_{p} denote the electron and hole conductivities, μ_{n} and μ_{p} are the electron and hole mobilities, while n and p represent the electron and hole densities, respectively. Thus, the electrical conductivity for the mixed conduction case can be obtained by substituting the expressions for σ_{n} and σ_{p} derived from the single-carrier conduction case into (7.104).

7.5.2. Hall Coefficient

The Hall coefficient for the mixed conduction case can be derived as follows. If an electric field is applied in the x-direction and a magnetic field in the z-direction,

then a Hall voltage is developed in the y-direction of the specimen. The total current flow due to both electrons and holes in the x-direction is given by

$$J_x = J_{nx} + J_{px}, \tag{7.105}$$

where J_x, J_{nx}, and J_{px} are given respectively by

$$J_x = \sigma \mathcal{E}_x, \quad J_{nx} = \sigma_{\mathrm{n}} \mathcal{E}_x, \quad J_{px} = \sigma_{\mathrm{p}} \mathcal{E}_x \,. \tag{7.106}$$

The Hall fields due to electrons and holes developed in the y-direction can be expressed by

$$\mathcal{E}_y = R_{\mathrm{H}} J_x B_z, \quad \mathcal{E}_{ny} = R_{\mathrm{Hn}} J_{nx} B_z, \quad \mathcal{E}_{py} = R_{\mathrm{Hp}} J_{px} B_z. \tag{7.107}$$

Since the electric current density in the y-direction is given by $J_y = J_{ny} + J_{py}$, one obtains

$$\sigma \mathcal{E}_y = \sigma_{\mathrm{n}} \mathcal{E}_{ny} + \sigma_{\mathrm{p}} \mathcal{E}_{py}. \tag{7.108}$$

Substituting (7.107) for $\mathcal{E}_y$, $\mathcal{E}_{ny}$, and $\mathcal{E}_{py}$ into (7.108) yields

$$\sigma R_{\mathrm{H}} J_x B_z = \sigma_{\mathrm{n}} R_{\mathrm{Hn}} J_{nx} B_z + \sigma_{\mathrm{p}} R_{\mathrm{Hp}} J_{px} B_z. \tag{7.109}$$

Substituting (7.106) for J_x, J_{nx}, and J_{px} into (7.109), one obtains

$$\sigma^2 R_{\mathrm{H}} \mathcal{E}_x B_z = \sigma_{\mathrm{n}}^2 R_{\mathrm{Hn}} \mathcal{E}_x B_z + \sigma_{\mathrm{p}}^2 R_{\mathrm{Hp}} \mathcal{E}_x B_z. \tag{7.110}$$

Thus the Hall coefficient due to electron and hole conduction can be obtained from (7.110), which yields

$$R_{\mathrm{H}} = \frac{\left(R_{\mathrm{Hn}} \sigma_{\mathrm{n}}^2 + R_{\mathrm{Hp}} \sigma_{\mathrm{p}}^2 \right)}{(\sigma_{\mathrm{n}} + \sigma_{\mathrm{p}})^2}, \tag{7.111}$$

where R_{Hn} and R_{Hp} are the Hall coefficients given by (7.68) and (7.70) for n- and p-type conduction, respectively; σ_{n} and σ_{p} denote the corresponding electron and hole conductivities.

It is interesting to note that the Hall coefficient for the mixed conduction case given above is not equal to the simple summation of the Hall coefficients due to electrons and holes. In fact, (7.111) shows that the Hall coefficient may become zero if the contribution of the Hall coefficient from electrons (i.e., the first term, negative) is equal to that from holes (i.e., the second term, positive) in the numerator. This phenomenon may be observed in the Hall coefficient versus temperature plot for an intrinsic semiconductor in which changes in conductivity type may occur from n-type to p-type while the Hall coefficient changes from negative to positive at a certain elevated temperature.

7.5.3. Seebeck Coefficient

The Seebeck coefficient for the mixed conduction case can be derived in a similar way as the Hall coefficient. First, one considers the electric current density contributed by both electrons and holes in the presence of an electric field and a

temperature gradient in the x-direction. The electric current density due to electrons is given by

$$J_{nx} = \sigma_{\mathrm{n}} \left(\mathcal{E}_x - S_{\mathrm{n}} \frac{\partial T}{\partial x} \right), \tag{7.112}$$

and the electric current density due to holes is

$$J_{px} = \sigma_{\mathrm{p}} \left(\mathcal{E}_x - S_{\mathrm{p}} \frac{\partial T}{\partial x} \right). \tag{7.113}$$

If the temperature gradient is zero, then the total electric current density in the x-direction is given by

$$J_x = J_{nx} + J_{px} = (\sigma_{\mathrm{n}} + \sigma_{\mathrm{p}}) \mathcal{E}_x = \sigma \mathcal{E}_x. \tag{7.114}$$

Note that the electrical conductivity is simply equal to the sum of the conductivities due to electrons and holes. Thus, in the mixed conduction case the thermoelectric effect is obtained from (7.112) and (7.113) by setting the total current density in the x-direction equal to zero (i.e., $J_x = 0$), which yields

$$(\sigma_{\mathrm{n}} + \sigma_{\mathrm{p}}) \mathcal{E}_x = (S_{\mathrm{n}} \sigma_{\mathrm{n}} + S_{\mathrm{p}} \sigma_{\mathrm{p}}) \frac{\partial T}{\partial x}. \tag{7.115}$$

From (7.115), the total Seebeck coefficient S for the mixed conduction case is given by

$$S = \left. \frac{\mathcal{E}_x}{\partial T / \partial x} \right|_{J_x = 0} = \frac{(S_{\mathrm{n}} \sigma_{\mathrm{n}} + S_{\mathrm{p}} \sigma_{\mathrm{p}})}{(\sigma_{\mathrm{n}} + \sigma_{\mathrm{p}})}, \tag{7.116}$$

where S_{n} denotes the Seebeck coefficient for n-type conduction given by (7.76). The Seebeck coefficient S_{p} for p-type conduction is similar to (7.76), except that n is replaced by p, and a plus sign is used instead.

7.5.4. Nernst Coefficient

When both electrons and holes are present in a semiconductor, the Nernst coefficient Q is not equal to the simple sum of Q_{n} and Q_{p} because in the mixed conduction case an additional temperature gradient is developed in the specimen. This temperature gradient causes a Seebeck voltage to appear across the specimen. The Seebeck voltage in turn creates an electric field, which results in a flow of charge carriers. This current flow can induce a Hall voltage when a transverse magnetic field is applied to the specimen. Therefore, it can be shown that the Nernst coefficient for the mixed conduction case can be expressed in terms of σ_{n}, σ_{p}, S_{n}, S_{p}, R_{Hn}, R_{Hp}, Q_{n}, and Q_{p} for the single-carrier conduction case as

$$Q = \frac{(Q_{\mathrm{n}} \sigma_{\mathrm{n}} + Q_{\mathrm{p}} \sigma_{\mathrm{p}})(\sigma_{\mathrm{n}} + \sigma_{\mathrm{p}}) + (S_{\mathrm{n}} - S_{\mathrm{p}}) \sigma_{\mathrm{n}} \sigma_{\mathrm{p}} (R_{\mathrm{Hn}} \sigma_{\mathrm{n}} - R_{\mathrm{Hp}} \sigma_{\mathrm{p}})}{(\sigma_{\mathrm{n}} + \sigma_{\mathrm{p}})^2}. \tag{7.117}$$

It is seen that the total Nernst coefficient for a given semiconductor depends on both the location of the Fermi level and the types of scattering mechanisms

involved. For an intrinsic semiconductor, the Nernst coefficient differs considerably from that of Q_n and Q_p for the single-carrier conduction case.

It should be noted that exact expressions for the transport coefficients for the mixed conduction case can be obtained by inserting the individual transport coefficients derived for the n- or p-type single-carrier conduction case into the transport coefficient formula given by (7.104) to (7.117) for the mixed conduction case.

7.6. Transport Coefficients for Some Semiconductors

Measurements of the transport coefficients for elemental and compound semiconductors have been widely reported in the literature. Some of these results are discussed in this section. Since silicon, germanium, and GaAs have been studied most extensively in the past, it is pertinent to describe some of the resistivity, Hall coefficient, Hall mobility, Seebeck coefficient, and magnetoresistance data for these materials.

Figure 7.4 shows the resistivity as a function of reciprocal temperature for several n-type As-doped germanium specimens of different doping concentrations.[1] Figure 7.5 shows the corresponding Hall coefficient curves.[1] The results indicate that in the high-temperature regime, values of resistivity for all samples are almost identical

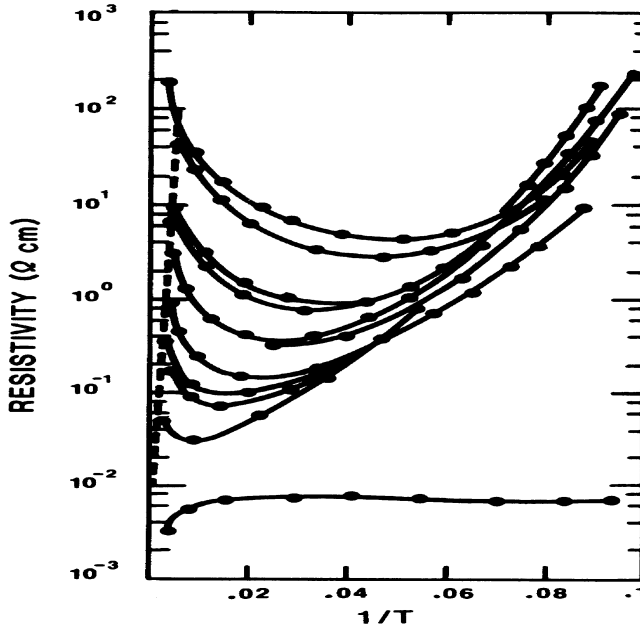


FIGURE 7.4. Resistivity versus inverse absolute temperature for the As-doped germanium samples of different donor concentrations (from low 10^{14} for the top curve to 10^{18} cm^{-3} for the bottom curve). After Debye and Conwell,[1] by permission.

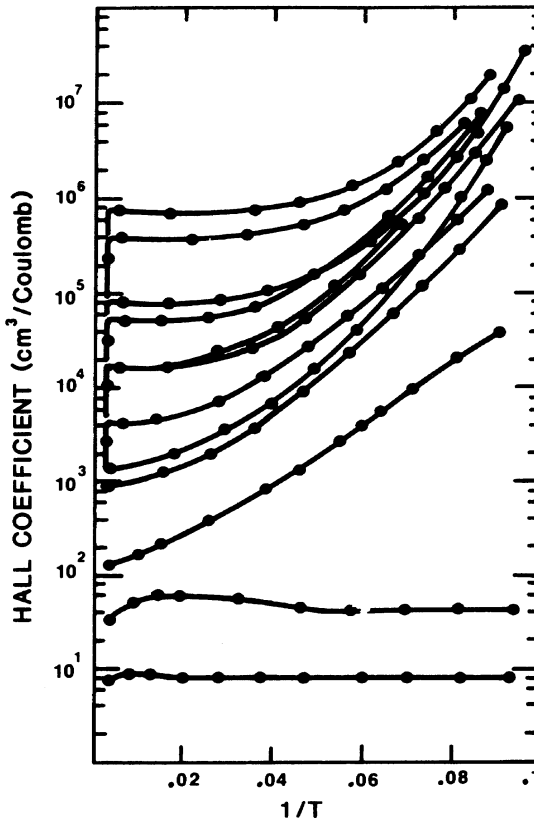


FIGURE 7.5. Hall coefficient versus inverse absolute temperature for the As-doped germanium samples of different donor concentrations (from low 10^{14} for the top curve to 10^{18} cm^{-3} for the bottom curve). After Debye and Conwell,[1] by permission.

and independent of the dopant densities. This is the intrinsic regime, and the carrier concentration is predicted by (5.20) for the intrinsic semiconductor case. In this regime, the densities of electrons and holes are equal and increase exponentially with temperature with a slope equal to $-E_g/2k_B$. As the temperature decreases, the material becomes an extrinsic semiconductor. In this temperature regime (i.e., the exhaustion regime), all impurity atoms are ionized, and the carrier density is equal to the net dopant density. As the temperature further decreases, carrier freeze-out occurs in the material. This is the so-called deionization regime. In this regime, the Hall coefficient increases again, and from the slope of the Hall coefficient versus temperature curve one can determine the shallow impurity activation energy. At very high doping concentrations, the carrier density becomes nearly constant over the entire temperature range, as is evident by the flatness of the resistivity and Hall coefficient curves (bottom curves) shown in Figures 7.4 and 7.5. The carrier concentration as a function of temperature can be deduced from Figure 7.5, and the Hall mobilities can be obtained from the product of the Hall coefficient and electrical resistivity curves shown in Figures 7.4 and 7.5.

Measurements of transport coefficients for both n- and p-type silicon have been widely reported in the literature, and some of these results are illustrated

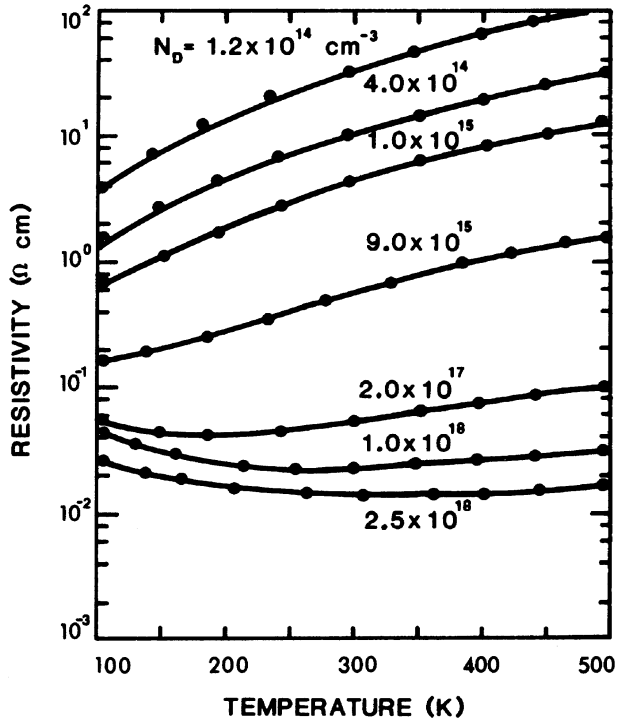


FIGURE 7.6. Resistivity versus temperature for phosphorus-doped silicon. Solid lines denote theoretical calculations and solid dots are the experimental data. After Li[2].

in Figures 7.6 to 7.11.[2–5] Figure 7.6 shows resistivity as a function of temperature for n-type silicon doped with different phosphorus impurity densities (N_D varying from 1.2×10^{14} to 2.5×10^{18} cm^{-3}).(2) Figure 7.7 shows the resistivity versus temperature for boron-doped silicon with boron impurity densities varying from 4.5×10^{14} to 3.2×10^{18} cm^{-3}.[3] Excellent agreement between theoretical calculations (solid lines) and experimental data (solid dots) was obtained in both cases. Figures 7.8a and b show resistivity versus dopant density for both n- and p-type silicon at 300 K.[2,3] The solid line shown in Figure 7.8a represents the theoretical calculations given by this author, while the dashed line corresponds to the experimental data compiled by Irvine.[4] In Figure 7.8b, the solid line represents the theoretical calculations by this author and the dashed line by these Irvine, while the broken line denotes the experimental data reported by Wagner.[5]

The resistivity and Hall effect measurements are often used in determining the carrier density and mobility in a semiconductor. The mobility determined from the product of electrical conductivity and Hall coefficient is known as the Hall mobility, which is the majority carrier mobility. The drift mobility determined by the Haynes–Shockley experiment is usually referred to as the minority carrier

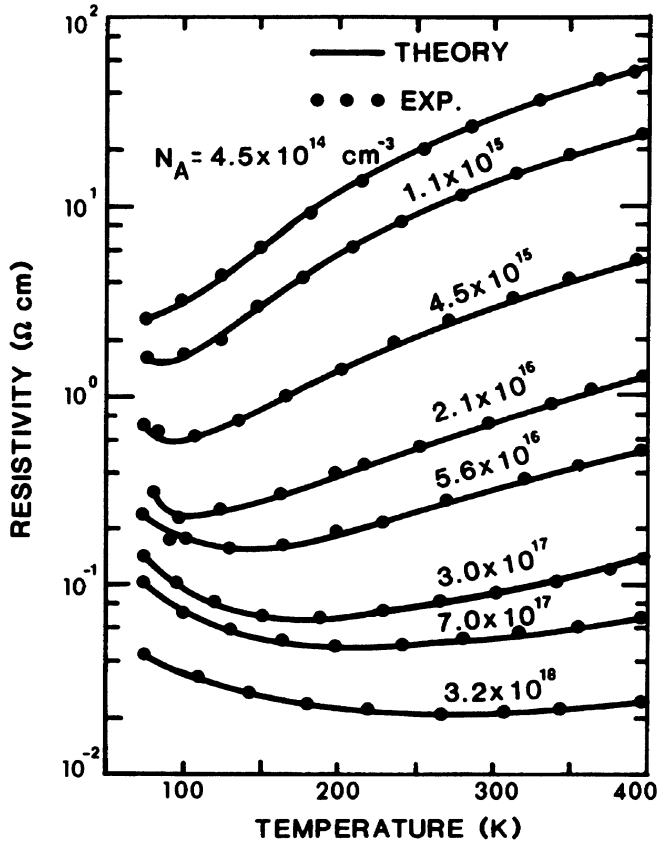


FIGURE 7.7. Resistivity versus temperature for boron-doped silicon. Solid lines denote theoretical calculations and solid dots denote experimental data. After Li[3].

mobility. These two quantities may or may not be equal, depending on the scattering processes and the dopant density of the semiconductor. The ratio of the Hall mobility and the conductivity mobility is equal to the Hall factor. Values of the Hall factor may vary between 1 and 1.93, depending on the types of scattering mechanisms involved in a semiconductor.

Figures 7.9 and 7.10 show the Hall mobility as a function of temperature for both n- and p-type silicon specimens with dopant density as a parameter, respectively.[6] The empirical formulas for the temperature dependence of Hall mobility for both n- and p-type silicon are given respectively by

$$\mu_{\mathrm{Hn}} = 5.5 \times 10^{6} T^{-3/2} \quad \text{and} \quad \mu_{\mathrm{Hp}} = 2.4 \times 10^{8} T^{-2.3}. \tag{7.118}$$

Equation (7.118) is valid for $T > 100$ K and $N_I < 10^{17}$ cm^{-3}.

Figures 7.11a and b show the electron and hole conductivity mobilities as a function of dopant density for n- and p-type silicon, respectively, at $T = 300$ K.[1]

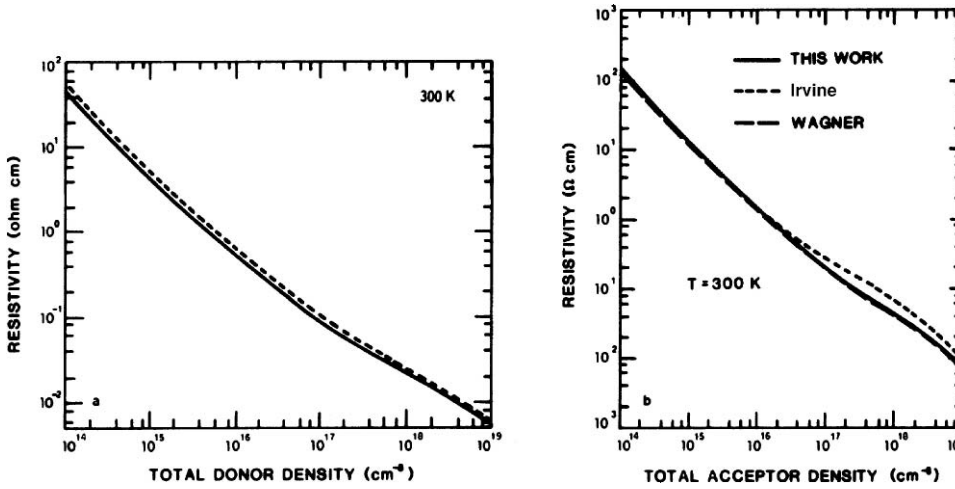


FIGURE 7.8. Resistivity versus dopant density for (a) n-type and (b) p-type silicon at 300 K. Solid lines correspond to calculated values by Li,[2,3] the dashed line data published by Irvine,[4] and the broken line data by Wager;[5] solid dots correspond to experimental data by Li.[2,3]

The experimental data are deduced from the resistivity and junction capacitance–voltage (CV) measurements on a specially designed test structure developed at the National Bureau of Standards for accurate determination of the conductivity mobility in silicon. The solid line corresponds to the theoretical calculations reported by this author using a more rigorous theoretical model.[2,3] The model takes into account all the scattering mechanisms contributed by acoustical and optical phonons, as well as scatterings due to ionized and neutral impurities. Furthermore, the intervalley and intravalley phonon scatterings and the effect of the nonparabolic band structure of silicon have also been taken into account in the calculations. The results show excellent agreement between theory and experiment for both n- and p-type silicon over a wide range of dopant densities and temperatures.

The Seebeck coefficient data for silicon and germanium are discussed next. The Seebeck coefficient for a nondegenerate n-type semiconductor is given by (7.77). If the acoustical phonon scattering is dominant, then the Seebeck coefficient for both n- and p-type semiconductors can be expressed by

$$S_{n,p} = \pm \left(\frac{k_B}{q}\right)\left(2 - \frac{E_f}{k_B T}\right), \tag{7.119}$$

where the plus sign is for p-type conduction and the minus sign for n-type conduction. Equation (7.119) shows that the Seebeck coefficient is directly related to the Fermi energy in the semiconductor. By measuring the Seebeck coefficient as a function of temperature, the Fermi level can be determined at different temperatures.

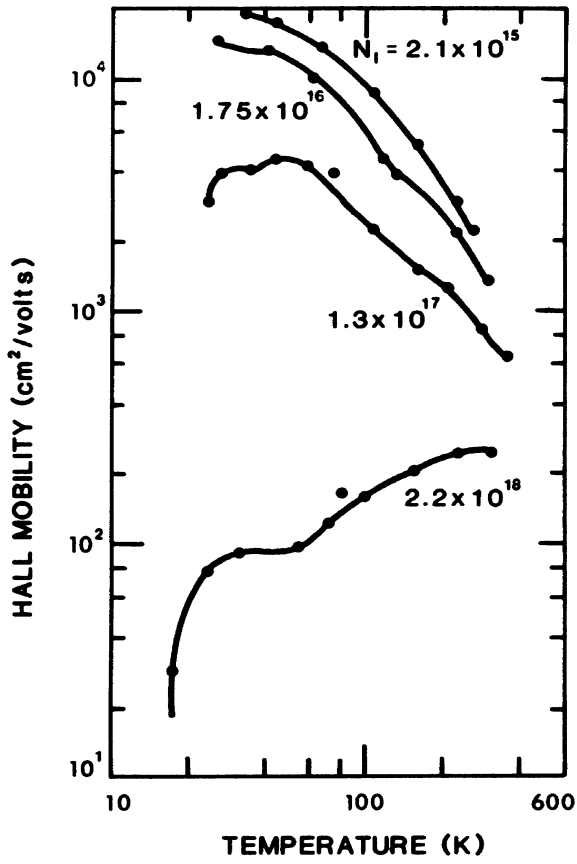


FIGURE 7.9. Hall mobility versus temperature for the As-doped silicon samples. After Morin and Maita,[6] by permission.

Figure 7.12a shows the Seebeck coefficient (thermoelectric power) versus temperature for n-type germanium with different resistivities,[7] and Figure 7.12b shows the Seebeck coefficient as a function of temperature for n- and p-type silicon.[8] As is clearly shown in Figure 7.12a, the Seebeck coefficient increases with increasing resistivity (or decreasing doping density) in n-type germanium. To obtain a large Seebeck coefficient (or large Peltier coefficient) in a semiconductor, it is necessary to use lightly doped semiconductors for the thermoelectric cooling or power generation elements. For p-type silicon, the Seebeck coefficient changes sign from positive to negative at high temperatures because of mixed conduction and becomes constant at the onset of the intrinsic regime. It is noted in Figure 7.12a that the measured and calculated Seebeck coefficients (dashed lines) for germanium are in good agreement for low-resistivity samples for $T > 200$ K. However, the Seebeck coefficients (curves B, C, and D shown in Figure 7.12a) increase rapidly with decreasing temperature for lightly doped samples for $T < 100$ K. Such

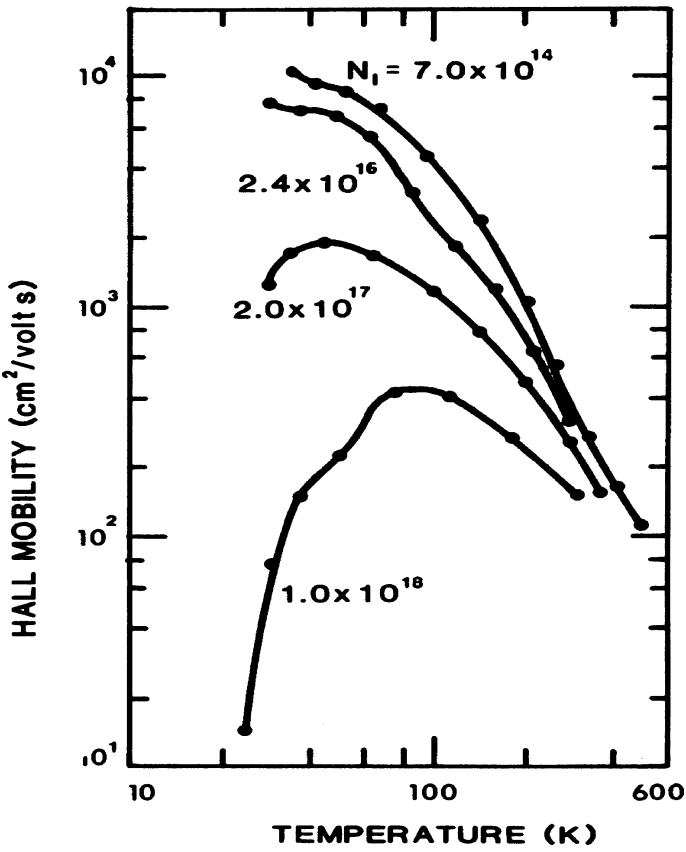


FIGURE 7.10. Hall mobility versus temperature for p-type silicon doped with different boron densities. After Morin and Maita,[6] by permission.

behavior cannot be explained by the theoretical expression derived above, and the so-called phonon drag effect must be considered in order to explain this anomalous behavior. Phonon drag has a striking effect on the Seebeck coefficient, particularly at low temperatures. This effect can be explained if one assumes that the flow of long-wavelength phonons in the presence of a temperature gradient leads to preferential scattering of electrons in the direction of the temperature gradient. It can be shown that the phonon-drag Seebeck coefficient can be expressed as

$$S_{\mathrm{pd}} = \pm \left(\frac{x \nu^2 \tau_d}{\mu T} \right), \tag{7.120}$$

where the minus sign denotes n-type conduction and the plus sign p-type conduction, x is the fraction of carrier collisions due to phonons, and τ_{d} is the relaxation time for loss of momentum from the phonon system.

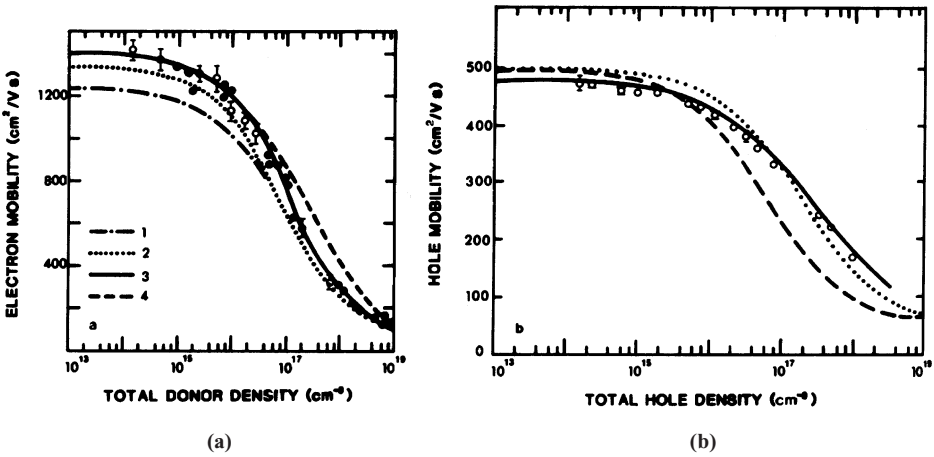


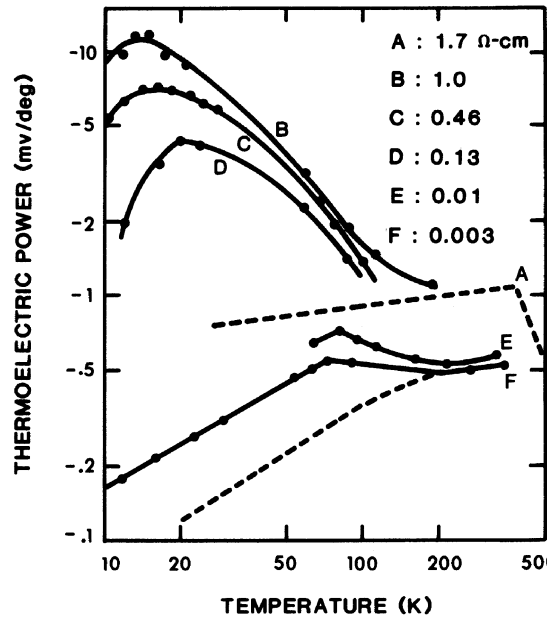
FIGURE 7.11. Electron and hole mobility versus dopant density for (a) n-type and (b) p-type silicon doped with phosphorus and boron impurities at $T = 300$ K, respectively. After Li.[2,3].

It is noted that the phonon-drag Seebeck coefficient has the same sign as the Seebeck coefficient in the absence of the phonon-drag effect. Therefore, the electron and phonon contributions to the Seebeck coefficient reinforce one another at low temperatures. It is seen from (7.77) that the Seebeck coefficient in a nondegenerate semiconductor can be quite large, on the order of a few mV/K, while for metals the Seebeck coefficient is on the order of a few tens of μV/K. Thus, at a metal–semiconductor junction, the total Seebeck coefficient is approximately equal to the absolute Seebeck coefficient of the semiconductor.

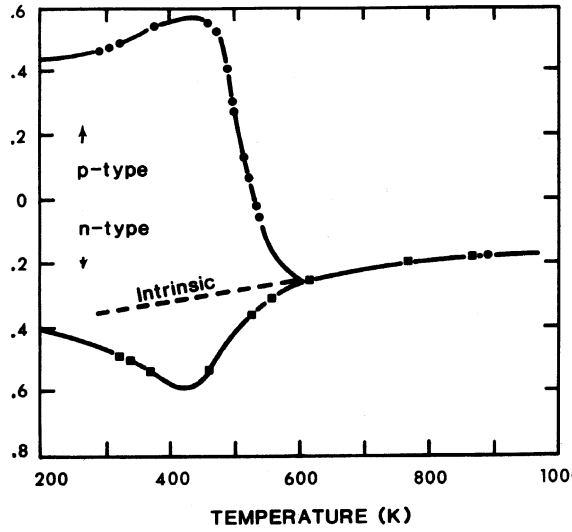
Transverse magnetoresistance for silicon and germanium is discussed next. The change in resistivity as a result of an applied magnetic field is usually referred to as the magnetoresistance effect. The magnetoresistance is defined by

$$\frac{(\rho - \rho_0)}{\rho_0} = \frac{(\sigma_0 - \sigma)}{\sigma_0}. \tag{7.121}$$

Early measurements of magnetoresistance in germanium and silicon were made on polycrystalline materials. Magnetoresistance measurements on single-crystal silicon samples by Pearson and Herring have produced some interesting results. The effect was found to depend not only on the relative orientations of the current and magnetic field, but also on the crystal orientations. This is illustrated in Figure 7.13 for an n-type silicon sample.[9] The results show that the theoretical derivation of magnetoresistance based on the assumption that the constant-energy surface is spherical and that σ depends only on the carrier energy are inadequate for the case of silicon. This is due to the fact that the constant-energy surfaces in the conduction band minimum and valence band maximum of silicon are not exactly spherical. Therefore, refinement



(a) n-Ge



(b) n- and p-Si

FIGURE 7.12. Seebeck coefficient as a function of temperature: (a) for n-type germanium and (b) for n- and p-type silicon. After Frederikse,[7] and Geballe and Hull,[8] by permission.

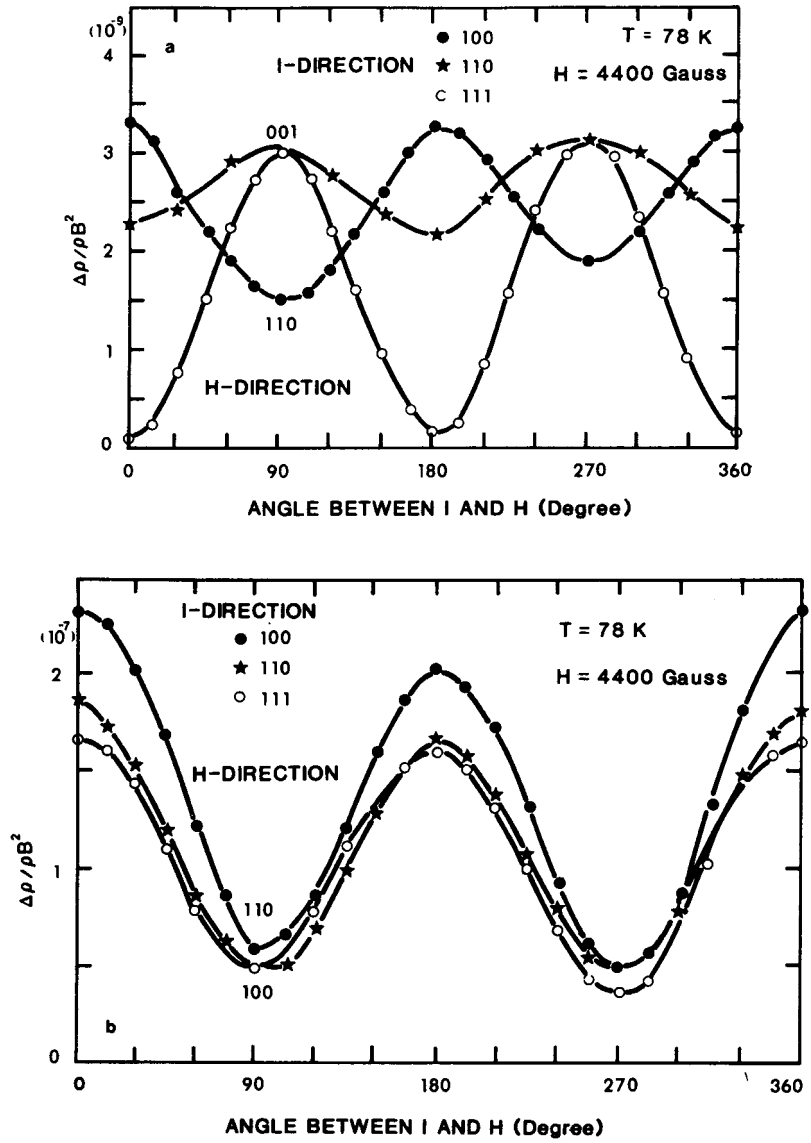


FIGURE 7.13. Variation of $\Delta\rho/\rho B^2$ as the magnetic field H is rotated with respect to the current flow I for (a) n-type and (b) p-type silicon samples. After Pearson and Herring,[9] by permission.

of the transport theories presented in this chapter by taking into account various effects cited above is needed in order to obtain an accurate prediction of the experimental results of different transport coefficients in semiconductor materials.

Problems

7.1. (a) Show that if an electric field is applied in the x-direction, the steady-state nonequilibrium distribution function can be expressed by

$$f(k_x, k_y, k_z) = f_0((k_x - q\mathcal{E}_x\tau/\hbar), k_y k_z),$$

where f_0 is the equilibrium distribution function.

(b) If a temperature gradient is applied in the x-direction, show that the steady-state nonequilibrium distribution function is given by

$$f(k_x, k_y, k_z) = f_0(k_x + \Delta k_x, k_y, k_z),$$

where $\Delta k_x = (\tau\hbar k_f/m^*T)(k - k_f)(dT/dx)$, and k_f is the wave vector of electrons at the Fermi level.

7.2. (a) Plot σ_n and S_n versus η (the reduced Fermi energy) for $-4 \leq \eta \leq 4$, assuming $s = -1/2$ and $\tau = \tau_0 E^s$.

(b) Repeat (a) for $s = +3/2$.

7.3. Using (7.53) and the Fermi–Dirac statistics, show that the electrical conductivity for a degenerate n-type semiconductor is given by

$$\sigma_n = \left(\frac{2nq^2\tau_0}{3m_n^*}\right)(k_B T)^s (s + 3/2)\frac{F_{(s+1/2)}}{F_{(1/2)}},$$

where

$$\tau = \tau_0 E^s$$

and

$$F_r(\eta) = \int_0^\infty \frac{\varepsilon^r \, d\varepsilon}{[1 + e^{(\varepsilon - \eta)}]},$$

where $F_r(\eta)$ is the Fermi integral of order r, $\varepsilon = E/k_B T$, and $\eta = E_f/k_B T$.

7.4. Show that the Seebeck coefficient for a degenerate n-type semiconductor can be expressed by

$$S_n = -\left(\frac{k_B}{q}\right)\left[\frac{(s + 5/2)F_{(s+3/2)}}{(s + 3/2)F_{(s+1/2)}} - \frac{E_f}{k_B T}\right].$$

7.5. If an electric current and a temperature gradient are applied simultaneously to an n-type semiconductor specimen in the x-direction, show that the electric current density and the heat flux density can be expressed by

$$\begin{aligned} J_x = -nqv_x &= -\int_0^\infty q v_x f(E) g(E) \, dE \\ &= -\left(\frac{2q}{3m_n^*}\right)\int_0^\infty \tau E g(E)\frac{\partial f_0}{\partial E}\left[q\mathcal{E}_x - \frac{(E_f - E)}{T}\frac{\partial T}{\partial x}\right] dE \end{aligned}$$

and

$$Q_x = -\left(\frac{2q}{3m_n^*}\right)\int_0^\infty \tau E^2 g(E)\frac{\partial f_0}{\partial E}\left[q\mathcal{E}_x - \frac{(E_f - E)}{T}\frac{\partial T}{\partial x}\right] dE.$$

(*Hint:* Solve $P(E)$ from (7.38) with $B = 0$.)

7.6. (a) Using the expressions given in Problem 7.5, show that the electronic thermal conductivity for a degenerate n-type semiconductor can be expressed by

$$K_{\mathrm{n}} = -\frac{Q_x}{(\mathrm{d}T/\mathrm{d}x)} = \left(\frac{n}{m^* T}\right)[\langle \tau E^2 \rangle - \langle \tau E \rangle^2 / \langle \tau \rangle].$$

(b) Show that for the nondegenerate case, the expression given by (a) can be simplified to

$$K_{\mathrm{n}} = \left(\frac{n\tau_0}{m^* T}\right)(k_{\mathrm{B}} T)^{s+2} \frac{\Gamma_{(7/2-s)}}{\Gamma_{(5/2)}}.$$

7.7. Show that the longitudinal magnetoresistance effect will vanish if the constant-energy surface of the conduction bands is spherical.

7.8. Using (7.65) and (7.66) and Fermi–Dirac statistics, derive the Hall coefficient for a degenerate n-type semiconductor and show that the result can be reduced to (7.70) if the M-B statistics are used instead. Derive the Hall factor for a degenerate n-type semiconductor.

7.9. If the total electron mobility of an n-type semiconductor is obtained using the reciprocal sum of the lattice scattering mobility and the ionized impurity scattering mobility (i.e., $\mu_{\mathrm{n}}^{-1} = \mu_{\mathrm{L}}^{-1} + \mu_{\mathrm{I}}^{-1}$), where $\tau_{\mathrm{L}} = aT^{-1}E^s$, $\tau_I = bE^s$, and a and b are constants, derive an expression for the total electron mobility when both the lattice and ionized impurity scatterings are dominated in this semiconductor.

7.10. Show that the Seebeck coefficient for the mixed conduction case is given by (7.116). If the lattice scattering is dominant, derive the Hall and Seebeck coefficients from (7.104) and (7.116) for a nondegenerate n-type semiconductor.

References

1. P. P. Debye and E. M. Conwell, *Phys. Rev.* **93**, 693 (1954).
2. S. S. Li, *The Dopant Density and Temperature Dependence of Electron Mobility and Resistivity in n-Type Silicon*, NBS Special Publication, 400–33 (1977). See also, S. S. Li and R. W. Thurber, *Solid-State Electron*. **20**, 609–616 (1977).
3. S. S. Li, *The Theoretical and Experimental Study of the Temperature and Dopant Density Dependence of Hole Mobility, Effective Mass, and Resistivity in Boron-Doped Silicon*, NBS Special Publication, 400–47 (1979). See also, S. Li, *Solid-State Electron*. **21**, 1109–1117 (1978).
4. J. C. Irvine, *Bell Syst. Tech. J.* **16**, 387 (1962).
5. S. Wagner, *J. Electrochem. Soc.* **119**, 1570 (1972).
6. F. T. Morin and J. P. Maita, *Phys. Rev.* **96**, 28 (1954).
7. H. P. R. Frederikse, *Phys. Rev.* **92**, 248 (1953).
8. T. H. Geballe and G. W. Hull, *Phys. Rev.* **98**, 940 (1955).
9. G. L. Pearson and C. Herring, *Physica* **20**, 975 (1954).

Bibliography

F. J. Blatt, *Physics of Electronic Conduction in Solids*, McGraw-Hill, New York (1968).
R. H. Bube, *Electronic Properties of Crystalline Solids*, Academic Press, New York (1974).
M. Dresden, *Rev. Mod. Phys.* **33**, 265 (1961).
A. F. Gibson and R. E. Burgess, *The Electrical Conductivity of Germanium*, Wiley, New York (1964).
G. L. Pearson and J. Bardeen, *Phys. Rev.* **75**, 865 (1949).
R. A. Smith, *Semiconductors*, Cambridge University Press, London (1960).
R. K. Willardson and A. C. Beer, *Transport Phenomena, Semiconductors and Semi-Metals*, Vol. 10, Academic Press, New York (1975).
A. H. Wilson, *The Theory of Metals*, Cambridge University Press, London (1954).
J. M. Ziman, *Electrons and Phonons*, Oxford University Press, London (1960).

8
Scattering Mechanisms and Carrier Mobilities in Semiconductors

8.1. Introduction

The relaxation time approximation introduced in Chapter 7 enables one to linearize the Boltzmann transport equation in that the collision term is expressed in terms of the ratio of the perturbed distribution function (i.e., $f - f_0$) and the relaxation time. This approximation allows one to obtain analytical expressions for different transport coefficients in semiconductors. However, detailed physical insights concerning the collision term and the validity of the relaxation time approximation were not discussed in Chapter 7. In this chapter, various scattering mechanisms associated with the collision term in the Boltzmann equation will be described, and the relaxation time constants due to different scattering mechanisms in a semiconductor will be derived.

The collision term in the Boltzmann transport equation represents the internal relaxation mechanisms, which are related to the collision of charged carriers (electrons or holes) with different scattering sources (e.g., scattering of electrons by acoustical phonons and ionized impurity) in a semiconductor under the influence of external forces. These scattering mechanisms are responsible for the charged carriers reaching steady-state conditions when external forces are applied to the semiconductor, and returning to equilibrium conditions when the external forces are removed from the semiconductor. In elastic scattering, the nonequilibrium distribution function will decay exponentially with time to its equilibrium value after the external force is removed. The time constant associated with this exponential decay is known as the relaxation time or the collision time.

In this chapter, several important scattering mechanisms such as acoustical phonon and optical phonon scatterings, ionized impurity scattering, and neutral impurity scattering, which play a key role in determining the carrier mobilities in a semiconductor, will be considered. Using quantum-mechanical treatments, the relaxation time expressions for these scattering mechanisms can be derived. In Section 8.2, the collision term is expressed in terms of the rate of transition probability and the distribution functions for the initial and final states in k-space.

The differential scattering cross section, which is defined in terms of the rate of transition probability and the incident flux of the scattering charged carriers, is also introduced in this section. Using the Brooks–Herring (B-H) model, the relaxation time for ionized impurity scattering is derived in Section 8.3. Section 8.4 describes neutral impurity scattering, which is an important scattering source at very low temperatures or at very high doping densities. Using deformation potential theory, the scattering of charge carriers by longitudinal-mode acoustical phonons is derived in Section 8.5. The scatterings of charge carriers by polar and nonpolar optical phonons in compound semiconductors as well as intervalley optical phonon scattering in a multivalley semiconductor are discussed in Section 8.6. The scattering of charge carriers by dislocations is described in Section 8.7. Finally, the measured Hall mobilities and drift mobilities for some elemental and compound semiconductors are presented in Section 8.8.

In general, the charge carriers in a semiconductor may be scattered by stationary defects (e.g., impurities and dislocations) and/or by dynamic defects (e.g., electrons, holes, and lattice phonons). Therefore, the transport properties of a semiconductor depend strongly on the types of scattering mechanisms involved. For example, the electrical conductivity of an n-type semiconductor can be expressed in terms of the electron mobility and electron concentration by

$$\sigma_{\mathrm{n}} = n_0 q \mu_{\mathrm{n}}, \tag{8.1}$$

where n_0 is the electron concentration, q is the electronic charge, and μ_{n} is the electron mobility. The electron mobility may be defined in terms of the conductivity effective mass m_{c}^* and the relaxation time τ by

$$\mu_n = \frac{q\langle\tau\rangle}{m_{\mathrm{c}}^*}, \tag{8.2}$$

where $\langle\tau\rangle$ is the average relaxation time defined by (7.54). Thus, the electron mobility is directly proportional to the average relaxation time and varies inversely with the conductivity effective mass. Since the average relaxation time given in (8.2) is directly related to the scattering mechanisms, in order to calculate the carrier mobility, it is necessary first to consider the scattering mechanisms in a semiconductor.

In the relaxation time approximation, the collision term in the Boltzmann equation can be expressed in terms of the perturbed distribution function divided by the relaxation time. From (7.25), one obtains

$$\left.\frac{\partial f}{\partial t}\right|_{\mathrm{c}} = -\frac{f - f_0}{\tau}\ , \tag{8.3}$$

where f is the nonequilibrium distribution function and f_0 is the equilibrium Fermi–Dirac distribution function. As mentioned earlier, the relaxation time approximation is valid only for the elastic-scattering case. This condition is satisfied as long as the change in energy of the charge carriers before and after each scattering event is small compared to the initial carrier energy. In fact, a generalized expression for the collision term given by (8.3) can be formulated in terms of the rate

of transition probability $P_{kk'}$ and the nonequilibrium distribution function $f(k, r)$, which is given by

$$-\left.\frac{\partial f_{k'}}{\partial t}\right|_{\mathrm{c}} = \sum_{k'} [P_{kk'} f_{k'} (1 - f_k) - P_{k'k} f_k (1 - f_{k'})], \tag{8.4}$$

where $P_{kk'}$ is the rate of transition probability from the final state k' to the initial state k, and $P_{k'k}$ is the rate of transition probability from the k-state to the k'-state. The electron distribution function in the k'-state is designated by $f_{k'}$, and the electron distribution function in the k-state is represented by f_k.

The right-hand side of (8.4) represents the net transition rates from the k- to the k'-state summed over all the final states k'. The summation in (8.4) can be replaced by integration over the entire conduction band if all the quantum states in the band are treated as a quasicontinuum. Since the density of quantum states in the conduction band is very large and the spacing between each quantum state is very small, such an assumption is usually valid. Therefore, it is a common practice to replace the summation in (8.4) by an integral, which can be written as

$$\begin{aligned} -\left.\frac{\partial f_{k'}}{\partial t}\right|_{\mathrm{c}} &= \frac{N\Omega}{(2\pi)^3} \int [P_{kk'} f_{k'} (1 - f_k) - P_{k'k} f_k (1 - f_{k'})] \, \mathrm{d}^3 k' \\ &= \frac{N\Omega}{(2\pi)^3} \int P_{kk'} (f_{k'} - f_k) \, \mathrm{d}^3 k', \end{aligned} \tag{8.5}$$

where $P_{kk'}$ is assumed equal to $P_{k'k}$, N is the total number of unit cells in the crystal, and Ω is the volume of the unit cell.

It is noted that the collision term given by (8.5) is a differential integral equation and cannot be solved analytically without further approximations. In order to derive an analytical expression for (8.5), it is useful to first consider the small-perturbation case (i.e., the low-field case). In this case, the nonequilibrium distribution function $f(k, r)$ can be expressed in terms of the equilibrium distribution function f_k^0 and a first-order perturbing distribution function f_k^1, which reads

$$\begin{aligned} f_k &= f_k^0 + f_k^1 + \cdots, \\ f_{k'} &= f_{k'}^0 + f_{k'}^1 + \cdots, \end{aligned} \tag{8.6}$$

where f_k^0 and $f_{k'}^0$ are the Fermi–Dirac distribution functions in the k- and k'-states, while f_k^1 and $f_{k'}^1$ denote the first-order correction terms of the distribution functions in the k- and k'-states, respectively.

If one assumes that the scattering is elastic, then the energy change during scattering processes is small compared to the average electron energy. Under this condition, the average energy of electrons in the initial and final states can be assumed equal (i.e., $E_k = E_{k'}$). Therefore, the equilibrium distribution functions for the initial and final states are identical, and the collision term can be simplified to

$$-\left.\frac{\partial f_{k'}}{\partial t}\right|_{\mathrm{c}} = \frac{f_{k'}^1}{\tau} = \frac{N\Omega}{(2\pi)^3} \int P_{kk'} \left(f_{k'}^1 - f_k^1\right) \mathrm{d}^3 k'. \tag{8.7}$$

Thus, the inverse relaxation time τ^{-1} can be written as

$$\frac{1}{\tau} = \frac{N\Omega}{(2\pi)^3} \int P_{kk'} \left(1 - \frac{f_k^1}{f_{k'}^1}\right) \mathrm{d}^3k'. \tag{8.8}$$

Furthermore, if one assumes that the scattering process is isotropic, then the ratio of f_k^1 and $f_{k'}^1$ can be expressed in terms of $\cos\theta'$, where θ' is the angle between the incident wave vector k and the scattered wave vector k' (see Figure 8.2b). Under this condition, (8.8) becomes

$$\frac{1}{\tau} = \frac{N\Omega}{(2\pi)^3} \int P_{kk'} \left(1 - \cos\theta'\right) \mathrm{d}^3k'. \tag{8.9}$$

Equation (8.9) shows that the scattering rate τ^{-1} of the charge carriers for isotropic elastic scattering depends only on the angle θ' between the k- and k'-states and the rate of transition probability $P_{kk'}$.

In order to derive the relaxation time for a specific scattering process, both the rate of transition probability and the differential scattering cross-section must be determined first. This is discussed next.

8.2. Differential Scattering Cross-Section

In the present treatment, it is assumed that the scattering of charge carriers is confined within a single energy band (e.g., electrons in the conduction band and holes in the valence band), as illustrated in Figure 8.1a. Other important scattering processes such as intervalley scattering for multivalley semiconductors such as Si and Ge and interband scattering in the heavy-hole and light-hole bands are also shown in Figure 8.1b and c, respectively.

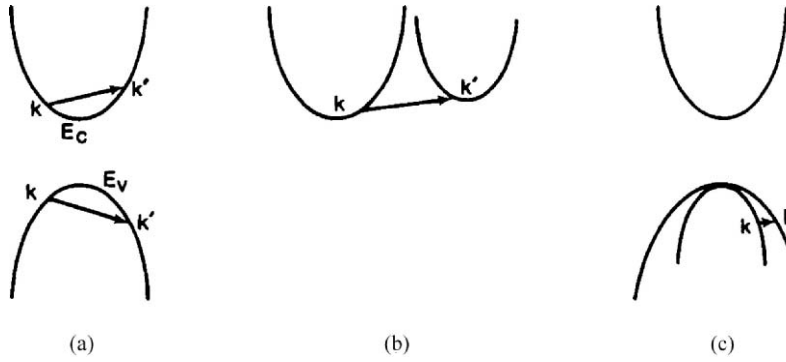


FIGURE 8.1. Scattering of electrons in the k-space of a semiconductor: (a) The intravalley scattering of electrons in the conduction band and the intraband scattering of holes in the valence band, (b) intervalley scattering of electrons in the conduction band, and (c) interband scatterng of holes in the valence bands; k denotes the wave vector of incident electrons and k' the wave vector of scattered electrons in k-space.

The intraband and intravalley scatterings shown in Figure 8.1a are usually accompanied by the absorption or emission of a longitudinal-mode acoustical phonon, and hence can be considered as elastic scattering. However, the interband and intervalley scatterings shown in Figure 8.1b and c are usually inelastic because the change in electron energy for these scatterings is no longer small compared to the average electron or hole energy. The intervalley and interband scattering processes are usually accompanied by the absorption or emission of optical phonons, which occur at high temperatures or high electric fields.

The rate of transition probability $P_{kk'}$ in a scattering event can be derived from the one-electron Schrödinger equation. The one-electron time-independent Schrödinger equation for the initial unperturbed states is given by

$$H_0\phi_k(r) = E_k\phi_k(r), \tag{8.10}$$

where

$$H_0 = -\frac{\hbar^2\nabla^2}{2m^*} + V(r). \tag{8.11}$$

Here H_0 is the unperturbed Hamiltonian, and $\phi_k(r)$ is the initial unperturbed electron wave function given by

$$\phi_k(r) = u_k(r)\,\mathrm{e}^{ik,r}, \tag{8.12}$$

where $u_k(r)$ is the Bloch function, which has the same periodicity as the crystal potential $V(r)$.

When a small perturbation (e.g., a small electric field) is applied to the crystal, the electron may be scattered from the initial state k into the final state k'. The perturbed Hamiltonian under this condition can be written as

$$H = H_0 + H', \tag{8.13}$$

where H_0 is the unperturbed Hamiltonian given by (8.11), and H' is the first-order correction due to perturbation. The time-dependent Schrödinger equation under the perturbed condition is given by

$$H\psi_k(r,t) = -i\hbar\frac{\partial\psi_k(r,t)}{\partial t}, \tag{8.14}$$

which has a solution given by

$$\psi_k(r,t) = \sum_k a_k(t)\,\mathrm{e}^{-iE_kt/\hbar}\phi_k(r), \tag{8.15}$$

where $a_k(t)$ is the time-dependent amplitude function, and $\phi_k(r)$ is the unperturbed electron wave function defined by (8.12).

According to time-dependent perturbation theory, the transition probability per unit time from the k- to the k'-state can be expressed in terms of the amplitude function $a_k(t)$ by

$$P_{k'k}(t) = \frac{|a_k(t)|^2}{t}. \tag{8.16}$$

Similarly, the transition probability per unit time from the k'- to the k-state is given by

$$P_{kk'}(t) = \frac{|a_{k'}(t)|^2}{t}. \tag{8.17}$$

From the principle of detailed balance, one can assume that $P_{kk'} = P_{k'k}$. Using (8.14) through (8.17) and the orthogonal properties of electron wave functions, it can be shown from quantum-mechanical calculations that the rate of transition probability $P_{kk'}$ in the presence of a step perturbation function (i.e., a constant H') is given by

$$P_{kk'} = \frac{|a_{k'}(t)|^2}{t} = \frac{2\pi}{\hbar} |H_{kk'}|^2 \, \delta(E_{k'} - E_k), \tag{8.18}$$

where

$$H_{kk'} = \langle k' | H' | k \rangle = \frac{1}{(N\Omega)} \int_{N\Omega} \phi_{k'}^* H' \phi_k \, \mathrm{d}^3 r \tag{8.19}$$

is the matrix element. In (8.19), H' is the perturbing Hamiltonian, ϕ_k is the electron wave function given by (8.12), and $\phi_{k'}^*$ is the complex conjugate of $\phi_{k'}$. The function $\delta(E_{k'} - E_k)$ is the Dirac delta function, which is equal to unity for $E_k = E_{k'}$ and vanishes otherwise.

The matrix element $H_{kk'}$, given by (8.19), has a finite value only if the golden selection (momentum conservation) rule is satisfied (i.e., $k = k'$ for a direct transition and $k' = k \pm q$ for an indirect transition). Calculations of relaxation time can be simplified by introducing a differential scattering cross-section $\sigma\ (\theta', \phi')$ in the relaxation time formula. It is noted that $\sigma(\theta', \phi')$ depends only on θ' if the scattering process is isotropic (i.e., independent of ϕ'). Under this condition, a simple relationship exists between $\sigma(\theta')$ and the rate of transition probability $P_{kk'}$. In general, the differential scattering cross-section $\sigma(\theta', \phi')$ is defined as the total number of particles that make transitions from the k- to the k'-state per unit solid angle per unit time divided by the incident flux density. This can be written as

$$\sigma(\theta', \phi') = \frac{\dfrac{N\Omega}{(2\pi)^3} P_{kk'} \dfrac{\mathrm{d}^3 k'}{\mathrm{d}\omega}}{\dfrac{v_k}{N\Omega}} = \frac{(N\Omega)^2 P_{kk'} \mathrm{d}^3 k'}{(2\pi)^3 v_k \sin\theta' \mathrm{d}\theta' \mathrm{d}\phi'}, \tag{8.20}$$

where v_k is the initial particle velocity, $N\Omega$ is the volume of the crystal, and $\mathrm{d}\omega = \sin\theta' \, \mathrm{d}\theta' \, \mathrm{d}\phi'$ is the solid angle between the incident wave vector k and the scattered wave vector k' (see Figure 8.2).

Now consider the case of isotropic elastic scattering. Substituting (8.18) and (8.19) into (8.20), and using the relationships $v_k = v_{k'}$, $k = k'$, and $\mathrm{d}^3 k' = k'^2 \sin\theta' \, \mathrm{d}\theta' \, \mathrm{d}\phi' \, \mathrm{d}k'$, one can obtain an expression for the differential scattering cross-section, which is given by

$$\sigma(\theta') = \frac{(N\Omega)^2 k'^2 |H_{kk'}|^2}{(2\pi\hbar v_{k'})^2}. \tag{8.21}$$

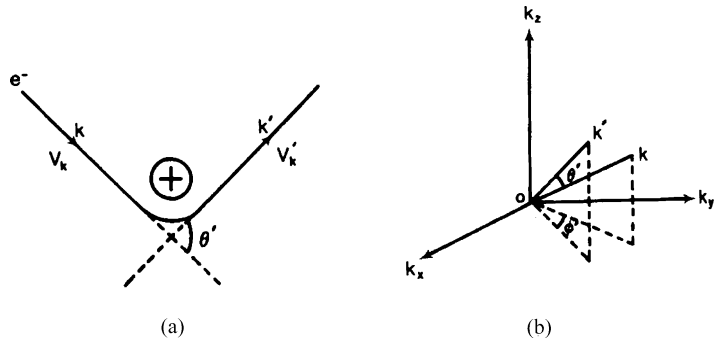


FIGURE 8.2. Scattering of electrons by a positively charged shallow-donor impurity atom in k-space.

The relaxation time τ is defined in terms of the total scattering cross-section by

$$\frac{1}{\tau} = N_{\mathrm{T}} \sigma_{\mathrm{T}} v_{\mathrm{th}}, \tag{8.22}$$

where N_{T} is the density of total scattering centers, σ_{T} is the total scattering cross-section, and v_{th} is the mean thermal velocity $[v_{\mathrm{th}} = (3k_{\mathrm{B}}T/m^*)^{1/2}]$. The total scattering cross-section (σ_{T}) for the isotropic elastic scattering process can be calculated from (8.23) using the differential scattering cross-section given by (8.21), which can be expressed by

$$\sigma_{\mathrm{T}} = 2\pi \int_0^{\pi} \sigma(\theta')(1 - \cos\theta') \sin\theta' \, \mathrm{d}\theta'. \tag{8.23}$$

Substituting (8.21) into (8.23), the total scattering cross-section can be calculated from (8.23), provided that the perturbing Hamiltonian H', and hence the matrix element $H_{kk'}$, is known. In the following sections, (8.20) through (8.23) will be used to derive the expressions of relaxation time constants and carrier mobilities for a semiconductor in which scatterings of electrons or holes are due to the ionized impurities, neutral impurities, or the longitudinal mode acoustical phonons.

8.3. Ionized Impurity Scattering

Scattering of electrons by ionized shallow-donor impurities is a classical example of elastic scattering in a semiconductor. This is due to the fact that the mass of a shallow-donor impurity atom is much larger than that of an electron. As a result, the change of electron energy during such a scattering process is negligible compared to the electron energy before the scattering. Therefore, the relaxation time approximation given by (8.22) is valid in this case. In order to derive the differential scattering cross-section and the relaxation time for the ionized impurity scattering, the matrix element $H_{kk'}$ and the perturbing Hamiltonian H' due to the

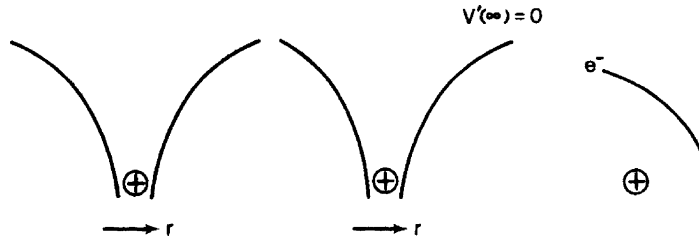


FIGURE 8.3. Potential due to a positively charged shallow-donor impurity atom: (a) bare Coulomb potential, (b) screening Coulomb potential, and (c) trajectory of electron scattering by a positively charged ion.

shallow-donor impurity potential must first be determined. Let us consider the scattering of electrons by a positively charged shallow-donor impurity in an n-type semiconductor, as shown in Figure 8.2a. If the donor impurity is ionized with a single net positive charge, then the potential due to this ionized donor atom, at a large distance from the impurity atom, can be approximated by a bare Coulomb potential

$$V(r) = \frac{q}{4\pi\varepsilon_0\varepsilon_s r}. \tag{8.24}$$

It should be noted that (8.24) did not consider the Coulomb screening effect due to electrons from the rest of positively charged donor ions in the semiconductor. To take into account the screening effect of these electrons, it is necessary to replace the bare Coulomb potential by a screening Coulomb potential in the derivation of ionized impurity scattering mobility. As shown in Figure 8.3b, if the screening effect of the shallow-donor ion by the surrounding conduction electrons is included, then the screening Coulomb potential (also known as the Yukawa potential) for the ionized impurity atom can be expressed by

$$V'(r) = \frac{q\mathrm{e}^{-r/\lambda_D}}{4\pi\varepsilon_0\varepsilon_s r}, \tag{8.25}$$

where

$$\lambda_D = \sqrt{\frac{\varepsilon_0\varepsilon_s k_B T}{q^2 n_0}} \tag{8.26}$$

is the Debye screen length.

In deriving the matrix element for ionized impurity scattering, Conwell and Weisskopf[1] used the bare Coulomb potential given by (8.24) as the perturbing Hamiltonian, while Brooks and Herring[2] employed the screening Coulomb (Yukawa) potential given by (8.25) as the perturbing Hamiltonian. It will be shown later that the relaxation-time formula derived from both models differs only by a constant but gives the same prediction concerning the energy dependence of

the relaxation time. Since the Brooks–Herring (B-H) model is based on the quantum mechanical principle, and is fundamentally much more sound and accurate than the Conwell–Weisskopf (C-W) model, it is pertinent here to use the B-H model for the derivation of ionized impurity scattering mobility in a semiconductor. The perturbing Hamiltonian due to the Yukawa potential, given by (8.25), can be written as

$$H' = qV'(r) = \frac{q^2 \mathrm{e}^{-r/\lambda_\mathrm{D}}}{4\pi\varepsilon_0\varepsilon_\mathrm{s} r}. \tag{8.27}$$

Based on the Bloch theorem, the electron wave functions for the k-state can be expressed by

$$\phi_k(r) = \left(\frac{1}{N\Omega}\right)^{1/2} u_k(r)\,\mathrm{e}^{ik\cdot r}. \tag{8.28}$$

The matrix element due to the Yukawa potential can be derived using (8.19), (8.27), and (8.28), with the result

$$\begin{aligned} H_{kk'} &= \frac{1}{N\Omega}\int \mathrm{e}^{-ik'\cdot r}\left(\frac{q^2\mathrm{e}^{-r/\lambda_\mathrm{D}}}{4\pi\varepsilon_0\varepsilon_\mathrm{s} r}\right)\mathrm{e}^{ik\cdot r}\,\mathrm{d}^3 r \\ &= \frac{q^2}{2N\Omega\varepsilon_0\varepsilon_\mathrm{s}}\int_0^\pi\int_0^\infty \mathrm{e}^{-iK\cdot r}\left(\frac{\mathrm{e}^{-r/\lambda_\mathrm{D}}}{r}\right) r^2\sin\theta_r\,\mathrm{d}\theta_r\,\mathrm{d}r \\ &= \frac{q^2\lambda_\mathrm{D}^2}{N\Omega\varepsilon_0\varepsilon_\mathrm{s}\left(1+K^2\lambda_\mathrm{D}^2\right)}, \end{aligned} \tag{8.29}$$

where $\mathrm{d}^3 r = 2\pi r^2 \sin\theta_r\,\mathrm{d}\theta_r\,\mathrm{d}r$ is the volume element, and

$$K = k' - k = 2|k|\sin\left(\frac{\theta'}{2}\right). \tag{8.30}$$

Here K is the reciprocal lattice vector. Figure 8.4 shows the relationship between the incident and scattered wave vectors k and k' in real space.

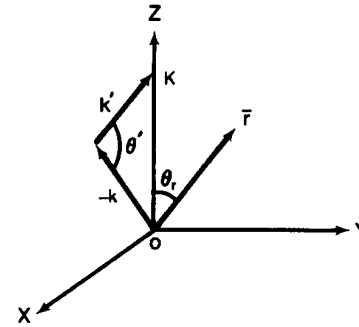


FIGURE 8.4. Coordinates for computing the matrix element of scattering by an ionized impurity, where k is the wave vector of the incident electron, k' is the wave vector of the scattered electron wave, and $K = k' - k = 2|k|\sin(\theta'/2)$ is the reciprocal lattice vector.

The differential scattering cross-section can be obtained by substituting Eq. (8.29) into (8.21), which yields

$$\sigma\left(\theta'\right) = \frac{\left(q^2 m^* \lambda_{\mathrm{D}}^2\right)^2}{\left(2\pi\hbar^2\varepsilon_0\varepsilon_{\mathrm{s}}\right)^2\left(1+K^2\lambda_{\mathrm{D}}^2\right)^2} = \frac{4\lambda_{\mathrm{D}}^4}{a_{\mathrm{B}}^2\left(1+K^2\lambda_{\mathrm{D}}^2\right)^2}, \tag{8.31}$$

where

$$a_{\mathrm{B}} = \frac{4\pi\varepsilon_0\varepsilon_{\mathrm{s}}\hbar^2}{m^* q^2} \tag{8.32}$$

is the Bohr radius for the ground state of the impurity atom.

The relaxation time for the ionized impurity scattering can be obtained by substituting (8.31) into (8.22) and (8.23), and the result is given by

$$\begin{aligned}\frac{1}{\tau_{\mathrm{I}}} &= (2\pi N_I v)\int_0^{\pi} \frac{4\lambda_{\mathrm{D}}^4(1-\cos\theta')\sin\theta'\,\mathrm{d}\theta'}{a_{\mathrm{B}}^2\left[1+4\lambda_{\mathrm{D}}^2k^2\sin^2(\theta'/2)\right]^2} \\ &= (2\pi N_{\mathrm{I}} v)\left(\frac{\lambda_{\mathrm{D}}^4}{a_{\mathrm{B}}^2}\right)\left(\frac{1}{k\lambda_{\mathrm{D}}}\right)^4 L(2k\lambda_{\mathrm{D}}),\end{aligned} \tag{8.33}$$

where

$$L(2k\lambda_{\mathrm{D}}) = \ln\left(1+4k^2\lambda_{\mathrm{D}}^2\right) - \frac{4k^2\lambda_{\mathrm{D}}^2}{\left(1+4k^2\lambda_{\mathrm{D}}^2\right)} \cong \ln\left(4k^2\lambda_{\mathrm{D}}^2\right), \quad \text{for } k\lambda_D \gg 1. \tag{8.34}$$

It is noted that $L(2k\lambda_{\mathrm{D}})$ is a slowly varying function of temperature and electron density, that $4k^2\lambda_{\mathrm{D}}^2 = 8m^* E\varepsilon_0\varepsilon_{\mathrm{s}}k_{\mathrm{B}}T/\hbar^2q^2n'$, and that $n' = n + (N_{\mathrm{D}} - N_{\mathrm{A}}^- - n)(N_{\mathrm{A}}^- + n)/N_{\mathrm{D}}$ is the density of screening electrons surrounding the ionized donor impurity. The integration of (8.33) can be carried out by letting $\sin(\theta'/2) = x$, $(1-\cos\theta') = 2x^2$, $\sin\theta'\mathrm{d}\theta' = 4x\,\mathrm{d}x$, and using a table of integrals.

Equation (8.33) is derived using the Brooks–Herring (B-H) model, and hence is known as the Brooks–Herring formula for ionized impurity scattering. By using the relation $E = \hbar^2k^2/2m^*$, and substituting (8.26) and (8.32) into (8.33), one obtains the inverse relaxation time as

$$\frac{1}{\tau_{\mathrm{I}}} = \frac{q^4 N_{\mathrm{I}} L(2k\lambda_{\mathrm{D}})}{16\pi(2m^*)^{1/2}\varepsilon_0^2\varepsilon_{\mathrm{s}}^2 E^{3/2}}. \tag{8.35}$$

Equation (8.35) shows that for ionized impurity scattering, the relaxation time τ_{I} is directly proportional to the energy to the 3/2 power (i.e., $\tau_{\mathrm{I}} \propto E^{3/2}$). The temperature dependence of τ_{I} comes only from the variation of $L(2k\lambda_{\mathrm{D}})$ with T, which is usually very small.

By substituting τ_{I}, given by (8.35), into (8.2) and averaging τ_{I} over the energy with the aid of (7.54), one obtains the ionized impurity scattering mobility μ_{I}, which reads

$$\mu_{\mathrm{I}} = \frac{q\left\langle\tau_I\right\rangle}{m^*} = \frac{64\sqrt{\pi}\varepsilon_0^2\varepsilon_{\mathrm{s}}^2(2k_{\mathrm{B}}T)^{3/2}}{N_I q^3\sqrt{m^*}\ln\left(\dfrac{12m^*k_{\mathrm{B}}^2T^2\varepsilon_0\varepsilon_{\mathrm{s}}}{q^2\hbar^2n'}\right)}, \tag{8.36}$$

which shows that the ionized impurity scattering mobility μ_I is directly proportional to the temperature to the 3/2 power (i.e., $\mu_I \propto T^{3/2}$). Good agreement has been found between the theoretical prediction given by (8.36) and mobility data for different semiconductors in which ionized impurity scattering is the dominant scattering mechanism.

Conwell and Weisskopf used the bare Coulomb potential as the perturbing Hamiltonian and derived a relaxation-time formula for the ionized impurity scattering given by

$$\frac{1}{\tau_I'} = \frac{q^4 N_I}{16\pi(2m^*)^{1/2}\varepsilon_0^2\varepsilon_s^2 E^{3/2}} \ln[1 + (2E/E_m)^2], \tag{8.37}$$

where $E_m = q^2/4\pi\varepsilon_0\varepsilon_s r_m$ and $N_I = (2r_m)^{-3}$. The ionized impurity scattering mobility derived from (8.37) is given by

$$\mu_I' = \frac{64\sqrt{\pi}\varepsilon_0^2\varepsilon_s^2(2k_B T)^{3/2}}{N_I q^3 m^{*1/2} \ln\left[1 + \left(12\pi\varepsilon_0\varepsilon_s k_B T/q^2 N_I^{1/3}\right)^2\right]}. \tag{8.38}$$

Equation (8.38) is known as the Conwell–Weisskopf formula for ionized impurity scattering. Comparing (8.36) and (8.38) reveals that both formulas are very similar except that the coefficient inside the logarithmic term is slightly different. It is of interest to note that both formulas predict the same temperature dependence for the ionized impurity scattering mobility and the same energy dependence for the relaxation time.

8.4. Neutral Impurity Scattering

Neutral impurity scattering is an important source of resistance in a semiconductor at very low temperatures. As the temperature decreases, carrier freeze-out occurs at the shallow-level impurity centers in an extrinsic semiconductor, and these shallow-level impurities become neutral at very low temperatures. The scattering potential due to a neutral shallow-level impurity center may be described by a square-well potential, which becomes the dominant scattering source for electrons or holes at very low temperatures.

In general, the scattering of charge carriers by neutral shallow-donor or shallow-acceptor impurities can be treated in a similar way to that of scattering of electrons by a hydrogen atom. The neutral shallow-donor atom in a semiconductor can be treated as a hydrogenic neutral atom immersed in the dielectric medium of the semiconductor whose dielectric constant is equal to that of the host semiconductor.

Erginsoy[3] derived the neutral impurity scattering mobility for a semiconductor using the partial wave technique to obtain the differential scattering cross-section. In the derivation, Erginsoy assumed that the electron velocity is low, and elastic scattering prevails in the semiconductor. Based on his derivation, the total differential scattering cross-section for neutral impurity scattering can be written

as

$$\sigma_{\mathrm{N}} \approx \frac{20a_{\mathrm{B}}}{k}, \tag{8.39}$$

where a_{B} is the Bohr radius given by (8.32). Using (8.22) and (8.39), the relaxation time for the neutral impurity scattering can be expressed by

$$\tau_{\mathrm{N}} = (N_{\mathrm{N}} v_{\mathrm{th}} \sigma_{\mathrm{N}})^{-1} = \frac{k}{20a_{\mathrm{B}} N_{\mathrm{N}} v_{\mathrm{th}}}. \tag{8.40}$$

Substituting a_{B} given by (8.32) and $k = m^* v/\hbar$ into (8.40), one obtains

$$\frac{1}{\tau_{\mathrm{N}}} = \frac{10\varepsilon_0 \varepsilon_{\mathrm{s}} N_{\mathrm{N}} h^3}{\pi^2 m^{*2} q^2}, \tag{8.41}$$

where N_{N} is the density of neutral impurities. Since the relaxation time for neutral impurity scattering is independent of energy, the mobility due to neutral impurity scattering can be readily obtained from (8.41), which yields

$$\mu_{\mathrm{N}} = \frac{q\tau_{\mathrm{N}}}{m^*} = \frac{\pi^2 m^* q^3}{10\varepsilon_0 \varepsilon_{\mathrm{s}} N_{\mathrm{N}} h^3}, \tag{8.42}$$

which shows that the carrier mobility due to neutral impurity scattering is independent of temperature. However, experimental results show that the carrier mobility is generally a weak function of temperature for many semiconductors at low temperatures.

8.5. Acoustical Phonon Scattering

Scattering of electrons by longitudinal-mode acoustical phonons is described in this section. The scattering of electrons by longitudinal-mode acoustical phonons is the most important scattering source in intrinsic or lightly doped semiconductors at room temperature. The scattering of electrons by longitudinal-mode acoustical phonons can usually be treated as an elastic scattering because the electron energy is much larger than the phonon energy and the change in electron energy during such a scattering process is small compared to the average energy of electrons. It can be shown that the maximum change of electron energy due to acoustical phonon scattering is given by

$$\Delta E \approx 4\left(\frac{u_{\mathrm{s}}}{\mathrm{v}_{\mathrm{th}}}\right) E_{\mathrm{e}}, \tag{8.43}$$

where $u_{\mathrm{s}} = 3 \times 10^5$ cm/s is the velocity of sound in a solid, and v_{th} is the mean thermal velocity of electrons ($\approx 10^7$ cm/s). Thus, the ratio of phonon energy to mean electron energy as given by (8.43) is usually much smaller than unity for $T > 100$ K. At very low temperatures, mean electron energy may become comparable to the acoustical phonon energy, and the assumption of elastic scattering may no longer be valid. Fortunately, at very low temperatures other types of scattering such as ionized impurity and neutral impurity scattering may become

dominant. It is noted that acoustical phonons may cause scattering in two different ways, either through deformation potential scattering or piezoelectric scattering. An acoustical wave may induce a change in the spacing of neighboring atoms in a semiconductor. This change in atomic spacing could result in the fluctuation of energy band gap locally on an atomic scale and is known as the deformation potential. The deformation potential is measured as the change of energy band gap per unit strain due to the acoustical phonons. This type of scattering is usually the most important scattering source for intrinsic or lightly doped silicon and germanium at room temperatures.

Piezoelectric scattering is another type of acoustical phonon scattering. This type of scattering is observed in III-V and II-VI compound semiconductors with the zincblende and wurtzite crystal structures. The lack of inversion symmetry in these semiconductors creates a strain-induced microscopic electric field perturbation, which leads to piezoelectric scattering with emission or absorption of an acoustical phonon. This type of scattering is important for pure III-V and II-VI compound semiconductors at low temperatures. These two types of acoustical phonon scattering are discussed next.

8.5.1. Deformation Potential Scattering

To derive an expression for relaxation time for nonpolar acoustical phonon scattering, the deformation potential technique developed originally by Bardeen and Shockley[4] for calculating the matrix element of longitudinal-mode acoustical phonon scattering will be discussed first. The perturbing Hamiltonian can be obtained from the deformation potential shown in Figure 8.5. Figure 8.5a shows the change of lattice spacing with respect to its equilibrium position due to lattice vibration. It is seen that thermal expansion and contraction of the lattice with temperature can lead to a change in the conduction and valence band edges or the energy band gap of the semiconductor, as shown in Figure 8.5b. Based on the deformation potential model proposed by Shockley, the fluctuation of the conduction

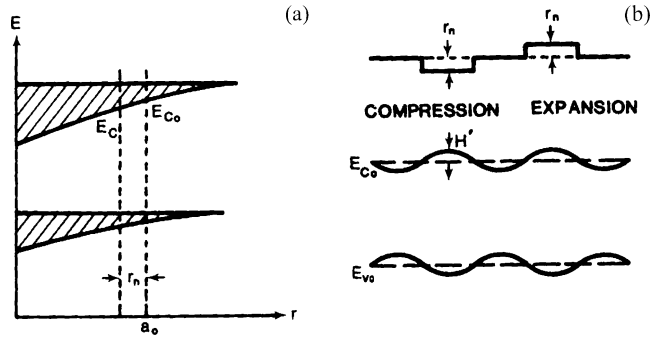


FIGURE 8.5. The change of conduction band edge and the deformation potential due to thermal expansion or contraction of the lattice spacing.

band edge due to lattice vibrations may be represented by a deformation potential. Therefore, the perturbing Hamiltonian can be related to the change of crystal volume caused by the lattice phonons and the deformation potential by the expression

$$H' = \Delta E_{\mathrm{c}} = E_{\mathrm{c}} - E_{\mathrm{co}} = \left(\frac{\Delta E_{\mathrm{c}}}{\Delta V}\right)\Delta V = E_{\mathrm{c1}}\left(\frac{\Delta V}{V}\right), \tag{8.44}$$

where E_{co} is the conduction band edge in thermal equilibrium, and

$$E_{\mathrm{c1}} = \frac{\Delta E_{\mathrm{c}}/\Delta T}{\Delta V/V\,\Delta T} \tag{8.45}$$

is the deformation potential constant. For silicon, $E_{\mathrm{c1}} = -16$ eV, and for germanium, $E_{\mathrm{c1}} = -9.5$ eV. The ratio $\Delta V/V$ represents the change of crystal volume to the total crystal volume due to temperature change in a semiconductor. Since $\Delta V/V$ can be expanded in terms of a Fourier series in the atomic displacement r_n, one can write

$$\frac{\Delta V}{V} = \nabla_{\mathrm{r}} \cdot r_{\mathrm{n}}, \tag{8.46}$$

where

$$r_{\mathrm{n}} = \sum_{j=1}^{3} (1/N)^{1/2} \zeta_j b_j(q)\, \mathrm{e}^{i(q\cdot R_{n0} - \omega t)}. \tag{8.47}$$

The lattice displacement r_{n} given by (8.47) can be expressed in terms of the normal coordinates and normal frequencies in three-dimensional form (i.e., two transverse branches and one longitudinal branch). It is also assumed that only the longitudinal-mode acoustical phonon scattering is important in the present case. Therefore, under this condition $\nabla_{\mathrm{r}} \cdot r_{\mathrm{n}}$ can be expressed by

$$\nabla_r \cdot r_{\mathrm{n}} = \sum_q q_{\mathrm{l}} r_{\mathrm{l}}, \tag{8.48}$$

where q_{l} is the wave vector of the longitudinal-mode acoustical phonon, and r_{l} represents the displacement due to longitudinal-mode acoustical phonons. Substituting (8.46) and (8.48) for $(\Delta V/V)$ into (8.44) yields the perturbing Hamiltonian H', which is given by

$$H' = E_{\mathrm{c1}}\left(\frac{\Delta V}{V}\right) = E_{\mathrm{c1}} \sum_q q_{\mathrm{l}} r_{\mathrm{l}}. \tag{8.49}$$

The matrix element $H_{kk'qq'}$ due to this perturbing Hamiltonian can be expressed as

$$H_{kk'qq'} = \left\langle k' n_{q'} \left| H' \right| k n_q \right\rangle = \int \phi^*_{k'} \varphi^*_{nq} \left(\sum_q E_{\mathrm{c1}} q_{\mathrm{l}} r_{\mathrm{l}}\right) \phi_k \varphi_{nq}\, \mathrm{d}^3 r\, \mathrm{d}^3 r_{\mathrm{l}}, \tag{8.50}$$

where φ_{nq} represents the phonon wave functions and ϕ_k is the electron wave functions. For phonon emission, the solution of (8.50) for the matrix element is given by

$$H_{kk'qq'} = \left(\frac{E_{c1}q_l}{N\Omega}\right)\left(\frac{\hbar}{M\omega}\right)^{1/2}\left(\frac{\langle n_q\rangle}{2}\right)^{1/2}, \tag{8.51}$$

and for phonon absorption it is given by

$$H_{kk'qq'} = \left(\frac{E_{c1}q_{\mathrm{l}}}{N\Omega}\right)\left(\frac{\hbar}{M\omega}\right)^{1/2}\left[\frac{\left(\langle n_q\rangle + 1\right)}{2}\right]^{1/2}. \tag{8.52}$$

In (8.51) and (8.52), $N\Omega$ is the volume of the crystal, M is the mass of the atom, and $\langle n_q\rangle$ is the average phonon population density given by

$$\langle n_q\rangle = \frac{1}{(\mathrm{e}^{\hbar\omega/k_{\mathrm{B}}T} - 1)} \approx \frac{k_{\mathrm{B}}T}{\hbar\omega}. \tag{8.53}$$

Equation (8.53) is valid for long-wavelength acoustical phonons (i.e., $k_{\mathrm{B}}T \gg \hbar\omega$ and $\mathrm{e}^{\hbar\omega/k_{\mathrm{B}}\mathrm{T}} \approx 1 + \hbar\omega/k_{\mathrm{B}}T$). The square of the matrix element due to deformation potential scattering can be obtained from the summation of the square of (8.51) and (8.52) and using the dispersionless relation $\omega = u_{\mathrm{s}}q_{\mathrm{l}}$, which yields

$$\left|H_{kk'qq'}\right|^2 = \frac{E_{\mathrm{cl}}^2 k_{\mathrm{B}}T}{M(u_{\mathrm{s}}N\Omega)^2}. \tag{8.54}$$

Substituting $\left|H_{kk'qq'}\right|^2$ given by (8.54) into (8.21) yields the differential scattering cross-section, which reads

$$\sigma_{\mathrm{a}} = \frac{m^{*2}E_{\mathrm{c1}}^2 k_{\mathrm{B}}T}{4\pi^2\hbar^4\rho u_{\mathrm{s}}^2} = \frac{m^{*2}E_{\mathrm{c1}}^2 k_{\mathrm{B}}T}{4\pi^2\hbar^4 c_{\mathrm{l}}}, \tag{8.55}$$

where $\rho = M/\Omega$ is the mass density of the atom, and Ω is the volume of the unit cell; $c_{\mathrm{l}} = \rho u_{\mathrm{s}}^2$ is the longitudinal elastic constant. For a cubic crystal, $c_{\mathrm{l}} = c_{11}$ for wave propagating along the (100) direction; for the (110) direction, $c_{\mathrm{l}} = (c_{11} + c_{12} + c_{44})/2$, and for the (111) propagation direction, $c_{\mathrm{l}} = (c_{11} + 2c_{12} + 4c_{44})/3$, where c_{11}, c_{12}, and c_{44} are components of the elasticity tensor.

The relaxation time due to longitudinal-mode acoustical phonon scattering can be obtained by substituting (8.55) into (8.23) and (8.22), yielding

$$\frac{1}{\tau_{\mathrm{a}}} = 2\pi v\int_0^{\pi}\sigma_{\mathrm{a}}\sin\theta'(1-\cos\theta')\,\mathrm{d}\theta' = \frac{m^{*2}vE_{\mathrm{c1}}^2 k_{\mathrm{B}}T}{\pi\hbar^4 c_{\mathrm{l}}} = \frac{v}{l_{\mathrm{a}}}, \tag{8.56}$$

where $l_a = \pi\hbar^4 c_{\mathrm{l}}/m^{*2}\mathrm{E}_{\mathrm{c1}}^2 k_{\mathrm{B}}T$ is the mean free path of electrons, which varies inversely with temperature. Substituting $v = (2E/m^*)^{1/2}$ into (8.56)

yields

$$\frac{1}{\tau_a} = \frac{m_n^{*3/2} k_B T E_{c1}^2 (2E)^{1/2}}{\pi \hbar^4 c_l}, \tag{8.57}$$

which shows that for acoustical phonon scattering τ_a varies with $E^{-1/2}$ and T^{-1}. The electron mobility due to acoustical phonon scattering can be obtained by substituting τ_a given by (8.57) into (7.64), with the result

$$\mu_a = \frac{q \langle \tau_a \rangle}{m_c^*} = \left(\frac{2\sqrt{2\pi} q \hbar^4 c_l}{3 m_n^{*3/2} m_c^* k_B^{3/2} E_{c1}^2} \right) T^{-3/2}, \tag{8.58}$$

where m_c^* is the conductivity effective mass of electrons. For cubic crystals with ellipsoidal constant-energy surfaces, the effective mass product $m_c m_n^{*3/2}$ is given by

$$\frac{1}{m_c m_n^{*3/2}} = \frac{1}{3 m_t m_l^{1/2}} \left(\frac{2}{m_t} + \frac{1}{m_l} \right), \tag{8.59}$$

where m_t and m_l are the transverse and longitudinal effective masses of electrons for the ellipsoidal conduction band valley, respectively. Equation (8.58) predicts that the electron mobility due to longitudinal-mode acoustical phonon scattering is directly proportional to $T^{-3/2}$. Figures 8.7 and 8.8 show the experimental results for electron mobilities in undoped germanium and silicon crystals, which were found to be in good agreement with theoretical predictions for $T < 200$ K. However, at high temperatures, intervalley optical phonon scattering contributes substantially to electron mobility, and hence μ_n varies as T^{-n}, where n lies between 1.5 and 2.7.

8.5.2. *Piezoelectric Scattering*

For polar semiconductors such as III-V and II-VI compound semiconductors, the bonds are partially ionic, and the unit cell does not possess inversion symmetry. As a result, charged carriers may be scattered by longitudinal-mode acoustical phonons due to piezoelectric scattering. In general, the strain-induced electric field due to the piezoelectric effect can be represented by

$$\mathcal{E}_{pz} = -\left(\frac{e_{pz}}{\varepsilon_0 \varepsilon_s} \right) (\nabla_r r_n) \ , \tag{8.60}$$

where e_{pz} is the piezoelectric constant. Thus, the perturbation potential due to piezoelectric scattering can be expressed by

$$H' = \frac{e\mathcal{E}_{pz}}{q} = \left(|e| \frac{e_{pz}}{\varepsilon_0 \varepsilon_s q} \right) (\nabla_r r_n), \tag{8.61}$$

where $q = |k' - k| = 2k \sin(\theta'/2) = (2m^* v/\hbar) \sin(\theta'/2)$ is the phonon wave vector, and $|e|$ is the electronic charge. A comparison of (8.61) with (8.49) for nonpolar

acoustical phonon scattering reveals that instead of the deformation potential constant E_{c1}, one has $|e|e_{pz}/\varepsilon_0\varepsilon_s q$, which is not a constant (since q depends on v and θ'). Thus, the matrix element $H_{kk'}$ due to piezoelectric scattering can be written as

$$H_{kk'} = \frac{|e|\, e_{pz}}{\varepsilon_0\varepsilon_s q}\left(\frac{k_B T}{2Vc_l}\right)^{1/2} = \left(\frac{e^2K^2k_BT}{2V\varepsilon_0\varepsilon_s q^2}\right)^{1/2}. \tag{8.62}$$

In (8.62) a dimensionless electromechanical coupling constant K^2 is introduced, which is defined by

$$\frac{K^2}{1-K^2} = \frac{e_{pz}^2}{\varepsilon_0\varepsilon_s c_l}. \tag{8.63}$$

The left-hand side of (8.63) reduces to K^2 if $K^2 \ll 1$. For most polar semiconductors, the value of K^2 is on the order of 10^{-3}.

The relaxation time due to piezoelectric scattering can be obtained by substituting (8.62) into (8.18) and (8.9), which yields

$$\frac{1}{\tau_{pz}} = \frac{V}{(2\pi)^2}\int 2\left(\frac{2\pi}{\hbar}\right)\left(\frac{e^2K^2k_BT}{2V\varepsilon_0\varepsilon_s q^2}\right)\delta(E_k - E_{k'})k'^2(1-\cos\theta')\sin\theta'\,d\theta'\,dk', \tag{8.64}$$

where $q^2 = 4k'^2\sin^2(\theta'/2)$ and $dk' = \hbar^{-1}(m^*/2E)^{1/2}\,dE$. Carrying out the integration in (8.64), one obtains

$$\tau_{pz} = \frac{2^{3/2}\pi\hbar^2\varepsilon_0\varepsilon_s}{m^{*1/2}e^2K^2k_BT}E^{1/2}, \tag{8.65}$$

which shows that the relaxation time for piezoelectric scattering is proportional to the square root of the energy. Thus, the carrier mobility due to piezoelectric scattering can be derived using the expression of τ_{pz} given by (8.65) and (8.2), and one has

$$\mu_{pz} = \frac{16\sqrt{2\pi}\hbar^2\varepsilon_0\varepsilon_s}{3m^{*3/2}|e|K^2(k_BT)^{1/2}}. \tag{8.66}$$

Equation (8.66) shows that the piezoelectric scattering mobility depends on $T^{-1/2}$. For a typical III-V compound semiconductor with $\varepsilon_s = 12$, $m^*/m_0 = 0.1$, and $K^2 = 10^{-3}$, a mobility value of 1.7×10^5 cm^2/(V·s) was obtained for piezoelectric scattering at $T = 300$ K. This value is significantly higher than the deformation potential scattering mobility for most polar semiconductors. Therefore, piezoelectric scattering is usually not as important as acoustical phonon scattering due to deformation potential or ionized impurity scattering. Thus, piezoelectric scattering has little influence on electron mobilities for most III-V compound semiconductors. However, piezoelectric scattering can become important for many II-VI compound semiconductors such as CdS and ZnSe, which have the wurtzite crystal structure. For example, ionic and polar crystals, including most of the II-VI compound semiconductors, show a strong piezoelectric effect because the wurtzite

crystal structure lacks inversion symmetry, and hence the piezoelectric stress tensor is nonvanishing. The microscopic origins of piezoelectricity are due to ionic polarization, strain-dependent ionization, and electronic polarization. It has been suggested that the strain-induced flow of covalent charge between sublattices may be the dominant source of piezoelectricity in II-VI compound semiconductors, since electronic polarization is usually accompanied by acoustical mode phonons in such a crystal. This polarization can lead to a periodic electric perturbation potential, which will contribute to electron scattering. The electron mobility due to piezoelectric scattering varies as $T^{-1/2}$, and the effects of piezoelectric scattering may be sufficiently large to be important in determining mobility in a piezoelectric crystal. For example, the temperature dependence of electron mobility for CdS crystal shows that contributions from optical-mode phonon scattering and piezoelectric scattering become dominant at high temperatures. In contrast, for III-V compound semiconductors, piezoelectric scattering becomes important only at very low temperatures.

8.6. Optical Phonon Scattering

Optical phonon scattering becomes the predominant scattering source at high temperatures or at high electric fields. Both polar and nonpolar optical phonons are responsible for this type of scattering. The scattering of electrons by nonpolar optical phonons may be treated as one type of deformation potential scattering process. Nonpolar optical phonon scattering becomes important for silicon and germanium crystals above room temperatures when intervalley scattering becomes the dominant process. However, intervalley scattering is generally not important for electrons in the conduction band minima located at the Γ-valley or along the $\langle 100 \rangle$ axes, but is important for conduction band minima located along the $\langle 111 \rangle$ axis (e.g., the Γ-valley in germanium and the L-valley in GaAs). Polar optical phonon scattering is the predominant scattering mechanism for ionic or polar crystals such as II-VI and III-V compound semiconductors. For these crystals the motion of negatively and positively charged atoms in a unit cell will produce an oscillating dipole, and the vibration mode is called the polar optical-mode phonon. Polar optical phonon scattering is associated with the atomic polarization arising from displacement caused by optical phonons. This is often the most important scattering mechanism at room temperature for III-V compound semiconductors. Optical phonon scattering is usually an inelastic process that cannot be treated by the relaxation time approximation because the optical phonon energy is comparable to that of mean electron energy (i.e., $\hbar\omega \approx k_B T$) at room temperature.

For a multivalley semiconductor such as silicon or germanium, intravalley scattering (i.e., scattering within a single conduction band minimum) near room temperature is usually accompanied by absorption or emission of a longitudinal-mode acoustical phonon. In this case, (8.58) is used to calculate the mobilities in these materials. However, at higher temperatures, intervalley scattering (i.e., scattering from one conduction band minimum to another) may become the dominant

scattering process. Intervalley scattering is usually accompanied by absorption or emission of a longitudinal-mode optical phonon. Since the energy of an optical phonon is comparable to that of the average electron energy, scattering of electrons by intervalley optical phonons is generally regarded as inelastic. In this case, the change in electron energy during scattering is no longer small, and hence the relaxation time approximation can be used only if certain assumptions are made for this type of scattering. For silicon and germanium, it is found that over the temperature range in which intervalley optical phonon scattering is comparable to acoustical phonon scattering, the temperature dependence of electron mobility can be described by an empirical formula given by

$$\mu_{\mathrm{n}} \propto T^{-n} \quad \text{with} \quad 1.5 < n < 2.5. \tag{8.67}$$

Figure 8.8 shows the temperature dependence of electron mobility in silicon at high temperatures. Theoretical calculations of hole mobility for p-type silicon show that hole mobility varies as $T^{-2.3}$ when both optical and acoustical phonon scatterings become dominant. This result compares favorably with the measured data.

In multivalley semiconductors such as silicon and germanium, intervalley scattering becomes important at high temperatures. In this case the scattering of electrons is controlled by nonpolar optical phonons, and the relaxation time is given by[5]

$$\frac{1}{\tau_{\mathrm{oi}}} = \left(\frac{m_{\mathrm{dn}}^{*3/2}}{\tau_0}\right) W\theta_{\mathrm{D}} T^{1/2} \left[\langle n_0 + 1\rangle \left(\mathcal{E}_0 - \frac{\theta_{\mathrm{D}}}{T}\right)^{1/2} + \langle n_0\rangle \left(\mathcal{E}_0 + \frac{\theta_{\mathrm{D}}}{T}\right)^{1/2}\right], \tag{8.68}$$

where θ_{D} is the Debye temperature, $\langle n_0\rangle = [\exp(\theta_{\mathrm{D}}/T) - 1]^{-1}$ is the average phonon distribution function, and W is a constant that determines the relative coupling strength between the electrons and optical phonons; $W = (D_0 h u_{\mathrm{s}})^2/2(k_0 a \theta_{\mathrm{D}})^2$, where D_0^2 is the optical deformation potential constant, $\mathcal{E}_0 = \hbar\omega/k_{\mathrm{B}}T$ is the reduced optical phonon energy, and a is the optical coupling constant. Note that the first term in (8.68) corresponds to the emission of an optical phonon, and the second term corresponds to the absorption of an optical phonon. Emission of optical phonons is important only when it is energetically possible (i.e., $\mathcal{E}_0 > \theta_{\mathrm{D}}/T$). The mobility due to intervalley optical phonon scattering can be calculated using (8.68) to find the average relaxation time $\langle\tau_{0i}\rangle$ and then substituting the result in the mobility formula $\mu_{0i} = q\langle\tau_{0i}\rangle/m_{\mathrm{c}}^*$. Based on (8.68) and the mobility formula, one can expect that the electron mobility due to intervalley optical phonon scattering will decrease exponentially with temperature (i.e., $\mu_{0\mathrm{i}} \sim \mathrm{e}^{\theta_{\mathrm{D}}/T}$).

In II-VI and III-V compound semiconductors, polar optical phonon scattering becomes the dominant scattering mechanism at room temperature. Coupling between the conduction electrons and the optical-mode phonons in a polar crystal such as GaAs is a very effective scattering source. Both perturbation theory and polaron theory have been employed to derive the polar optical phonon scattering mobility. The theoretical expression of electron mobility derived by Petritz and

Scanlon for polar optical-mode phonon scattering is given by[6]

$$\mu_{\text{po}} = \frac{8qa_0}{3(2\pi m k_{\text{B}}\Theta)^{1/2}} \left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_{\text{s}}}\right)^{-1} \left(\frac{m_0}{m^*}\right)^{1/2} \frac{\chi(Z_0)[\exp(Z_0) - 1]}{Z_0^{1/2}}, \qquad (8.69)$$

where ε_∞ is the high-frequency dielectric constant, ε_{s} is the low-frequency dielectric constant [$\varepsilon_{\text{s}} = \varepsilon_\infty(\omega_{\text{l}}/\omega_{\text{s}})^2$], $\Theta = \hbar\omega_{\text{l}}/k_{\text{B}}$, $a_0 = \hbar^2/mq^2$, and $\text{Z}_0 = \Theta/T$; ω_{l} is the angular frequency of the longitudinal optical phonon (LO) modes and $\chi(Z_0)$ is a quantity defined by Howarth and Sondheimer.[5] For pure GaAs crystal, with a longitudinal optical phonon temperature $\Theta = 416$ K (i.e., LO phonon energy $\hbar\omega_{\text{l}} \sim 36$ meV), the mobility μ_{po} is roughly equal to 10,000 $\text{cm}^2/\text{V}\cdot\text{s}$ at 300 K.

Due to the exponential dependence of μ_{po} on temperature, the scattering of electrons by polar optical phonons becomes very unlikely at low temperatures. For example, at room temperature, the electron mobility in a lightly to moderately doped GaAs is contributed to both the longitudinal acoustical phonon and polar optical phonon scatterings, while ionized impurity scattering becomes dominant at low temperatures.

8.7. Scattering by Dislocations

Dislocations in a semiconductor can act as scattering centers for both electrons and holes. The scattering of electrons by a dislocation may be attributed to two effects. First, a dislocation may be viewed as a line charge, and hence has an effect similar to that of a charged impurity center. Second, the strain field created by the dislocations in a crystal can produce a scattering potential similar to that of a deformation potential. However, it is generally known that scattering by dislocations can become important only if the density of dislocations is greater than 10^8 cm^{-2}.

To deal with scattering of electrons by dislocations, one may consider the dislocation line as a space charge cylinder of radius R and length L, as shown in Figure 8.6. The probability that an electron is scattered into an angle $\text{d}\theta'$ by a dislocation

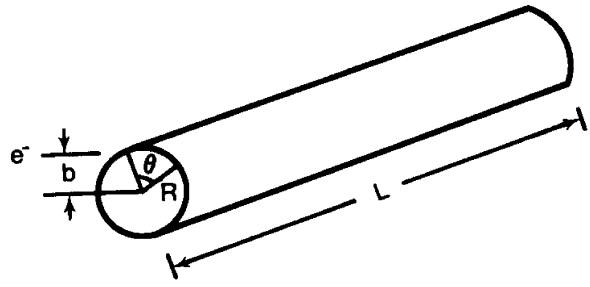


FIGURE 8.6. Scattering of electrons by a dislocation line.

line can be expressed by

$$P_{\mathrm{d}} = \frac{\mathrm{d}\,(b/R)}{\mathrm{d}\theta'} = \frac{1}{2}\sin\left(\frac{\theta'}{2}\right), \tag{8.70}$$

where b is the scattering impact parameter. The differential scattering cross-section per unit length of dislocation line charge is thus given by

$$\sigma_{\mathrm{d}}(\theta') = R\sin\left(\frac{\theta'}{2}\right). \tag{8.71}$$

The total scattering cross-section can be obtained by substituting (8.71) into (8.23) and integrating over θ' from 0 to π, which yields

$$\sigma_{\mathrm{T}} = \frac{8R}{3}. \tag{8.72}$$

Therefore, the relaxation time due to scattering of electrons by dislocations is given by

$$\tau_{\mathrm{d}} = \frac{1}{N_{\mathrm{d}}\sigma_{\mathrm{T}}\upsilon} = \frac{3}{(8N_{\mathrm{d}}R\upsilon)}. \tag{8.73}$$

The electron mobility due to scattering by dislocations can be obtained directly from (8.73), yielding

$$\mu_{\mathrm{d}} = \frac{q\tau_{\mathrm{d}}}{m^*} = \left(\frac{3q}{8N_{\mathrm{d}}R}\right)\frac{1}{(3m^*k_{\mathrm{B}}T)^{1/2}}, \tag{8.74}$$

where N_{d} is the density of dislocation lines. Equation (8.74) shows that the electron mobility due to scattering by dislocations is directly proportional to $T^{-1/2}$. For single-crystal silicon and germanium the dislocation density is usually very low, and hence scattering of electrons by dislocations is negligible. It should be pointed out that scattering of carriers by dislocations could also take place by virtue of their surrounding strain fields. The effect of strain fields can be calculated by finding a deformation potential from the known strain field. The scattering due to these strain fields is usually not important for n-type semiconductors, but could become important for p-type semiconductors.

8.8. Electron and Hole Mobilities in Semiconductors

Using the relaxation time approximation and the mobility formulas derived in this chapter for different scattering mechanisms, the electron and hole mobilities in a semiconductor could in principle be calculated over a wide range of temperatures and doping concentrations. However, one must realize that these mobility formulas are derived for the isotropic elastic scattering case. Some modifications may be needed so that these mobility formulas can be applied to practical semiconductors. In general, it is not a simple task to fit theoretical calculations with experimental data for electron and hole mobilities in a semiconductor over a wide range of doping concentrations and temperatures because in most semiconductors the total carrier

mobility is usually controlled by several scattering mechanisms, such as acoustical phonons, optical phonons, and ionized impurities. An exception may exist for ultrapure semiconductors in which longitudinal acoustical phonon scattering may prevail over a wide range of temperatures, and hence allows for direct comparison between the theoretical calculations and measured values. In general, electron mobility in a scmiconductor due to mixed scattering processes can be calculated using the expression

$$\mu_{\mathrm{n}} = \frac{q\langle\tau\rangle}{m_{\mathrm{c}}^*}, \tag{8.75}$$

where

$$\frac{1}{\tau} = \sum_i \frac{1}{\tau_i} \tag{8.76}$$

and τ_i denotes the relaxation time due to a particular scattering process. For example, if the scattering mechanisms are due to acoustical phonons, ionized impurities, and neutral impurities, then the total scattering time constant can be obtained by employing the reciprocal sum of the relaxation times due to these scattering processes, namely,

$$\tau^{-1} = \tau_{\mathrm{a}}^{-1} + \tau_{\mathrm{I}}^{-1} + \tau_{\mathrm{N}}^{-1}. \tag{8.77}$$

The electron mobility for the mixed scattering case can be calculated as follows: (1) find the total relaxation time τ due to different scattering mechanisms using (8.76); (2) calculate the average relaxation time $\langle\tau\rangle$ from (7.54); and (3) calculate the total electron mobility using (8.75). It should be pointed out here that computing the carrier mobility using the above procedure can be quite tedious if the relaxation time due to different scattering mchanisms is energy-dependent. In this case it may not be possible to obtain an analytical expression for the average relaxation time, and instead a numerical solution may be needed for finding the mean relaxation time and the total carrier mobility. On the other hand, if the relaxation time due to different scattering mechanisms is independent of energy, then one could use the simplified reciprocal sum formula to obtain the total electron mobility, which is given by

$$\mu_{\mathrm{n}}^{-1} = \sum_i \mu_i^{-1}. \tag{8.78}$$

Figures 8.7 through 8.15 show the calculated and measured values of electron and hole mobilities versus temperature for pure Ge, Si, GaAs, GaP, InSb, InP, InAs, CdS, and CdTe crystals, respectively. The solid lines are theoretical calculations, while the solid dots are the measured values.[7] Figure 8.16 shows (a) the carrier concentration versus reciprocal temperature and (b) Hall mobility versus temperature for the undoped (sample one) and silicon-doped (samples two to five) n-type GaN films grown by the MOCVD technique. The symbols refer to experimental data. Figure 8.17 shows (a) the hole concentration versus reciprocal temperature and (b) Hall mobility versus temperature for the Mg-doped, p-type GaN films. The symbols refer to the experimental data.

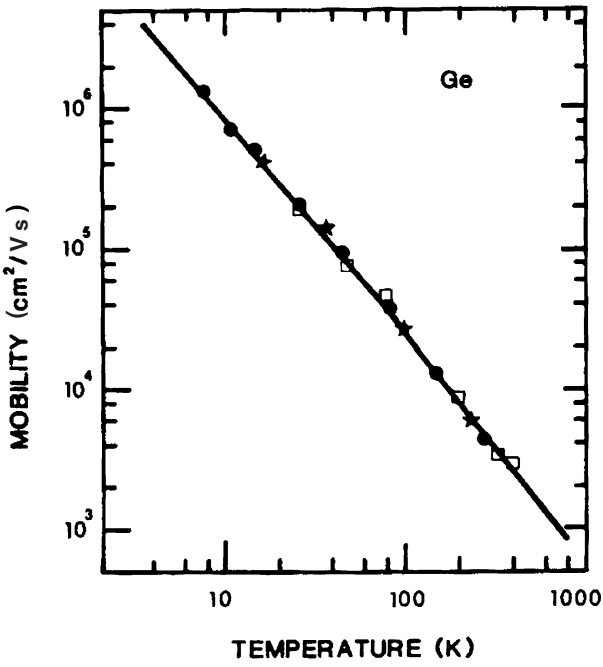


FIGURE 8.7. A comparison of the calculated drift mobility of electrons (solid curve) and the measured Hall mobility in a pure germanium specimen. The results show that acoustical phonon scattering is the dominant scattering mechanism in this sample. After Rode,[7] p. 83, by permission.

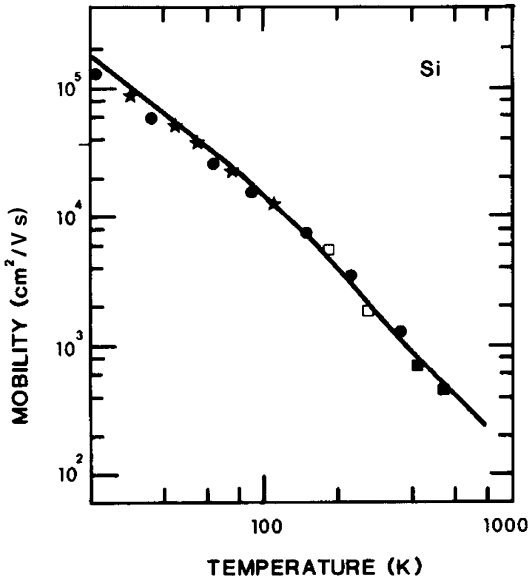


FIGURE 8.8. A comparison of the calculated drift mobility of electrons (solid curve) and the measured Hall mobility in a pure silicon specimen. The results show that acoustical phonon scattering is dominant for $T < 80$ K, and intervalley scattering becomes comparable to acoustical phonon scattering for $T \geq 300$ K. After Rode,[7] p. 81, by permission.

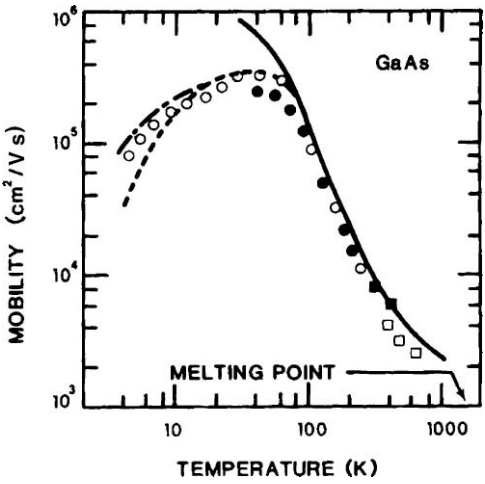


FIGURE 8.9. Comparison of calculated drift mobility of electrons (solid curve) and measured Hall mobility for a pure GaAs crystal. After Rode,[7] by permission.

The solid lines in Figure 8.17a result from a least-squares fit to the experimental data, which yields parameters for the shallow acceptors.

Table 8.1 lists electron drift mobilities for Ge, Si, GaP, and GaAs measured at 77 and 300 K. A comparison of the mobility data of these semiconductors shows that

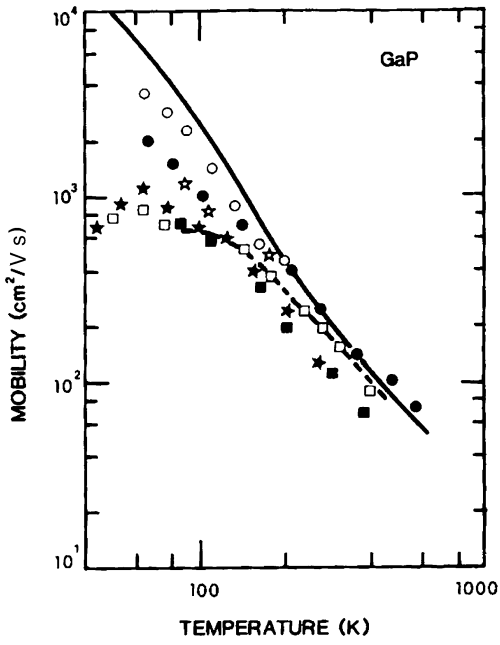


FIGURE 8.10. Comparison of calculated drift mobility of electrons (solid curve) and measured Hall mobility for a pure GaP crystal. After Rode,[7] by permission.

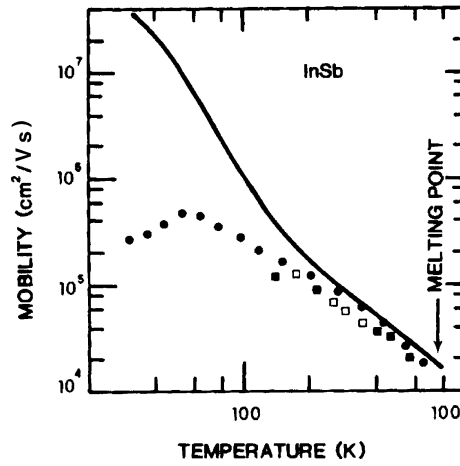


FIGURE 8.11. Comparison of calculated drift mobility of electrons (solid curve) and measured Hall mobility for a pure InSb crystal. After Rode,[7] by permission.

InSb has the highest electron mobility, while CdS has the lowest electron mobility. In general, the electron mobilities for III-V compound semiconductors such as GaAs, InP, and InAs are higher than for Si and Ge. Therefore, various electronic and photonic devices fabricated from III-V compound semiconductors are expected to operate at much higher frequencies and speeds than those of silicon devices. To facilitate mobility calculations in GaAs due to various scattering mechanisms,

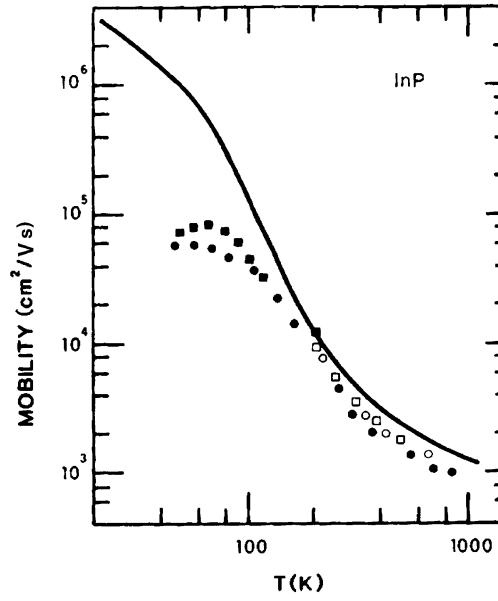


FIGURE 8.12. Comparison of calculated drift mobility of electrons (solid curve) and measured Hall mobility for a pure InP crystal. After Rode,[7] by permission.

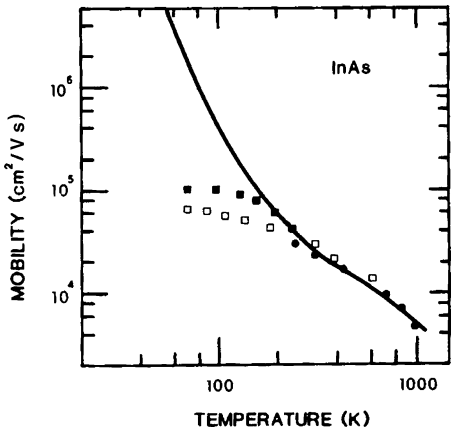


FIGURE 8.13. Comparison of calculated drift mobility of electrons (solid curve) and measured Hall mobility for a pure InAs crystal. After Rode,[7] by permission.

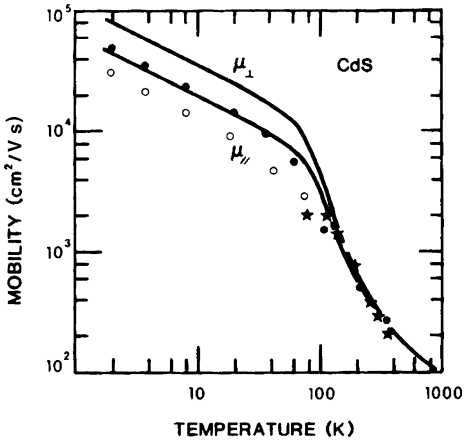


FIGURE 8.14. Comparison of calculated drift mobility of electrons (solid curve) and measured Hall mobility for a pure CdS specimen. After Rode,[7] by permission.

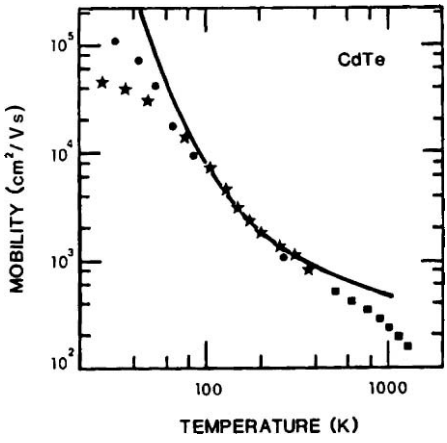


FIGURE 8.15. Comparsion of calculated drift mobility of electrons (solid curve) and measured Hall mobility for a pure CdTe specimen. After Rode,[7] by permission.

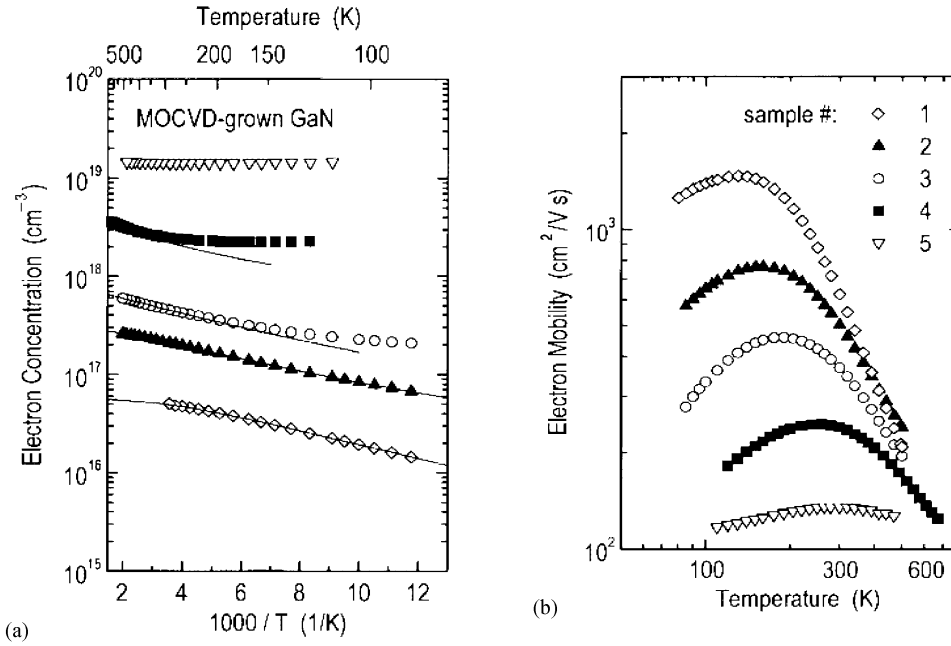


FIGURE 8.16. (a) Electron concentration versus reciprocal temperature and (b) Hall mobility versus temperature for the undoped (sample one) and silicon-doped (samples two through five) n-type GaN films grown by MOCVD technique. The symbols refer to the experimental data.

Table 8.2 lists some bulk- and valley-dependent material parameters for the GaAs crystal.

In n-type silicon, the important scattering mechanisms for electrons are mainly due to acoustical phonon and ionized impurity scatterings. At room temperature, the longitudinal-mode acoustic phonons are the dominant scattering source for undoped silicon, while ionized impurity scattering becomes important for $N_D \geq 10^{17}$ cm^{-3}. Optical deformation potential scattering is negligible for electron scattering within a particular conduction band minimum, since the matrix element vanishes due to symmetry. However, scattering between different conduction band minima (i.e., intervalley optical phonon scatterings) may become important at higher temperatures. The scattering mechanisms in the conduction band of GaAs crystal are different from that of silicon. Due to the spherical symmetry of the electron wave functions at the Γ-band, optical deformation potential scattering is zero in the conduction band minimum. Furthermore, due to the small electron effective mass (i.e., $m^* = 0.067m_0$) at the Γ-band minimum, the contribution of acoustical phonon scattering to electron mobility is also negligible in GaAs. As a result, electron mobility in GaAs is much higher than in silicon. Important scattering mechanisms for GaAs are polar optical phonon scattering (for

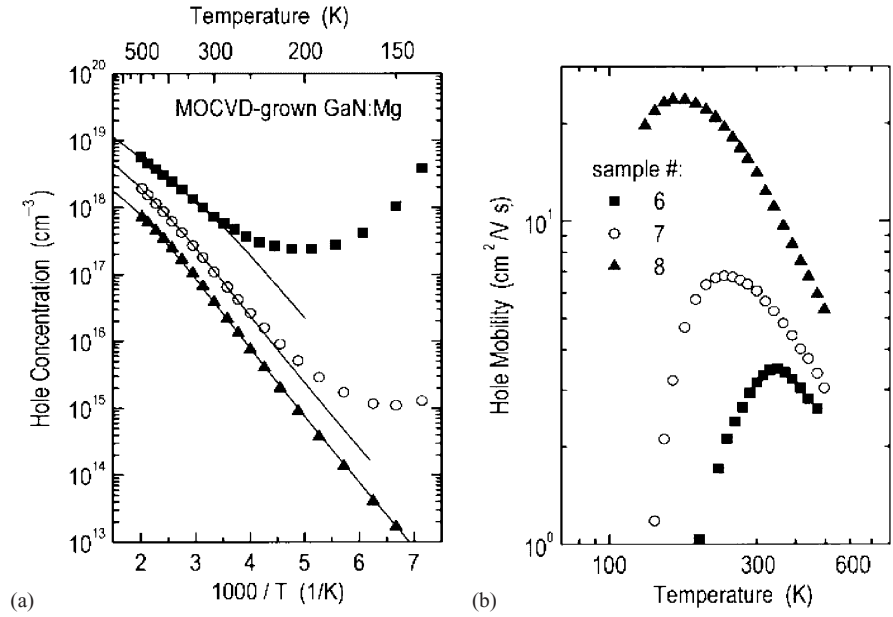


FIGURE 8.17. (a) Hole concentration versus reciprocal temperature and (b) Hall mobility versus temperature for Mg-doped, p-type GaN films. The symbols refer to the experimental data. The solid lines in (a) are a least-squares fit to the experimental data, which yields parameters for shallow acceptors.

pure and lightly doped GaAs), ionized impurity scattering (for $N_D \geq 10^{17}$ cm^{-3}), and intervalley optical phonon scattering (at high fields).

For p-type silicon and GaAs, the valence band maxima for both silicon and GaAs are located at the Γ-point (i.e., the zone center), and wave functions of holes do not possess spherical symmetry. Thus, optical deformation potential scattering is important for holes in p-type GaAs. In addition, both acoustical phonon scattering and ionized impurity scattering may also play an important role in the valence bands for both materials.

TABLE 8.1. Electron Drift Mobilities μ_n(cm^2/V·s) for Ge, Si, GaP, and GaAs.

	Ge	Si	GaP	GaAs
		$T = 300$ K		
Calculated	4080	1580	183	8920
Measured	3800–4200	1350–1450	120–200	3500–9000
		$T = 77$ K		
Calculated	37,400	22,800	4370	2.9×10^5
Measured	35,000–47,000	18,000–24,000		2.2×10^5

TABLE 8.2. Bulk- and valley-dependent material parameters for mobility calculations in GaAs.

(a) Bulk parameters:			
Density (g/cm)	5.36		
Piezoelectric constant (C/m^2)	0.16		
LO phonon energy (eV)	0.36		
Longitudinal sound velocity (cm/s)			
Optical dielectric constant	10.92		
Static dielectric constant	12.90		
(b) Valley material parameters:	Γ[100]	L[111]	X[100]
Electron effective mass (m^*/m_0)	0.067	0.222	0.58
Energy band gap E_g(eV)	1.43	1.77	1.96
Acoustic deformation potential (eV)	7.0	9.2	9.7
Optical deformation potential (eV/cm)	0	3×10^8	0
Number of equivalent valleys	1	4	3
Intervally deformation potential constant D (eV/cm)			
Γ	0	1×10^9	1×10^9
L	1×10^9	1×10^9	1×10^8
X	1×10^9	5×10^8	7×10^8

8.9. Hot Electron Effects in a Semiconductor

As discussed in Chapter 7, Ohm's law prevails in low-electric-field conditions, and the current density varies linearly with the applied electric field. This can be expressed as

$$J_n = \sigma_n \mathcal{E}_x = q n_0 \mu_n \mathcal{E}_x, \tag{8.79}$$

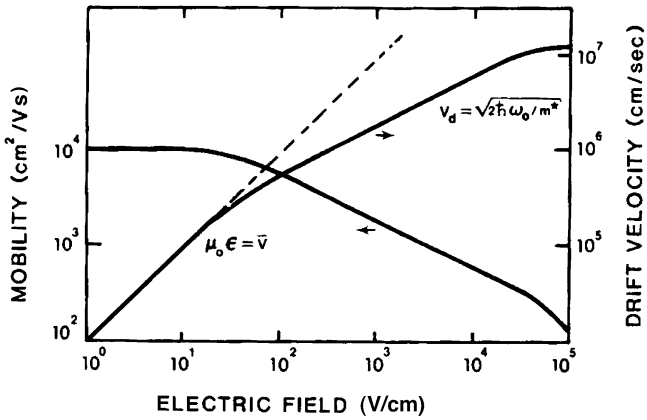


FIGURE 8.18. Electron mobility that drift velocity versus electric field calculated for a typical semiconductor, assuming that longitudinal acoustical phonon scattering dominates at low and intermediate fields, and optical phonon scattering dominates at high fields. Values of parameters used in the calculations are $\mu_0 = 104$ cm^2/V·s, $u_s = 2 \times 10^5$ cm/s, $\hbar\omega_0 = 0.04$ eV, and $m^* = m_0$. After Bube,[8] by permission.

where σ_n is the electrical conductivity, and $\mathcal{E}_x$ is the applied electric field. As the electric field continues to increase, a point is reached at which the electric current density will no longer vary linearly with the electric field. This means that either the electron density or the electron mobility becomes a function of the electric field. An increase in the electron density is possible if the electric field is high enough to cause (1) impact ionization (i.e., ionization of other imperfections or crystal atoms upon impact by hot electrons), (2) field ionization (i.e., ionization of imperfections by quantum-mechanical tunneling to the nearest band), or (3) electrical injection (i.e., injection of electrons from contacts into the semiconductor). These processes may lead to a change of electric current with applied electric field that is faster than that predicted by (8.79). It will be shown later that these effects are usually observed in a p-n junction diode operating under a large reverse bias condition. Another high-field effect, which has been found in many III-V compound semiconductor devices, is that the electric current density will increase with the electric field at a slower rate than that predicted by (8.79) under high-field conditions. This effect arises from the decrease in electron mobility with increasing electric field resulting from scattering of electrons by optical phonons under high-electric-field conditions.

In this section, only the effect of applied electric fields on electron mobility is considered. The mobility versus electric field relation can be derived by assuming that the scattering of electrons is dominated by the longitudinal-mode acoustical phonons.

It is generally known that as the electric field increases, the electrons will gain energy from the applied electric field. Furthermore, scattering of electrons is associated with the absorption or emission of phonons. Thus, in order to calculate energy loss by electrons due to phonon scattering, it is necessary to determine the average energy resulting from either absorption or emission of phonons under high electric-field conditions. The electron energy will increase if there is a net gain in energy due to phonon absorption.

Under high-electric-field conditions, electron energy can be described in terms of an effective electron temperature T_e. For a nondegenerate semiconductor, an increase in energy on the order of $k_B T$ represents a large change in the mean electron energy. An effective electron temperature T_e for such an energetic electron may be defined by the Maxwellian mean velocity, which is given by

$$\langle v \rangle = \left(\frac{k_B T_e}{8\pi m^*} \right)^{1/2} . \tag{8.80}$$

If the effective electron temperature defined by (8.80) is equal to the lattice temperature, then the electron mobility is independent of the electric field. On the other hand, if there is a net gain of energy due to the effects of applied electric field and acoustical phonon scattering, then the electrons will heat up. Under this condition T_e becomes larger than the lattice temperature of the crystal, and the electric current will no longer vary linearly with the electric field. Derivation of the current–electric field relation under high-field conditions is quite complicated, and only the relation between the electron mobility and the electric field is discussed in this section.

Under steady-state conditions, electron mobility as a function of electric field can be expressed in terms of the effective electron temperature T_e. The low-field electron mobility for longitudinal acoustical phonon scattering is given by

$$\mu_0 = \frac{4ql}{3\,(2\pi m^* k_B T)^{1/2}}, \tag{8.81}$$

where l is the mean free path of electrons, which is inversely proportional to the temperature. If the electron mobility under high-field conditions is expressed in terms of the low-field electron mobility μ_0 and the effective electron temperature T_e, then one can write the field-dependent electron mobility as

$$\mu = \mu_0 \left(\frac{T}{T_e}\right)^{1/2}. \tag{8.82}$$

The condition for T_e to exceed the lattice temperature T is that $\mu_0 \mathcal{E}_x > u_s$. This means that the effective electron temperature starts to rise when the drift velocity becomes comparable to the velocity of sound (u_s) in the semiconductor. Since μ_0 is proportional to $T^{-3/2}$ for acoustical phonon scattering, it follows that μ is proportional to $T^{-1}T_e^{-1/2}$. For scattering by acoustical mode phonons, the effective electron temperature versus electric field can be written as[8]

$$T_e = \left(\frac{T}{2}\right)\left\{1 + \left[1 + \left(\frac{3\pi}{8}\right)\left(\frac{\mu_0 \mathcal{E}_x}{u_s}\right)^2\right]^{1/2}\right\}. \tag{8.83}$$

In the relatively low field regime, where $\mu_0 \mathcal{E}_x \ll u_s$, T_e can be simplified to

$$T_e \approx T\left[1 + \left(\frac{3\pi}{32}\right)\left(\frac{\mu_0 \mathcal{E}_x}{u_s}\right)^2\right]. \tag{8.84}$$

Now by substituting (8.85) into (8.82) and carrying out binomial expansion, one obtains the corresponding field-dependent electron mobility, which reads

$$\mu = \mu_0\left[1 - \left(\frac{3\pi}{64}\right)\left(\frac{\mu_0 \mathcal{E}_x}{u_s}\right)^2\right]. \tag{8.85}$$

It is noted from (8.85) that in the intermediate-field regime, the differential mobility $(\mu_0 - \mu)$ varies as the square of the applied electric field. In the high-field regime, with $\mu_0 \mathcal{E}_x \gg u_s$, (8.83) becomes

$$T_e = T\left[\left(\frac{3\pi}{32}\right)^{1/2}\left(\frac{\mu_0 \mathcal{E}_x}{u_s}\right)\right], \tag{8.86}$$

and the corresponding electron mobility is given by

$$\mu = \left(\frac{32}{3\pi}\right)^{1/4}\left(\frac{\mu_0 u_s}{\mathcal{E}_x}\right)^{1/2}. \tag{8.87}$$

Equation (8.87) shows that electron mobility at high fields is inversely proportional to the square root of the electric field. Since the drift velocity v_d is equal to

the product of electron mobility and electric field, it will increase with the square root of the electric field at high fields.

It should be noted that the results obtained above are for the case that scattering of electrons is due to longitudinal acoustical phonons. For such scattering, increasing electron energy with the applied electric field will result in an increase of phonon scattering, which in turn will lead to the reduction of electron mobility with increasing electric field. On the other hand, if scattering is dominated by ionized impurity scattering, then an increase in electron energy with increasing electric field will result in an increase of electron mobility. This is due to the fact that for ionized impurity scattering, the probability of scattering decreases with increasing electron energy (i.e., $\tau_{\mathrm{I}}^{-1} \sim E^{-3/2}$).

Figure 8.16 shows a plot of electron mobility and drift velocity versus electric field calculated for a typical semiconductor at 300 K.[2] In this figure, it is assumed that scattering of electrons is dominated by the longitudinal-mode acoustical phonons at low and intermediate electric fields and by optical-mode phonons at high electric fields. The results clearly show that for scattering by acoustical phonons, the electron mobility will decrease with the square of the applied electric fields, and the high-field electron mobility will vary inversely with the square root of the electric fields. At very high fields, hot electrons will start interacting with optical phonons, which in turn will limit the drift velocity to a saturation value.

The most widely used method to study the hot electron effects in a semiconductor is the Monte Carlo approach. It consists in a simulation of the motion of one or more electrons inside a semiconductor subject to the action of an external applied electric field and given scattering mechanisms. The basic principle of the Monte Carlo method relies on the generation of a sequence of random numbers with given distribution probabilities. When charge transport is analyzed on submicrometer scales under very high electric field conditions, the conventional semiclassical approach of transport processes in terms of the Boltzmann equation can be substituted by a full quantum-mechanical description, namely, the Monte Carlo approach.

A brief description of the general procedure governing the Monte Carlo method is given as follows: Consider the case of a cubic semiconductor under a very high electric field $\mathcal{E}_x$. The simulation starts with a set of given initial conditions with initial wave vector k_0. The duration of the first free flight is determined stochastically from a probability distribution determined by the scattering probabilities. The simulation of all quantities of interest, such as velocity and energy, are recorded. A dominant scattering mechanism is then selected as being responsible for the end of the free flight according to the relative probabilities of all possible scattering mechanisms. From the transition rate of this scattering mechanism, The value of a new wave vector k after scattering is determined stochastically as the initial state of the new free flight, and the entire process is repeated iteratively. The results of the calculation become more and more accurate, and the simulation ends when the quantities of interest are known with the desired precision. A detailed description of this method can be found in a monograph edited by Reggiani.[9] The Monte

Carlo method allows one to extract derived phyical information from simulated experiments, and is a powerful tool for analyzing stationary and transient transport effects in semiconductors under high-field conditions. It is particularly useful for analyzing high-field transport properties in submicron devices.

Problems

8.1. Using (8.18), (8.19), and (8.20) derive (8.21), assuming that $v_k = v_{k'}$, $k = k'$, and $\mathrm{d}^3k' = k'^2 \sin\theta' \mathrm{d}\theta' \mathrm{d}\phi' \mathrm{d}k'$.

8.2. Using the Conwell–Weisskopf model, derive (8.37) and (8.38) (i.e., $V(r) = \mathrm{q}/4\pi\varepsilon_0\varepsilon_\mathrm{s} r$ for ionized impurity scattering).

8.3. Show that the maximum change of electron energy due to acoustical phonon scattering is given by (8.43). Does this satisfy the condition of elastic scattering?

8.4. Calculate the Debye screen lengths for Si, Ge, and GaAs for $N_\mathrm{D} = 10^{15}$, 10^{17}, and 10^{19} cm^{-3}, given $\varepsilon_\mathrm{s} = 11.7$ for Si, 12 for GaAs, and 16 for Ge.

8.5. If the electron mobility in silicon is due to scattering of acoustical phonons and ionized impurities, show that the mixed scattering mobility can be approximated by

$$\mu_{LI} = \mu_L \left\{1 + \chi^2 \left[\mathrm{Ci}(\chi)\cos\chi + \sin\chi \left(\mathrm{Si}(\chi) - \frac{\pi}{2}\right)\right]\right\},$$

where $\chi^2 = 6\mu_\mathrm{L}/\mu_\mathrm{L}$, and μ_L and μ_I are the acoustical phonon scattering and ionized impurity scattering mobilities; $\mathrm{Ci}(\chi)$ and $\mathrm{Si}(\chi)$ are the cosine and sine integrals of χ, respectively. (See the paper by P. P. Debye and E. M. Conwell, *Phys. Rev.* **93,** 693 (1954).)

8.6. Using (8.58), calculate the electron mobility due to acoustical phonon scattering for pure silicon for $100\ \mathrm{K} \leq T \leq 300\ \mathrm{K}$, given $(m_0/m^*)^{5/2} = 20.4$, $E_{\mathrm{c}1} = 12.8$ eV, and $lu_\mathrm{s}^2 = 1.97 \times 10^{12}$ dynes/cm^2.

8.7. Using the expression for τ_I given by (8.35) and $\mu_\mathrm{I} = q\langle\tau_\mathrm{I}\rangle/m^*$, show that the ionized impurity scattering mobility is given by (8.36).

8.8. The inverse scattering relaxation time for piezoelectric scattering in a nondegenerate semiconductor with a parabolic band is given by

$$\tau_\mathrm{pe}^{-l} = \frac{3q^2\kappa T P^2 m^*}{6\pi\hbar^3\varepsilon_0 k'},$$

where k' is the electron wave vector and P is the piezoelectric coefficient. Derive an expression for the piezoelectric scattering mobility, and show that the mobility is proportional to $T^{1/2}$.

8.9. For a GaAs crystal, the polar optical phonon scattering mobility μ_p0 given by (8.69) can be simplified to

$$\mu_\mathrm{po} = 5.3 \times 10^3 \left(\frac{\chi(Z_0)\left[\exp(Z_0) - 1\right]}{Z_0^{1/2}}\right),$$

where $\Theta = \hbar\omega_l / k_B$, $a_0 = \hbar^2/mq^2$, and $Z_0 = \Theta/T$; ω_l is the angular frequency of the longitudinal optical modes, and $\chi(Z_0)$ is a quantity defined by Howarth and Sondheimer.[5] For pure GaAs, the longitudinal optical phonon temperature Θ is equal to 416 K (i.e., the LO phonon energy $\hbar\omega_l \approx 36$ meV), and $\mu_{p0} \approx 10{,}000$ $cm^2/(V{\cdot}s)$ at 300 K.

The ionized impurity scattering mobility μ_i is given by

$$\mu_i = \frac{1.5 \times 10^{18}}{N_I[\ln(1+b) - b/(1+b)]} T^{3/2},$$

where

$$b = \frac{9.1 \times 10^{13}}{n_0} T^2.$$

The piezoelectric scattering mobility is given by

$$\mu_{pz} = 4.89 \times 10^5 \left(\frac{100}{T}\right)^{1/2}.$$

Assuming that Matthiessen's rule prevails, the total electron mobility for this GaAs crystal can be approximated by

$$\mu_n^{-1} = \mu_{po}^{-1} + \mu_i^{-1} + \mu_{pz}^{-1}.$$

Using the above expression, plot the electron mobility versus temperature for this GaAs crystal for 100 K $< T <$ 600 K for $N_I = 10^{16}$, 10^{17}, and 10^{18} cm^{-3}.

References

1. E. M. Conwell and V. F. Weisskopf, *Phys. Rev.* **77,** 388–390 (1950).
2. H. Brooks, in: *Advances in Electronics and Electron Physics* (L. Marton, ed.), Vol. 7, Academic Press, New York (1955), pp. 85–182.
3. C. Erginsoy, *Phys. Rev.* 1013–1017 (1956).
4. J. Bardeen and W. Shockley, *Phys. Rev.* **80,** 72–84 (1950).
5. D. Howarth and E. Sondheimer, *Proc. R. Soc. Lond. Ser. A* **219,** 53 (1953).
6. R. L. Petritz and W. W Scanlon, *Phys. Rev.* **97,** 1620 (1955).
7. D. L. Rode, in: *Semiconductors and Semimetals* (R. K. Willardson and A. C. Beer, eds.), Vol. 10, Academic Press, New York (1975).
8. R. H. Bube, *Electronic Properties of Crystalline Solids*, Chapter 8, Academic Press, New York (1974) p. 289.
9. R. Reggiani, *Hot Electron Transport in Semiconductors*, Springer-Verlag, New York (1985).

Bibliography

F. J. Blatt, *Physics of Electronic Conduction in Solids*, McGraw-Hill, New York (1968).
C. Herrings and E. Vogt, "Transport and Deformation Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering," *Phys. Rev.* **101,** 944–961 (1956).

S. S. Li, *The Dopant Density and Temperature Dependence of Electron Mobility and Resistivity in n-type Silicon*, NBS Special Publication, (1977), pp. 400–433.
S. S. Li, *The Dopant Density and Temperature Dependence of Hole Mobility and Resistivity in p-type Silicon*, NBS Special Publication (1979), pp. 400–447.
K. Seeger, *Semiconductor Physics*, Springer-Verlag, New York (1973).

9
Optical Properties and Photoelectric Effects

9.1. Introduction

This chapter presents the fundamental optical properties and bulk photoelectric effects in a semiconductor. The optical properties associated with the fundamental and free-carrier absorption processes and the internal photoelectric effects such as photoconductive (PC), photovoltaic (PV), and photomagnetoelectric (PME) effects in a semiconductor are described. Important fundamental physical and electronic properties such as energy band structures, excess carrier phenomena, and recombination mechanisms can be understood by studying the optical absorption processes and photoelectric effects in a semiconductor. Many practical applications have been developed using internal photoelectric effects such as PV and PC effects in semiconductors. Future trends are moving toward further development of various optoelectronic devices for a wide variety of applications in PV devices (solar cells), light-emitting diodes (LEDs) and laser diodes (LDs), and optoelectronic integrated circuits (OEICs) for use in optical computing, optical communications, signal processing, and data transmission.

Depending on the energy of incident photons, there are two types of optical absorption processes that may occur in a semiconductor. The first type involves the absorption of photons, which have energies equal to or greater than the band gap energy of a semiconductor. This type of optical absorption is called the fundamental or interband absorption process. The fundamental absorption process is usually accompanied by an electronic transition across the forbidden gap, and as a result, excess electron–hole pairs are generated in the semiconductor. The absorption coefficient due to the interband transition is usually very large. For example, in the ultraviolet (UV) to visible spectral range, typical values of the absorption coefficient for most semiconductors vary from 10^6 cm^{-1} near the UV wavelength to around 1 cm^{-1} near the cutoff wavelength of the semiconductor. However, the absorption coefficient becomes very small (e.g., less than 1 cm^{-1}) when the photon energies fall below the band gap energy of the semiconductor. In this case, another type of optical absorption process takes place in the

semiconductor. This type of optical absorption results in electronic transitions only within the allowed energy band, and is called the free-carrier absorption process. The fundamental absorption process, which leads to an interband transition, must be treated quantum mechanically, while the free-carrier absorption process can be described by classical electromagnetic (EM) wave theory. Finally, absorption of photons with energies below the band gap energy of the semiconductor may also lead to electronic transitions from localized impurity states to the conduction or valence band states. For example, the extrinsic photoconductivity observed at low temperatures is due to the photoexcitation of free carriers from shallow-impurity states to conduction or valence band states. Since the energy band gap varies between 0.1 and 6.2 eV for most semiconductors, the fundamental optical absorption may occur in the UV, visible, and infrared (IR) spectral regimes (i.e., 0.3 to 10 μm). Therefore, most semiconductors are opaque from the UV to the IR spectral range, and become transparent in the IR spectral regime for $\lambda > 10\,\mu m$.

In order to better understand the optical absorption processes in a semiconductor, it is necessary first to consider two optical constants, namely, the index of refraction and the extinction coefficient. These two optical constants may be derived by solving the Maxwell wave equations for the EM waves in a solid, as will be described in Section 9.2. The free-carrier absorption process is presented in Secton 9.3. Section 9.4 deals with the fundamental absorption process in a semiconductor. The internal photoelectric effects such as PC, PV, and PME effects in a semiconductor are described in Sections 9.5, 9.6, and 9.7, respectively. Section 9.5 presents both the intrinsic and extrinsic PC effects in a semiconductor. The internal PV effect also known as the Dember effect is discussed in Section 9.6. The PME effect in a semiconductor is presented in Section 9.7.

9.2. Optical Constants of a Solid

Optical constants such as the index of refraction and extinction coefficient can be derived by solving the Maxwell equations for EM waves propagating in a solid. It is well known that some solids are transparent while others are opaque, that some solid surfaces are strongly reflective while others tend to absorb optical radiation that falls on them. The degree of optical absorption depends on the wavelength of the incident optical radiation. For example, most semiconductors show strong absorption from UV ($\lambda < 0.4$ μm), visible to near-IR ($0.4 < \lambda < 2$ μm), mid-wavelength IR (3–5 μm), and long-wavelength IR (8–12 μm), and become transparent in the far-IR ($\lambda > 14$ μm) spectral regime. Therefore, in order to obtain a better understanding of the optical absorption process in a semiconductor, it is important to derive the expressions of the two basic optical constants (i.e., index of refraction and extinction coefficient) in the UV to IR spectral range.

The propagation of EM waves in a solid can be described by the Maxwell wave equations, which are given by

$$\nabla \cdot \mathcal{E} = 0, \tag{9.1}$$

$$\nabla \times \mathcal{E} = -\frac{\partial B}{\partial t}, \tag{9.2}$$

$$\nabla \cdot B = 0, \tag{9.3}$$

$$\nabla \times H = \sigma \mathcal{E} + \varepsilon_0 \varepsilon_s \frac{\partial \mathcal{E}}{\partial t}. \tag{9.4}$$

In free space, the EM wave equation can be obtained from (9.1) through (9.4) by setting $B = \mu_0 H$, $\sigma = 0$, and $\varepsilon_s = 1$, which yields

$$\nabla^2 \mathcal{E} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathcal{E}}{\partial t^2} = \left(\frac{1}{c^2}\right) \frac{\partial^2 \mathcal{E}}{\partial t^2}\ , \tag{9.5}$$

where $c = 1/\sqrt{\varepsilon_0 \mu_0}$ is the speed of light in free space. Inside the solid, the wave equation can also be obtained by solving (9.1) through (9.4), and the result yields

$$\nabla^2 \mathcal{E} = \mu_0 \varepsilon_0 \varepsilon_s \frac{\partial^2 \mathcal{E}}{\partial t^2} + \mu_0 \sigma \frac{\partial \mathcal{E}}{\partial t}. \tag{9.6}$$

A comparison of (9.5) and (9.6) shows that the difference between the waves propagating in free space and in a solid is due to the difference in the dielectric constant and electrical conductivity in both media. It is clear that (9.6) will reduce to (9.5) if the dielectric constant ε_s equals 1 and the electrical conductivity σ is zero. The first term on the right-hand side of (9.6) is the displacement current density, while the second term represents the conduction current density. An EM wave with frequency ω propagating in the z-direction and polarizing in the x-direction can be expressed by

$$\mathcal{E}_x = \mathcal{E}_0 \exp\left[i\omega\left(\frac{z}{v} - t\right)\right] = \mathcal{E}_0 \exp\left[i\left(k^* \cdot z - \omega t\right)\right], \tag{9.7}$$

where k^* is the complex wave vector and v is the velocity of the EM waves inside the solid, which could be a complex number. It is noted that k^* and v are related by

$$k^* = \frac{\omega}{v}. \tag{9.8}$$

Now, by substituting (9.7) into (9.6), one obtains

$$k^{*2} = \frac{\omega^2}{v^2} = \mu_0 \varepsilon_0 \varepsilon_s \omega^2 + i\mu_0 \sigma \omega, \tag{9.9}$$

or

$$k^* = \frac{\omega}{v} = \left(\frac{\omega}{c}\right)\left(\varepsilon_s + \frac{i\sigma}{\omega \varepsilon_0}\right)^{1/2} = \left(\frac{\omega}{c}\right) n^*, \tag{9.10}$$

where

$$n^* = \left(\varepsilon_s + \frac{i\sigma}{\omega \varepsilon_0}\right)^{1/2} = \varepsilon_s^{*1/2} \tag{9.11}$$

is the complex refractive index of the solid and ε_s^* is the complex dielectric constant; the complex refractive index n^* can be expressed by

$$n^* = n + ik_e, \tag{9.12}$$

where n is the index of refraction of the medium and k_e is the extinction coefficient, which is a constant relating the attenuation of the incident EM wave inside the solid to its penetration depth. For example, if an incident EM wave propagates into a solid at a distance equal to one wavelength in free space (i.e., $\lambda_0 = 2\pi c/\omega$), then its amplitude is decreased by a factor of $e^{-2\pi k_e}$, where k_e is the extinction coefficient of the solid. It is noted that the fundamental optical absorption coefficient α is related to k_e by $\alpha = 4\pi k_e/\lambda$, as will be shown later. Solving (9.11) and (9.12), one obtains the real and imaginary parts of the complex refractive index, which are given respectively by

$$n^2 - k_e^2 = \varepsilon_s, \tag{9.13}$$

$$2nk_e = \frac{\sigma}{\omega\varepsilon_0}. \tag{9.14}$$

Thus, the optical properties of a solid, as observed macroscopically, may be described in terms of the complex refractive index n^*. Now substituting (9.10) and (9.12) into (9.7), we see that the solution for the EM waves inside the solid becomes

$$\mathcal{E}_x = \mathcal{E}_0 \exp\left(\frac{-k_e\omega z}{c}\right) \exp\left[i\omega\left(\frac{nz}{c} - t\right)\right], \tag{9.15}$$

which shows that the speed of incident electric waves in a solid is reduced by a factor of n (n is the refractive index), and its amplitude decreases exponentially with distance. The attenuation of incident electric waves is associated with the absorption of EM energy by the dissipating medium. However, the optical constant commonly measured in a solid is not the extinction coefficient k_e, but the absorption coefficient α. The optical absorption coefficient is related to the Poynting vector of the EM wave energy flow by

$$S(z) = S_0 e^{-\alpha z}, \tag{9.16}$$

where $S(z)$ is the Poynting vector, which is proportional to the square of the amplitude of the electric waves (i.e., $|\mathcal{E}_x^2|$) given by (9.15). Thus, from (9.15) and (9.16), one obtains the optical absorption coefficient

$$\alpha = \frac{2k_e\omega}{c} = \frac{4\pi k_e}{\lambda_0}, \tag{9.17}$$

where λ_0 is the wavelength of the EM waves in free space. Thus, the extinction coefficient k_e can be determined from the optical absorption coefficient α of the semiconductor. It is noted that both the real ($n^2 - k_e^2$) and imaginary ($2nk_e$) parts of the complex refractive index n^* are quantities measured in a solid. In practice, ($n^2 - k_e^2$) and $2nk_e$ can be obtained by measuring the reflection and transmission coefficients of a solid.

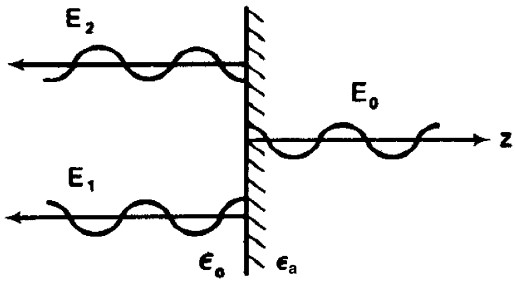


FIGURE 9.1. An eletromagnetic wave propagating into a solid under normal incidence.

To derive the reflection coefficient in a solid let us consider the case of normal incidence, as shown in Figure 9.1. If $\mathcal{E}_x(H_y)$ and $\mathcal{E}''_x(H''_y)$ denote the incident and reflected electric (magnetic) waves, and $\mathcal{E}'_x(H'_y)$ is the transmitted electric (magnetic) wave into the solid in the z-direction, then the transmitted wave for $z > 0$ can be expressed by

$$\mathcal{E}'_x = \mathcal{E}_0 \exp\left[i\omega\left(\frac{n^* z}{c} - t\right)\right]. \tag{9.18}$$

For $z < 0$ (i.e., in free space), the electric waves are composed of the incident and reflected waves, which can be expressed by

$$\mathcal{E}_x = \mathcal{E}_1 \exp\left[i\omega\left(\frac{z}{c} - t\right)\right] + \mathcal{E}_2 \exp\left[-i\omega\left(\frac{z}{c} + t\right)\right]. \tag{9.19}$$

The magnetic wave components polarized in the y-direction (i.e., H_y) may be related to the electric wave components in the x-direction by the characteristic impedance of the medium, which is given by

$$\frac{\mathcal{E}_x}{H_y} = \sqrt{\frac{\mu_0}{\mathcal{E}_0}} = Z_0, \quad \frac{\mathcal{E}'_x}{H'_y} = \sqrt{\frac{\mu'}{\varepsilon_0 \varepsilon_s}} = Z', \quad \frac{\mathcal{E}''_x}{H''_y} = -\sqrt{\frac{\mu_0}{\varepsilon_0}} = -Z_0. \tag{9.20}$$

Equation (9.20) relates the incident, transmitted, and reflected EM waves to the characteristic impedances Z_0 and Z' in free space and in the solid. The boundary conditions at the plane $z = 0$ requires that the tangential components of both $\mathcal{E}_x$ and H_y be continuous. Thus, one can write

$$\mathcal{E}'_x = \mathcal{E}_x + \mathcal{E}''_x \quad \text{and} \quad H'_y = H_y + H''_y. \tag{9.21}$$

Now solving (9.20) and (9.21) yields

$$\frac{\mathcal{E}''_x}{\mathcal{E}_x} = \frac{Z' - Z_0}{Z' + Z_0} = \frac{\sqrt{\varepsilon_0/\mu_0} - \sqrt{\varepsilon_0 \varepsilon_s/\mu'}}{\sqrt{\varepsilon_0/\mu_0} + \sqrt{\varepsilon_0 \varepsilon_s/\mu'}}. \tag{9.22}$$

From (9.20) one obtains

$$\frac{H''_y}{H_y} = -\frac{\mathcal{E}''_x}{\mathcal{E}_x} = -\frac{Z' - Z_0}{Z' + Z_0}. \tag{9.23}$$

Since the Poynting vector is equal to the product of the electric and magnetic field strengths (i.e., $|S| = \mathcal{E}_x H_y$, $|S''| = \mathcal{E}_x'' H_y''$), the reflection coefficient R can be obtained from (9.22) and (9.23) using the definition $R = |S''/S|$, and the result is

$$R = \left|\frac{S''}{S}\right| = \left(\frac{\mathcal{E}_x''}{\mathcal{E}_x}\right)\left(\frac{H_y''}{H_y}\right) = \left(\frac{Z' - Z_0}{Z' + Z_0}\right)^2$$
$$= \left[\frac{\sqrt{\varepsilon_0/\mu_0} - \sqrt{\varepsilon_0\varepsilon_s/\mu_0}}{\sqrt{\varepsilon_0/\mu_0} + \sqrt{\varepsilon_0\varepsilon_s/\mu_0}}\right]^2 = \left(\frac{n' - n_0}{n' + n_0}\right)^2. \tag{9.24}$$

For nonmagnetic materials, $\mu' = \mu_0$, $\varepsilon = \varepsilon_0\varepsilon_s$, and $n' = n + ik_e$; for free space, $\mathcal{E} = \varepsilon_0$ and $n_0 = 1$. Thus, the absolute value of the reflection coefficient for normal incidence can be written as

$$R = \left|\frac{(n-1)^2 + k_e^2}{(n+1)^2 + k_e^2}\right|, \tag{9.25}$$

where n and k_e are the index of refraction and the extinction coefficient of the solid, respectively.

The transmission coefficient T, defined as the ratio of the transmission power and the incident power, can be derived in a similar way as that of the reflection coefficient described above or using the relation $T = 1 - R$, which yields

$$T = \left|\frac{\mathcal{E}_x' H_y'}{\mathcal{E}_x H_y}\right| = \frac{4Z_0 Z'}{(Z' + Z_0)^2} = \frac{4n_0 n'}{(n_0 + n')^2}. \tag{9.26}$$

Since $n_0 = 1$ for free space, the absolute value of the transmission coefficient can be obtained from (9.26), which yields

$$T = \frac{4n}{(n+1)^2 + k_e^2}. \tag{9.27}$$

For normal incidence, it is seen from (9.25) and (9.27) that by measuring T and R one can determine both n and k_e. However, for incident angles other than normal incidence, the reflection coefficient will, in general, depend on the polarization, and from observation of different angles of incidence, both n and k_e values can be determined if k_e is not too small. If both n and k_e values are large, then R will approach unity.

An inspection of (9.13) reveals that the dielectric constant ε_s can also be determined directly from the refractive index n, provided that k_e is much smaller than unity. Values of n can be found directly from measurements of the reflection coefficient if k_e is very small.

There is considerable practical interest in measuring the transmission and reflection coefficients in free space under normal incidence using a thin plane-parallel sheet of crystal with refractive index n and thickness d. If I_0, I_t, and I_r denote the incident, transmitted, and reflected wave intensities through the thin specimen, then the normalized transmitted and reflected wave intensities can be expressed,

respectively, by

$$\frac{I_t}{I_0} = \frac{(1-R)^2\,e^{-\alpha d}\left(1+k_e^2/n^2\right)}{1-R^2\,e^{-2\alpha d}}, \tag{9.28}$$

$$\frac{I_r}{I_0} = \frac{R\left(1-e^{-2\alpha d}\right)}{1-R^2\,e^{-2\alpha d}}. \tag{9.29}$$

Equations (9.28) and (9.29) show that both n and k_e can be found by measuring I_t and I_r. For most transmission experiments, it is valid to assume that $k_e^2 \ll n^2$. If the sample thickness d is chosen such that $R^2\,e^{-2\alpha d} \ll 1$, then (9.28) becomes

$$\frac{I_t}{I_0} = (1-R)^2\,e^{-\alpha d}. \tag{9.30}$$

From (9.30), it is noted that the optical absorption coefficient α of a semiconductor near the band edge can be determined by measuring the transmission coefficient as a function of wavelength on two thin samples of different thicknesses without knowledge of the reflectance. This is valid as long as both samples have the same reflection coefficients at the front surface of the sample. For elemental semiconductors such as Si and Ge, the main contribution to the dielectric constant arises from electronic polarization. However, in compound semiconductors (such as III-V and II-VI compounds), both electronic and ionic polarizations can contribute to the dielectric constant. The increase in the degree of ionicity in these compounds relative to the group IV elements will lead to a significant difference between the static and optical (high-frequency) dielectric constants. The high-frequency dielectric constant ε_s^∞ is equal to n^2. The static dielectric constant ε_s can be calculated using the relation

$$\varepsilon_s = \varepsilon_s^\infty \left(\frac{\omega_l}{\omega_t}\right)^2, \tag{9.31}$$

where ω_l and ω_t are the longitudinal- and transverse-mode optical phonon frequencies, respectively. Table 9.1 lists values of dielectric constants and refractive indices for Si, Ge, and some III-V and II-VI compound semiconductors.

9.3. Free-Carrier Absorption Process

When the energy of incident EM radiation is smaller than the band gap energy (i.e., $h\nu \le E_g$) of the semiconductor, excitation of electrons from the valence band into the conduction band will not occur. Instead, the absorption of incident EM radiation will result in the excitation of lattice phonons and the acceleration of free electrons inside the conduction band. In the conduction band, free-carrier absorption is proportional to the density of conduction electrons. Since free-carrier absorption involves only electronic transitions within the conduction band, one can apply the classical equations of motion to deal with the interaction between the EM waves and the conduction electrons. The equation of motion for an electron due

TABLE 9.1. Refractive indices and dielectric constants for Si, Ge, and some III-V and II-VI semiconductors.

Materials	n	ε_s	ε_s^∞
Si	3.44	11.8	11.6
Ge	4.00	16	15.8
InSb	3.96	17	15.9
InAs	3.42	14.5	11.7
GaAs	3.30	12.5	10.9
GaP	2.91	10	8.4
CdS	2.30	8.6	5.2
CdSe	2.55	9.2	6.4
CdTe	2.67	9.7	7.1
ZnS	2.26	8.1	5.1
ZnSe	2.43	8.7	5.9

to a time-varying electric wave (i.e., $\mathcal{E}_0 \, e^{i\omega t}$) of frequency ω propagating in the z-direction is given by

$$m^* \frac{\partial^2 z}{\partial t^2} + \left(\frac{m^*}{\tau} \right) \frac{\partial z}{\partial t} = q \mathcal{E}_0 \, e^{i\omega t}, \tag{9.32}$$

where τ is the relaxation time and m^* is the effective mass of electrons in the conduction band. The solution of (9.32) is given by

$$z = \frac{(q\mathcal{E}_0/m^*) \, e^{i\omega t}}{\left(i\omega/\tau - \omega^2 \right)}. \tag{9.33}$$

If the electron density in the conduction band is equal to N_0, then the total polarization P, which is equal to the product of displacement z and electron density N_0, can be expressed by

$$P = q N_0 z. \tag{9.34}$$

The polarizability p^*, which is defined as the polarization per unit electric field, can be written as

$$p^* = \frac{q N_0 z}{\mathcal{E}_0}. \tag{9.35}$$

The complex dielectric constant ε_s^* given by (9.11) is related to the polarizability p^* by

$$\varepsilon_s^* = \varepsilon_s' - i\varepsilon_s'' = n^{*^2} = \varepsilon_s + \frac{p^*}{\varepsilon_0} = \varepsilon_s + \frac{\left(N_0 q^2/m^* \mathcal{E}_0 \right)}{\left(i\omega/\tau - \omega^2 \right)}. \tag{9.36}$$

Note that the second term on the right-hand side of (9.36) is due to the contribution of free-carrier absorption. Thus, from (9.36), the real and imaginary parts of the complex dielectric constant can be written as

$$\varepsilon_s' = n^2 - k_e^2 = \varepsilon_s - \frac{\tau \sigma_0}{\varepsilon_0 \left(1 + \omega^2 \tau^2 \right)} \tag{9.37}$$

and

$$\varepsilon_s'' = 2nk_e = \frac{\sigma_0}{\omega\varepsilon_0\left(1+\omega^2\tau^2\right)}, \tag{9.38}$$

where $\sigma_0 = N_0q^2\tau/m^*$ is the dc electrical conductivity. In (9.37) and (9.38), it is assumed that τ is a constant and is independent of energy.

Solving (9.17) and (9.38), one obtains the optical absorption coefficient as

$$\alpha = \frac{4\pi k_e}{\lambda_0} = \frac{\sigma_0}{nc\varepsilon_0\left(1+\omega^2\tau^2\right)}, \tag{9.39}$$

which shows the frequency dependence of the optical absorption coefficient. Two limiting cases, for $\omega\tau \gg 1$ and $\omega\tau \ll 1$, are discussed next.

(i) Long-wavelength limit ($\omega\tau \ll 1$). In this case, the absorption coefficient given by (9.39) becomes

$$\alpha = \frac{\sigma_0}{nc\varepsilon_0}. \tag{9.40}$$

The real part of the dielectric constant in (9.37) is reduced to

$$\varepsilon_s' = \varepsilon_s - \frac{\tau\sigma_0}{\varepsilon_0}. \tag{9.41}$$

Equation (9.40) shows that the absorption coefficient is independent of frequency, but depends on temperature through σ_0. For example, for an n-type germanium sample with $\tau = 10^{-12}$ s this corresponds to a wavelength of about 2 mm. For a lightly doped semiconductor with large dielectric constant, the contribution of $\tau\sigma_0/\varepsilon_0$ in (9.41) to the real part of the dielectric constant ε_s' is quite small, and hence ε_s' is equal to the dielectric constant ε_s of the semiconductor.

For heavily doped semiconductors with large σ_0, the value of $\tau\sigma_0/\varepsilon_0$ becomes much larger than ε_s and hence ε_s' becomes negative. This corresponds to the metallic case. If one assumes $\varepsilon_s \ll \tau\sigma_0/\varepsilon_0$, then solving (9.41) and (9.37), one obtains the real part of the dielectric constant as

$$\varepsilon_s' = n^2 - k_e^2 = -\frac{\tau\sigma_0}{\varepsilon_0}. \tag{9.42}$$

Similarly, from (9.38) one obtains the imaginary part of the dielectric constant as

$$\varepsilon_s'' = 2nk_e = \frac{\sigma_0}{\omega\varepsilon_0}. \tag{9.43}$$

Now, solving (9.42) and (9.43) yields

$$\omega\tau = -\frac{\left(n^2 - k_e^2\right)}{2nk_e}. \tag{9.44}$$

From (9.44), for $\omega\tau \ll 1$ we have $k_e \approx n$. Thus, setting $n = k_e$ in (9.43), we see that the refractive index is given by

$$n = \left(\frac{\sigma_0}{2\omega\varepsilon_0}\right)^{1/2}. \tag{9.45}$$

The absorption coefficient can be deduced from (9.40) and (9.45), and the result yields

$$\alpha = \left(\frac{2\sigma_0\omega}{\varepsilon_0 c^2}\right)^{1/2}, \tag{9.46}$$

which shows that α is proportional to the square root of the frequency. Therefore, in this case, the material exhibits metallic behavior. This corresponds to the well-known skin effect, in which the penetration depth (δ) of the incident EM wave is inversely proportional to the square root of the frequency and the electrical conductivity.

(ii) Short-wavelength limit ($\omega\tau \gg 1$). This usually occurs in the wavelength regime extending upward from far-IR toward the fundamental absorption edge of the semiconductor. The short-wavelength free-carrier absorption becomes negligible when the photon energy exceeds the band gap energy of the semiconductor. To understand the free-carrier absorption process in the short-wavelength limit, one can solve (9.37) to (9.39) to obtain

$$\varepsilon_s' = \varepsilon_s - \frac{\sigma_0}{\varepsilon_0\omega^2\tau} = n^2 - k_e^2, \tag{9.47}$$

$$\varepsilon_x'' = 2nk_e = \frac{\sigma_0}{\varepsilon_0\omega^3\tau^2}, \tag{9.48}$$

$$\alpha = \frac{\sigma_0}{nc\varepsilon_0\omega^2\tau^2} = \frac{N_0 q^3 \lambda_0^2}{4\pi^2 c^3 m^{*^2} \mu n \varepsilon_0}. \tag{9.49}$$

Equation (9.49) shows that for $\omega\tau \gg 1$, the absorption coefficient is directly proportional to the square of the wavelength, which has been observed in a number of semiconductors. Figure 9.2 shows a graph of the absorption coefficient versus the square of the wavelength for two n-type InSb specimens with different dopant concentrations.[1] The results are in good agreement with the prediction given by (9.49).

It is seen from (9.47) that ε_s' changes sign from positive to negative as ω decreases. The condition for which $\varepsilon_s' = 0$ corresponds to total internal reflection, and the frequency at which this occurs is called the plasma resonance frequency, ω_p. Solving (9.47), one obtains

$$\omega_p = \left(\frac{\sigma_0}{\varepsilon_0\varepsilon_s\tau}\right)^{1/2} = \left(\frac{N_0 q^2}{m^*\varepsilon_0\varepsilon_s}\right)^{1/2}, \tag{9.50}$$

where ω_p is the frequency at which the classical undamped plasma of free electrons exhibits its normal mode of oscillation. For a germanium sample with $N_0 = 10^{16}\,\text{cm}^{-3}$, $m^* = 0.12\,m_0$, and $\varepsilon_s = 16$, one finds that ω_p is equal to 2×10^2 GHz, which falls into the rather difficult millimeter-wavelength regime. In order

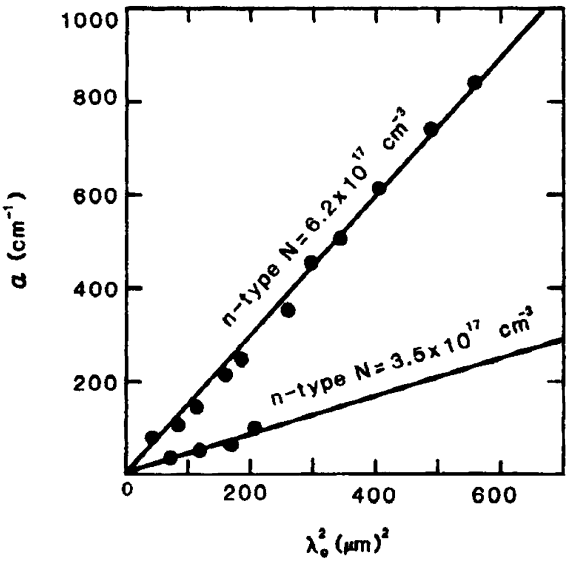


FIGURE 9.2. Optical absorption coefficient versus the square of the wavelength for two InSb specimens with different doping densities over the wavelength range in which free-carrier absorption is dominant. After Moss,[1] by permission.

to observe plasma resonance in the microwave frequency range, one should use an ultrapure semiconductor specimen for the experiment. Otherwise, the experiment must be performed at extremely low temperatures. For a germanium crystal, a carrier concentration of 10^{13} cm^{-3} or less is required for the plasma resonance to be observed in the microwave-frequency range. For metals, since the electron concentration is very high ($\approx 10^{22}$ cm^{-3}), the plasma resonance frequency usually falls in the solar blind UV spectral range, which corresponds to photon energies of 10–20 eV. Free-carrier absorption has been used extensively in determining the relaxation time constant and the conductivity effective mass of electrons in a semiconductor.

9.4. Fundamental Absorption Process

The fundamental absorption process takes place when photons with energies greater than the band gap energy of the semiconductor (i.e., $h\nu \geq E_g$) are absorbed in a semiconductor. This process usually results in the generation of electron–hole pairs in the semiconductor. For most semiconductors, the fundamental absorption process may occur in the UV, visible, and IR wavelength regimes. It is the most important optical absorption process because important photoelectric effects for generating excess electron–hole pairs in a semiconductor are based on such absorption processes.

There are two types of optical transition associated with the fundamental absorption process, namely, direct and indirect band-to-band transitions, as shown in Figures 9.3a and b. In a direct transition only one photon is involved, while in an

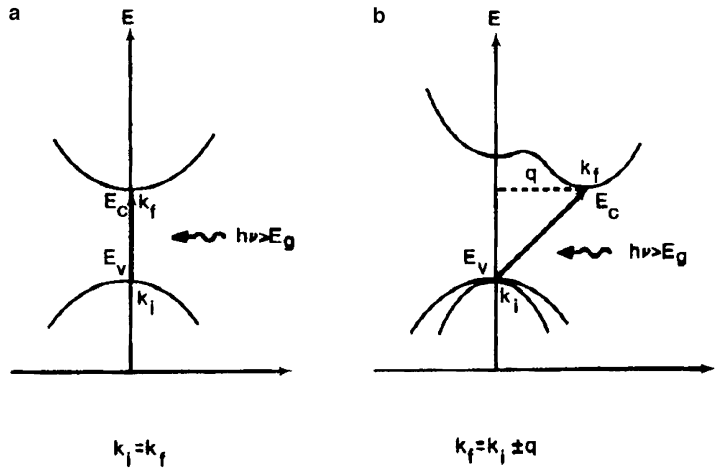


FIGURE 9.3. Direct and indirect transitions associated with the fundamental absorption processing in a semiconductor.

indirect transition additional energy is supplied or released in the form of phonons. The absorption coefficients associated with these two transition processes depend on the probability per unit time that an electron makes a transition from the valence band into the conduction band when an incident photon is absorbed. The transition probability P_i can be calculated using first-order time-dependent perturbation theory. We have

$$P_\mathrm{i} = \left(\frac{2\pi}{h}\right) |M_\mathrm{if}|^2 \, g_\mathrm{n}(E), \tag{9.51}$$

where P_i is the transition probability per unit time from the initial state k_i in the valence band to the final state k_f in the conduction band, M_if denotes the matrix element due to perturbation that connects the initial states k_i and the final states k_f of the system, and $g_\mathrm{n}(E)$ is the density of final states in the conduction band.

In the present case, the perturbation is due to incident EM radiation, and the matrix element corresponding to the electric dipole transition is given by

$$\begin{aligned} M_\mathrm{if} &= \int \psi_\mathrm{i}^* \nabla_\mathrm{r} \psi_\mathrm{f} \, \mathrm{d}^3 r \\ &= \int u_\mathrm{v}^*(k_\mathrm{i}, r) \, \mathrm{e}^{-ik_\mathrm{i}\cdot r} \nabla_\mathrm{r} u_\mathrm{c}(k_\mathrm{f}, r) \, \mathrm{e}^{ik_\mathrm{f}\cdot r} \mathrm{d}^3 r \\ &= \int u_\mathrm{v}^*(k_\mathrm{i}, r) \nabla_\mathrm{r} u_\mathrm{c}(k_\mathrm{f}, r) \, \mathrm{e}^{i(k_\mathrm{f}-k_\mathrm{i})\cdot r} \, \mathrm{d}^3 r + ik_\mathrm{f} \int u_\mathrm{v}^*(k_\mathrm{f}, r) u_\mathrm{c}(k_\mathrm{f}, r) \, \mathrm{e}^{i(k_\mathrm{f}-k_\mathrm{i})\cdot r} \mathrm{d}^3 r, \end{aligned} \tag{9.52}$$

where ψ_i and ψ_f denote the electron wave functions in the valence and conduction bands, respectively; $u_\mathrm{v}(k_\mathrm{i}, r)$ and $u_\mathrm{c}(k_\mathrm{f}, r)$ are the Bloch functions for the valence and conduction bands, respectively. Both terms on the right-hand side of (9.52)

contain the factor $e^{i(k_f - k_i)\cdot r}$, which oscillates rapidly. Thus, the integrand of (9.52) will vanish unless

$$k_f = k_i, \tag{9.53}$$

which is the condition of momentum conservation for such a transition. In fact, the contribution of photon momentum in (9.53) is negligible, because it is very small compared to the crystal momentum. It is noted that the first term on the right-hand side of (9.52) is known as the allowed transition; its value is real and independent of the wave vector k_f of the final state. The second term on the right-hand side of (9.52) is imaginary, and it depends on the wave vector of the final state k_f. Transitions associated with the second term are called the forbidden transitions. Since the absorption coefficient is directly related to the rate of transition probability P_{if}, equation (9.52) can be employed to derive the absorption coefficient for the direct and indirect interband transitions taking place between the valence band and the conduction band of a semiconductor.

9.4.1. Direct Transition Process

The direct (or vertical) transition shown in Figure 9.3a is the dominant absorption process taking place in a direct band gap semiconductor when the conduction band minimum and the valence band maximum are located at the same k-value in the reciprocal space (i.e., typically at the Γ-point of the Brillouin zone center). In order to derive an expression of the absorption coefficient near the conduction band minimum, it is necessary to find the density of final states $g_n(E)$ in (9.51). It is noted that electron energy in the conduction band can be expressed by

$$E_n = E_c + \frac{\hbar^2 k^2}{2m_n^*}, \tag{9.54}$$

and in the valence band by

$$E_p = E_v - \frac{\hbar^2 k^2}{2m_p^*}. \tag{9.55}$$

The photon energy corresponding to such a vertical transition can be written as

$$h\nu = E_n - E_p = E_g + \frac{\hbar^2 k^2}{2m_r^*}, \tag{9.56}$$

where E_g is the band gap energy of the semiconductor and $m_r^* = m_n^* m_p^*/(m_n^* + m_p^*)$ is the reduced electron effective mass.

Equations (9.54) through (9.56) allows one to express the density of final states for the conduction band with a parabolic band structure as

$$g_n(E) = \left(\frac{4\pi}{h^3}\right)(2m_r^*)^{3/2}(h\nu - E_g)^{1/2}. \tag{9.57}$$

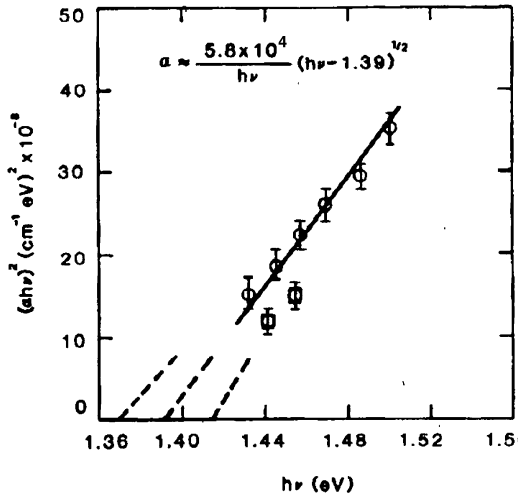


FIGURE 9.4. Direct transition in a p-type GaAs specimen with $N_A = 10^{17}$ cm^{-3} and an absorption coefficient greater than 9×10^3 cm^{-1}. The threshold energy is (1.39 ± 0.02) eV. After Kudman and Seidel,[2] by permission.

From (9.52) and (9.57), we see that the absorption coefficient for a direct allowed transition can be expressed by

$$\alpha_d^a = K_d^a (h\nu - E_g)^{1/2}. \tag{9.58}$$

In the direct allowed transitions, the square of the matrix element (i.e., $|M_{if}|^2$) is independent of the wave vector, and hence K_d^a is a constant, independent of electron energy.

Equation (9.58) shows that for an allowed direct optical transition, the optical absorption coefficient α_d^a varies as $(h\nu - E_g)^{1/2}$. Therefore, a plot of α_d^2 versus $h\nu$ near the fundamental absorption edge allows one to determine the band gap energy of a semiconductor. This is illustrated in Figure 9.4 for a p-type GaAs[2] specimen, the intercept of this plot with the horizontal axis yields the band gap energy of GaAs.

In the forbidden direct transitions, as given by the second term of (9.52), the matrix element M_{if} is proportional to k_f, and hence the optical absorption coefficient in this case is given by

$$\alpha_d^f = K_d^f (h\nu - E_g)^{3/2}. \tag{9.59}$$

The energy dependence ($\sim E^{3/2}$) of α_d^f given by (9.59) is due to the fact that the transition probability for the direct forbidden transitions varies with the product of k_f^2 ($\sim E = (h\nu - E_g)$) and the density-of-states function ($\sim E^{1/2}$).

9.4.2. *Indirect Transition Process*

For an indirect band gap semiconductor, the conduction band minimum and the valence band maximum are not located at the same k-value in the reciprocal space. Therefore, the indirect optical transition induced by photon absorption is

usually accompanied by the simultaneous absorption or emission of a phonon. As illustrated in Figure 9.3b, conservation of momentum in this case is given by

$$\boldsymbol{k}_{\mathrm{f}} = \boldsymbol{k}_{\mathrm{i}} \pm \boldsymbol{q}, \tag{9.60}$$

where $\boldsymbol{k}_{\mathrm{f}}$ and $\boldsymbol{k}_{\mathrm{i}}$ denote the wave vectors of the final and initial states of electrons, respectively, and $\boldsymbol{q}$ is the phonon wave vector. The plus sign in (9.60) corresponds to phonon emission, and the minus sign is for phonon absorption. The conservation of energy for the indirect optical transitions requires that

$$h\nu = E_{\mathrm{n}} - E_{\mathrm{p}} \pm \hbar\omega_{\mathrm{q}} = E_{\mathrm{g}} + \frac{\hbar^2(k_{\mathrm{n}} - k_{\mathrm{c}})^2}{2m_{\mathrm{n}}^*} + \frac{\hbar^2 k_{\mathrm{p}}^2}{2m_{\mathrm{p}}^*} \pm \hbar\omega_{\mathrm{q}} \tag{9.61}$$

and

$$E_{\mathrm{n}} = E_{\mathrm{c}} + \frac{\hbar^2(k_{\mathrm{n}} - k_{\mathrm{c}})^2}{2m_{\mathrm{n}}^*}, \tag{9.62}$$

$$E_{\mathrm{p}} = E_{\mathrm{v}} - \frac{\hbar^2 k_{\mathrm{p}}^2}{2m_{\mathrm{p}}^*}, \tag{9.63}$$

where E_{n} is the electron energy in the conduction band, E_{p} is the electron energy in the valence band, and $(k_{\mathrm{n}} - k_{\mathrm{c}}) \ll k_{\mathrm{c}}$.

It is noted from (9.61) that in an indirect optical transition the conservation of energy is accompanied by the emission or absorption of a phonon. The plus sign in (9.61) is for phonon emission and the minus sign is for phonon absorption; k_{p} denotes the initial state in the valence band, k_{c} is the state at the conduction band minimum, and k_{n} is the final state in the conduction band. Now consider the case in which transition from the k_{p} state in the valence band is induced by a photon with energy $h\nu$. The density of states in the valence band can be described by

$$g_v(E_{\mathrm{p}}) = A_v E_{\mathrm{p}}^{1/2}\,, \tag{9.64}$$

where $E_{\mathrm{p}} = E_v - \Delta E$, ΔE is a small energy interval in the valence band in which transitions can take place, and $A_\nu = (4\pi/h^3)\,(2m_{\mathrm{p}}^*)^{3/2}$.

The density of final conduction band states involving phonon absorption is given by

$$\begin{aligned} g_{\mathrm{c}}(E_{\mathrm{n}}) &= A_{\mathrm{c}}(E_{\mathrm{n}} - E_{\mathrm{c}})^{1/2} = A_{\mathrm{c}}(h\nu - E_{\mathrm{g}} - E_{\mathrm{p}} + \hbar\omega_{\mathrm{q}})^{1/2} \\ &= A_{\mathrm{c}}(\Delta E - E_{\mathrm{p}})^{1/2}\,, \end{aligned} \tag{9.65}$$

where $A_{\mathrm{c}} = (4\pi/h^3)\,(2m_{\mathrm{n}}^*)^{3/2}$. Equation (9.65) is obtained by solving (9.61) through (9.64). In (9.65), the relation $h\nu = (E_{\mathrm{g}} \pm \hbar\omega_{\mathrm{q}} + \Delta E)$ is used in the derivation. Therefore, the total effective density of states for transitions involving absorption and emission of a phonon can be expressed by

$$\begin{aligned} g(h\nu) &= \int_0^{\Delta E} g_{\mathrm{c}}(E_{\mathrm{n}}) g_{\mathrm{v}}(E_{\mathrm{p}})\,\mathrm{d}E_{\mathrm{p}} = A_{\mathrm{c}} A_{\mathrm{v}} \int_0^{\Delta E} (\Delta E - E_{\mathrm{p}})^{1/2} E_{\mathrm{p}}^{1/2}\;\mathrm{d}E_{\mathrm{p}} \\ &= K_{\mathrm{i}}^a \Delta E^2 = K_{\mathrm{i}}^a (E_{\mathrm{v}} - E_{\mathrm{p}})^2 = K_{\mathrm{i}}^a (h\nu - E_{\mathrm{g}} \pm \hbar\omega_{\mathrm{q}})^2\,, \end{aligned} \tag{9.66}$$

where the plus sign denotes phonon absorption and the minus sign phonon emission. Note that the integral on the right-hand side of (9.66) is carried out by letting $u = E_p^{1/2}$, so that

$$\begin{aligned}
&\int_0^{\Delta E} (\Delta E - E_p)^{1/2} E_p^{1/2}\, dE_p \\
&\quad = 2\int_0^{\Delta E^{1/2}} u(\Delta E - u^2)^{1/2}\, du \\
&\quad = 2\left\{-\frac{u}{4}(\Delta E - u^2)^{3/2} + \frac{\Delta E}{8}\left[u(\Delta E - u^2)^{1/2} + \Delta E \sin^{-1}\left(\frac{u}{\Delta E^{1/2}}\right)\right]\right\}_0^{\Delta E^{1/2}} \\
&\quad = \frac{\pi \Delta E^2}{8} = \frac{\pi(h\nu - E_g + \hbar\omega_q)^2}{8}.
\end{aligned} \tag{9.67}$$

The probability of phonon absorption and phonon emission is directly proportional to average phonon density, which is given by

$$\langle n_q \rangle = \left(e^{\hbar\omega_q / k_B T} - 1\right)^{-1}. \tag{9.68}$$

Combining (9.66) and (9.68) one obtains the optical absorption coefficient due to the indirect transitions with phonon absorption as

$$\alpha_{ia} = \langle n_q \rangle g(h\nu) = K_{ia} \frac{\left(h\nu - E_g + \hbar\omega_q\right)^2}{\left(e^{\hbar\omega_q / k_B T} - 1\right)}. \tag{9.69}$$

Similarly, for transitions involving phonon emission, the optical absorption coefficient can be expressed by

$$\alpha_{ie} = \langle n_q + 1 \rangle g(h\nu) = K_{ie} \frac{\left(h\nu - E_g - \hbar\omega_q\right)^2}{\left(1 - e^{-\hbar\omega_q / k_B T}\right)}. \tag{9.70}$$

Now combining (9.69) and (9.70), we see that the optical absorption coefficients for the indirect allowed transitions involving both the emission and absorption of a phonon can be written as

$$\alpha_i = \alpha_{ie} + \alpha_{ia} = K_i \left\{ \frac{(h\nu - E_g - \hbar\omega_q)^2}{\left(1 - e^{-\hbar\omega_q / k_B T}\right)} + \frac{(h\nu - E_g + \hbar\omega_q)^2}{\left(e^{\hbar\omega_q / k_B T} - 1\right)} \right\}. \tag{9.71}$$

The first term in (9.71) is due to phonon emission, while the second term is attributed to phonon absorption. It is clearly shown in Figure 9.5 that the optical absorption coefficient curve for the indirect allowed transitions involving phonon absorption will extend to longer wavelengths than those associated with phonon emission. The optical absorption coefficient for the indirect allowed transitions varies with the square of photon energy. A plot of $\alpha_i^{1/2}$ versus $h\nu$ at different temperatures should yield a straight line, and its intercept with the horizontal axis allows one to determine the phonon energy and the energy band gap of a semiconductor.

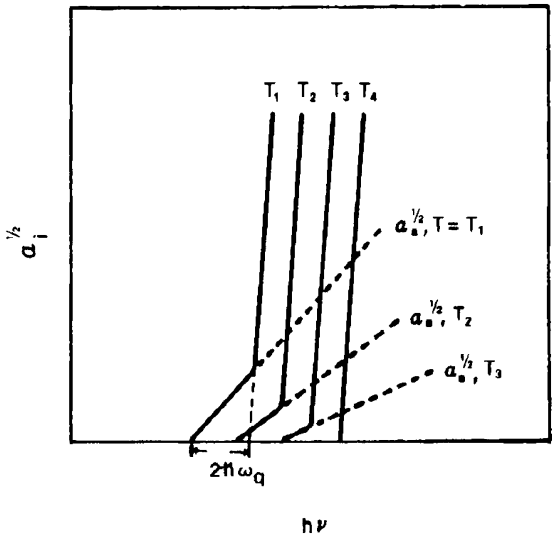


FIGURE 9.5. $\alpha_i^{1/2}$ versus photon energy $h\nu$ for indirect optical transitions with temperature as a parameter. Note that $T_1 > T_2 > T_3 > T_4$.

Figure 9.5 shows a plot of $\alpha_i^{1/2}$ versus $h\nu$ involving the emission and absorption of a phonon for four different temperatures. According to (9.70), two straight-line segments can be observed in the $\alpha_i^{1/2}$ versus $h\nu$ plot. For small photon energy, only α_{ia} (i.e., associated with phonon absorption) contributes, and the $\alpha_{ia}^{1/2}$ versus $h\nu$ plot intersects the axis at $h\nu = E_g - h\omega_q$. For $h\nu > E_g + \hbar\omega_q$, α_{ie} becomes dominant at lower temperatures. Since the intersection of $\alpha_{ie}^{1/2}$ versus $h\nu$ occurs at $h\nu = E_g + \hbar\omega_q$, one can determine both the energy band gap and the phonon energy from this plot. Figure 9.6 shows the square root of the absorption coefficient versus photon energy near the fundamental absorption edge of a germanium crystal with temperature as a parameter.[3] The results show that the phonon emission process becomes dominant for $T < 20$ K.

Several effects could influence the accuracy of determining the band gap energy from the optical absorption measurements in a semiconductor. The first effect is due to the Burstein shift in a degenerate semiconductor. In a heavily doped n-type semiconductor, the Fermi level lies inside the conduction band. Therefore, in order for photon-generated electrons to make transitions from the valence band into the conduction band, the photon energy must be greater than the band gap energy of the semiconductor so that electrons can be excited into the empty states above the Fermi level in the conduction band. This shifts the optical absorption edge to a higher energy with increased doping concentration. This problem is particularly severe for small-band gap semiconductors such as InSb and InAs, since the electron effective masses and densities of states in the conduction band are small for these materials. In calculating the Burstein shift, the effect of energy-band nonparabolicity should also be considered. The second effect is related to the formation of impurity band-tail states (or impurity bands) arising from high concentrations of shallow

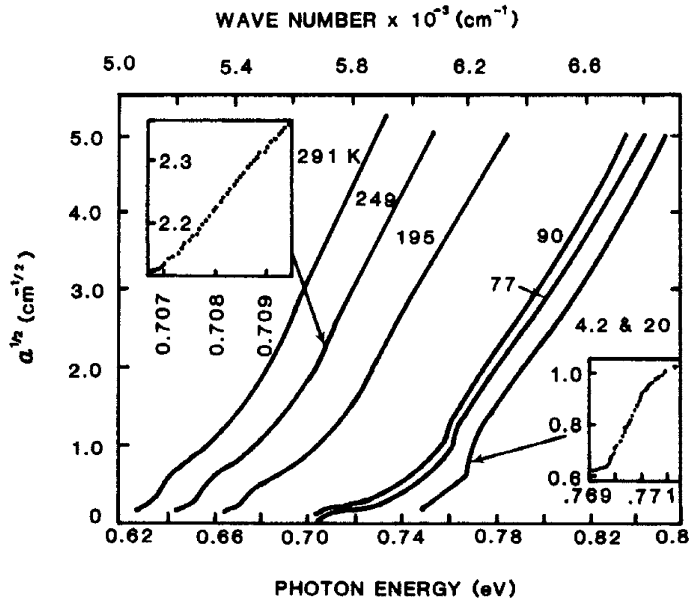


FIGURE 9.6. Square root of the absorption coefficient versus photon energy for a germanium specimen with temperature as a parameter. The inserts show the spectral resolution. After Macfarlane et al.,[3] by permission.

impurities or defects, which can merge into the conduction band (for n-type) or the valence band (for p-type). This effect will result in an exponential absorption edge in the semiconductor. The third effect is associated with exciton formation in the semiconductor. An exciton is an electron–hole pair bound together by Coulombic interaction. Excitons may be free, bound, or constrained to a surface, or associated with a defect complex. The binding energies for excitons are slightly below the conduction band edge, and hence exciton features are sharp peaks just below the absorption edge. Excitons are usually observed at low temperatures and become dissociated into free carriers at room temperature.

It is seen that the absorption coefficient increases rapidly above the fundamental absorption edge (i.e., $h\nu \geq E_g$). In the visible spectral range, values of the absorption coefficient for most semiconductors may vary from 10^3 to 10^5 cm^{-1}. In general, the magnitude of the absorption coefficient represents the degree of interaction between the semiconductor and the incident photons. The internal photoelectric effects in a semiconductor are closely related to the optical absorption coefficient. Experimental results of absorption coefficient versus photon energy for some elemental and compound semiconductors (Si, Ge, GaAs, GaP, and InSb) are shown in Figures 9.7 through Figure 9.10.[4–7] Information concerning the optical absorption coefficient versus photon energy is essential for analyzing the photoelectric effects in a semiconductor.

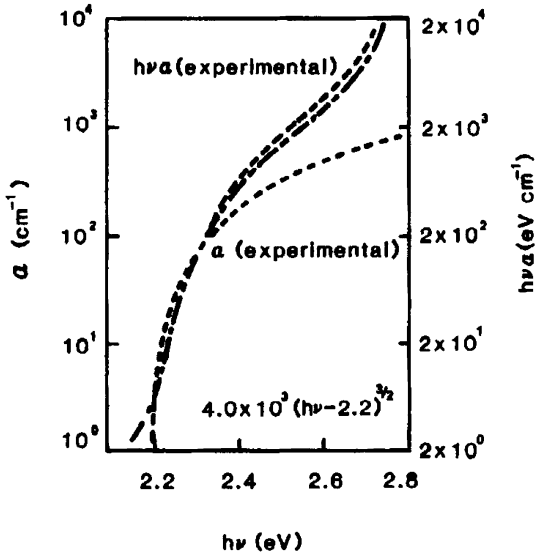


FIGURE 9.7. Absorption coefficient versus photon energy for a GaP sample at room temperature. After Spitzer et al.,[4] by permission.

9.5. The Photoconductivity Effect

In this section, the photoconductivity effect in a semiconductor is described. In the absence of illumination, the dark conductivity of a semiconductor is given by

$$\sigma_0 = q(n_0\mu_n + p_0\mu_p), \tag{9.72}$$

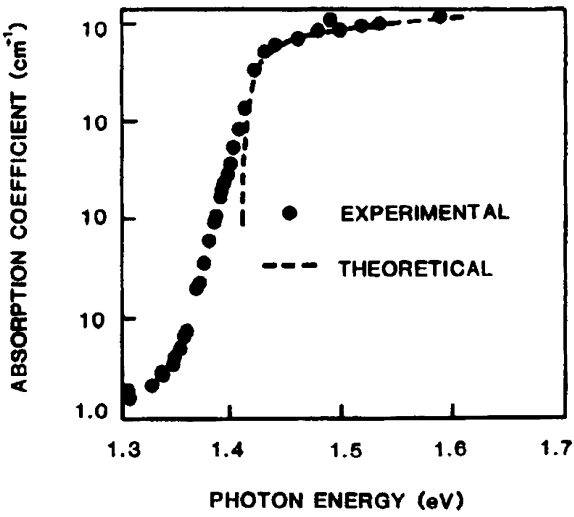


FIGURE 9.8. Absorption coefficient versus photon energy for a GaAs sample at room temperature. After Moss and Hawkins,[5] by permission.

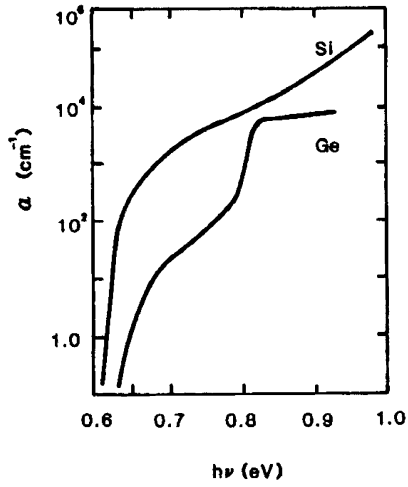


FIGURE 9.9. Absorption coefficient versus photon energy for silicon and germanium crystals measured at 300 K. After Dash and Newman,[6] by permission.

where n_0 and p_0 denote the densities of electrons and holes in thermal equilibrium, while μ_n and μ_p are the electron and hole mobilities, respectively.

When photons with energies equal to or greater than the band gap energy ($h\nu \geq E_g$) of a semiconductor are absorbed in a semiconductor, intrinsic photoconductivity results. The absorbed photons create excess electron–hole pairs (i.e., Δn and Δp), and as a result the densities of electrons and holes (i.e., n and p) increase above their equilibrium values of n_0 and p_0 (i.e., $n = n_0 + \Delta n$, $p = p_0 + \Delta p$).

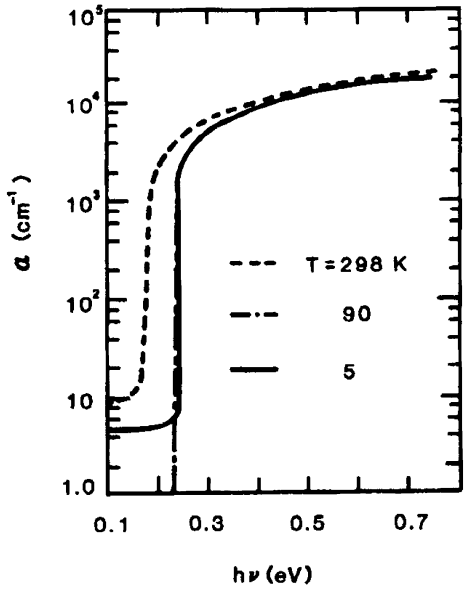


FIGURE 9.10. Absorption coefficient versus photon energy for a pure InSb sample measured at three different temperatures. After Johnson,[7] by permission.

The photoconductivity is defined as the net change in electrical conductivity under illumination and can be expressed by

$$\Delta\sigma = \sigma - \sigma_0 = q\left(\Delta n\mu_{\mathrm{n}} + \Delta p\mu_{\mathrm{p}}\right), \tag{9.73}$$

where Δn and Δp are the excess electron and hole densities, respectively.

In a degenerate semiconductor, Δp and Δn are generally much smaller than p_0 and n_0, and the effect of incident photons can be considered as a small perturbation. However, in an insulator or a nondegenerate semiconductor, values of Δn and Δp can become comparable or larger than their equilibrium carrier densities. If the effect of electron or hole trapping by the defect levels is negligible and the semiconductor remains neutral under illumination, then $\Delta n = \Delta p$ holds throughout the specimen.

Depending on the incident photon energies, there are two types of photoconduction processes that are commonly observed in a semiconductor. One type of photoconduction process is known as the intrinsic photoconductivity (PC), in which the excess electron–hole pairs are generated in the semiconductor by the absorption of photons with energies greater than the band gap energy of the semiconductor (i.e., $h\nu \geq E_{\mathrm{g}}$). This type of photoconduction process is illustrated in Figure 9.11a. The other type of photoconduction process is known as extrinsic photoconductivity, in which electrons (or holes) are excited from the localized donor (or acceptor) states into the conduction (or valence) band states by the absorption of photons with energy equal to or greater than the activation energy of the donor (or acceptor) levels, but is less than the band gap energy of the semiconductor (i.e., $E_{\mathrm{D}} \leq h\nu \leq E_{\mathrm{g}}$ for n-type conduction, and $E_{\mathrm{A}} \leq h\nu \leq E_{\mathrm{g}}$ for p-type conduction). This is illustrated in Figure 9.11b.

In intrinsic photoconduction, both the photogenerated electrons and holes participate in the photoconduction process, and the photoconductivity is described by (9.73). However, for extrinsic photoconductivity, the photoconduction process

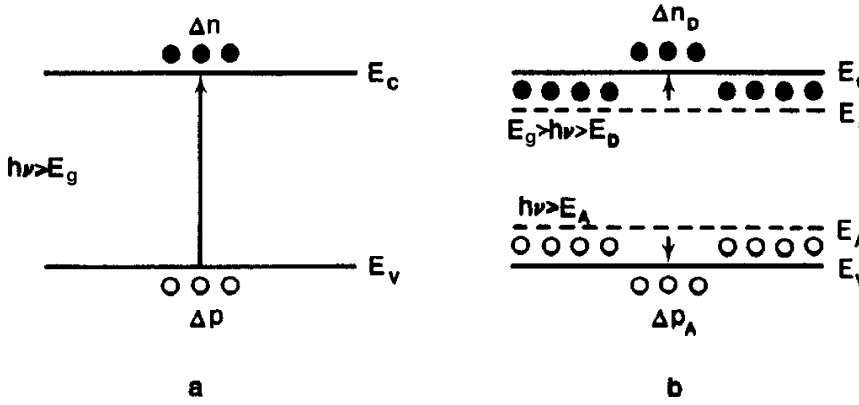


FIGURE 9.11. (a) Intrinsic and (b) extrinsic photoconductivity in a semiconductor.

usually involves only one type of charge carrier (i.e., either electrons or holes), and the expressions for the extrinsic photoconductivity are given by

$$\Delta\sigma_{\mathrm{n}} = qn_{\mathrm{D}}\mu_{\mathrm{n}} \quad \text{for n-type,} \tag{9.74}$$

$$\Delta\sigma_{\mathrm{p}} = qp_{\mathrm{A}}\mu_{\mathrm{p}} \quad \text{for p-type,} \tag{9.75}$$

where n_{D} and p_{A} are the photogenerated excess electron and hole densities from the donor and acceptor centers, respectively.

An extrinsic photoconductor usually operates at cryogenic temperatures because at very low temperatures freeze-out occurs for electrons in the conduction band states or for holes in the valence band states. The transition of electrons from the conduction band states to the shallow-donor states or of holes from the valence band states to the shallow-acceptor states because of the freeze-out effect is the basis for extrinsic photoconductivity. At very low temperatures, the electrical conductivity under dark condition and background noise are generally very low. When photons with energies of $E_{\mathrm{D}} \leq h\nu \leq E_{\mathrm{g}}$ impinge on an n-type specimen, the electrical conductivity of the sample will increase dramatically by the absorption of these incident photons. These photons excite the electrons in the shallow-donor impurity states into the conduction band states, resulting in an increase of electrical conductivity in the sample. The sensitivity of extrinsic photoconductivity depends greatly on the density of sensitizing shallow-impurity centers and the thickness of the specimen. Extrinsic photoconductivity has been widely used in long-wavelength IR detection. For example, a Cu-doped germanium extrinsic photoconductor operating at 4.2 K can be used to detect photons with wavelengths ranging from 2.5 to 30 μm, while a Hg-doped germanium photodetector operating at 28 K can be used for 10.6-μm wavelength detection.

Figure 9.12 shows a schematic diagram of an intrinsic photoconductor under illumination and bias conditions. In the intrinsic photoconduction process, electron–hole pairs are generated in a semiconductor when photons with energies exceeding the band gap energy of the semiconductor are absorbed. The rate of generation of electron–hole pairs per unit volume per unit time can be written as

$$g_{\mathrm{E}} = \begin{cases} \alpha\phi_0(1-R) & \text{for } \alpha d \ll 1, \quad (9.76) \\ \alpha\phi_0(1-R)\,\mathrm{e}^{-\alpha y} & \text{for } \alpha d \gg 1, \quad (9.77) \end{cases}$$

where R is the reflection coefficient of the semiconductor defined by (9.25), α is the absorption coefficient, ϕ_0 is the photon flux density (i.e., $\phi_0 = I_0/h\nu$), and I_0 is the incident light intensity per unit area (W/cm^2).

Equation (9.76) is valid for a very thin photoconductor (i.e., $\alpha d \ll 1$) in which photons are uniformly absorbed throughout the sample, while (9.77) is applicable for a thick specimen (i.e., $\alpha d \gg 1$) in which the photogeneration rate decays exponentially with penetration distance. These are discussed next.

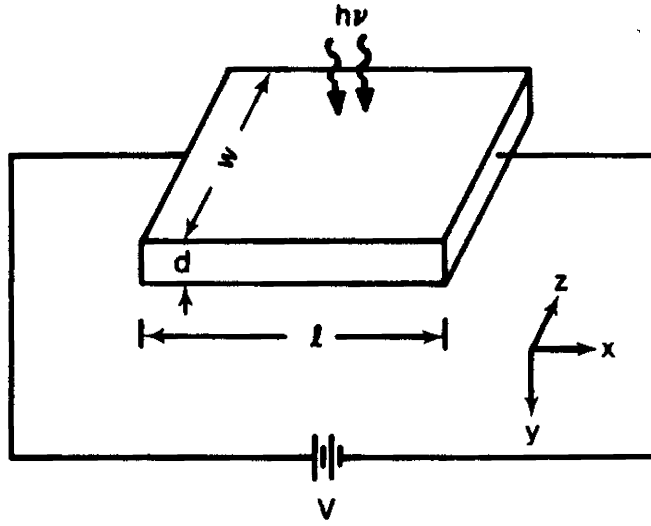


FIGURE 9.12. Photoconductivity process in a semiconductor specimen.

First consider the case of a thin specimen with $\alpha d \ll 1$. Here, the excess electron and hole densities are related to the generation rate g_E by

$$\Delta n = g_E \tau_n, \tag{9.78}$$

$$\Delta p = g_E \tau_p, \tag{9.79}$$

where τ_n and τ_p denote the electron and hole lifetimes, respectively. As shown in Figure 9.12, the change of electrical conductance as a result of the incident photons can be expressed by

$$\begin{aligned} \Delta G = \Delta\sigma \left(\frac{A}{l}\right) &= q\left(\Delta n \mu_n + \Delta p \mu_p\right)\left(\frac{Wd}{l}\right) \\ &= q g_E \left(\tau_n \mu_n + \tau_p \mu_p\right)\left(\frac{Wd}{l}\right) \end{aligned} \tag{9.80}$$

and

$$\Delta G = q G_E \frac{\left(\tau_n \mu_n + \tau_p \mu_p\right)}{l^2}, \tag{9.81}$$

where $G_E = g_E(Wdl) = g_E V_0$ is the total volume generation rate (i.e., total number of carriers generated per second), and $A = Wd$ is the cross-sectional area. If V is the applied voltage, the photocurrent I_{ph} can be expressed as

$$I_{ph} = V \Delta G = q V G_E \frac{\left(\tau_n \mu_n + \tau_p \mu_p\right)}{l^2} = q V G_E S, \tag{9.82}$$

where

$$S = \frac{(\tau_n \mu_n + \tau_p \mu_p)}{l^2} = \frac{\mu \tau}{l^2} \tag{9.83}$$

is the photosensitivity factor. It is seen that the value of S is directly proportional to the product $\mu\tau$. This means that in order to obtain a high photosensitivity factor, the lifetimes and mobilities of the excess carriers must be as large as possible and the sample length l between two electrodes should be as small as possible. As an example, consider a silicon photoconductor. If the wavelength of the incident photon is $\lambda = 0.5\ \mu m$, the absorption coefficient $\alpha = 10^4\ cm^{-1}$, $\tau_n = 100\ \mu s$, the reflection coefficient $R = 0.3$, and the photon flux density $\phi_0 = 10^{14}\ cm^{-2} \cdot s^{-1}$, then the excess electron density can be calculated using the following formula:

$$\Delta n = \alpha \phi_0 (1 - R) \tau_n = 7 \times 10^{13} cm^{-3}, \tag{9.84}$$

which shows that a relatively large density of excess electrons can be generated even with a relatively small incident light intensity. Another parameter that has often been used to assess the performance of a photoconductor is the photoconductivity gain G_p. This figure of merit (G_p) is defined as the ratio of the excess carrier lifetime τ to the carrier transit time t_r across the specimen, which can be written as

$$G_p = \frac{\tau}{t_r} = SV, \tag{9.85}$$

where $t_r = l/v_d = l^2/\mu V$ is the transit time for the excess carriers to drift across the photoconductor specimen, v_d is the drift velocity, and μ is the carrier mobility. A photoconductivity gain of 10^4 can be readily obtained for a CdS photoconductor.

In the above formulation, loss due to surface recombination was neglected. For a thin-film photoconductor, the effect of surface recombination can be incorporated into an effective excess carrier lifetime as

$$\frac{1}{\tau'} = \frac{1}{\tau_B} + \frac{1}{\tau_s}, \tag{9.86}$$

where τ' is the effective excess carrier lifetime, τ_B is the bulk carrier lifetime, and τ_s is the surface recombination lifetime given by

$$\tau_s = \frac{d}{2s}, \tag{9.87}$$

where s is the surface recombination velocity and d is the sample thickness. For example, for a chemically polished silicon specimen, with $s = 500$ cm/s and thickness $d = 2\ \mu m$, the surface recombination lifetime τ_s is found to be equal to 2×10^{-7} s. Therefore, for a thin-film photoconductor, if τ_s is less than τ_B, then the surface recombination lifetime rather than the bulk lifetime may control the effective excess carrier lifetime.

In general, the photocurrent for a thin-film photoconductor can be derived from (9.76) and (9.82), yielding

$$I_{\text{ph}} = \Delta G V_{\text{a}} = q\,(1-R)\,\alpha\phi_0\tau'\,(1+b)\,\mu_{\text{p}}\left(\frac{V}{l}\right)(Wd)\,, \tag{9.88}$$

where τ' is given by (9.86) and $b = \mu_{\text{n}}/\mu_{\text{p}}$ is the electron-to-hole mobility ratio. From (9.88) it is seen that for constant τ', the photocurrent I_{ph} is directly proportional to the light intensity $I_0\,(=\phi_0/h\nu)$ or photon flux density ϕ_0. This is generally true under low- and high-injection conditions (i.e., for $\Delta n \ll n_0$ or $\Delta n \ll n_0$). However, for the intermediate-injection range (i.e., $\Delta n \leq n_0$), τ may become a function of the injected excess carrier density Δn, and hence the photocurrent is no longer a linear function of the light intensity. Depending on the relationship between the excess carrier lifetime and the injected excess carrier density, a superlinear or sublinear region may exist in the intermediate-injection regime.

Next consider the case of a thick photoconductor with $\alpha d \gg 1$. Here, the generation rate is given by (9.77). Because of nonuniform absorption the diffusion of excess carriers along the direction of incident photons plays an important role in this case. As shown in Figure 9.12, the excess carrier densities as a function of distance along the y-direction can be obtained by solving the continuity equation for excess electron density:

$$D_{\text{n}}\frac{\partial^2\Delta n}{\partial y^2} - \frac{\Delta n}{\tau_{\text{n}}} = -g_{\text{E}} = -\alpha\phi_0\,(1-R)\,\mathrm{e}^{-\alpha y}. \tag{9.89}$$

If the electron diffusion length $L_{\text{n}}(=(D_{\text{n}}\tau_{\text{n}})^{1/2})$ is much smaller than the sample thickness d, then (9.89) has the solution

$$\Delta n = \frac{\alpha I_0\,(1-R)\,\tau_{\text{n}}}{h\nu\left(\alpha^2 L_{\text{n}}^2 - 1\right)}\left[\left(\frac{\alpha L_{\text{n}}^2 + s\tau_{\text{n}}}{L_{\text{n}} + s\tau_{\text{n}}}\right)\mathrm{e}^{-y/L_{\text{n}}} - \mathrm{e}^{-\alpha y}\right], \tag{9.90}$$

where $I_0 = \phi_0/h\nu$ is the incident light intensity. Note that (9.90) was obtained using the boundary condition

$$D_{\text{n}}\frac{\partial\Delta n}{\partial y}\bigg|_{y=0} = s\Delta n\bigg|_{y=0}. \tag{9.91}$$

The photocurrent can be obtained by integrating (9.90) with respect to y from $y = 0$ to $y = \infty$, yielding

$$\begin{aligned} I_{\text{ph}} &= \left(\frac{W}{l}\right)Vq\mu_{\text{p}}\,(1+b)\int_0^\infty \Delta n\,\mathrm{d}y \\ &= \frac{qI_0WL_{\text{n}}\mu_{\text{p}}(1+b)\tau_{\text{n}}(1-R)V}{l(L_{\text{n}}+s\tau_{\text{n}})h\nu}\left[1 + \frac{s\tau_{\text{n}}}{L_{\text{n}}(1+\alpha L_{\text{n}})}\right], \end{aligned} \tag{9.92}$$

where l is the sample length. In (9.92), the upper limit of the integral $y = d$ is replaced by $y = \infty$ in the integration. This is valid as long as the sample thickness is much larger than the diffusion length of electrons.

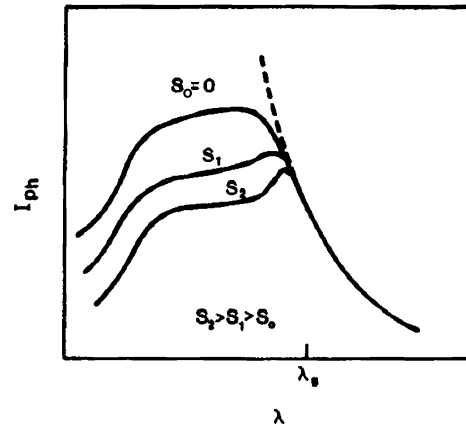


FIGURE 9.13. Relative photoresponse versus wavelength for different surface recombination velocities s_i.

Figure 9.13 shows the photocurrent versus wavelength of incident photons for different surface recombination velocities. For $s\tau_n \gg L_n$, the photocurrent I_{ph} reaches a maximum for $\alpha \approx 1/L_n$. However, if the surface recombination velocity is small and $s\tau_n \ll L_n$, then the photocurrent will increase monotonically with decreasing wavelength. This is shown in Figure 9.13 for the case $s = 0$. The sharp decrease in photocurrent is usually observed near the absorption edge (i.e., $h\nu \approx E_g$), in which the absorption coefficient decreases sharply with increasing wavelength. However, in the very short wave length regime (i.e., near the UV regime), the absorption coefficient is usually very large (i.e., $\alpha \geq 10^5$ cm^{-1}) and $\alpha L_n \gg 1$. In this regime, the excess carriers are generated near the surface of the photoconductor, where the excess carrier lifetime is controlled by the surface recombination. Thus, the photocurrent is expected to decrease rapidly with increasing surface recombination velocity in the short-wavelength regime. In order to improve the short-wavelength photoresponse, careful preparation of the sample surface is necessary so that the surface recombination velocity of the photoconductor can be kept low.

9.5.1. Kinetics of Photoconduction

Since the photocurrent is directly related to the excess carrier densities generated by the incident photons, a study of photocurrent as a function of light intensity usually yields useful information concerning the recombination mechanisms of the excess carriers in a semiconductor. As an example, consider an n-type direct band gap semiconductor. If the band-to-band radiative recombination dominates the excess carrier lifetimes, then the kinetic equation for the photoconduction process can be expressed by

$$\frac{dn}{dt} = g_E - U, \tag{9.93}$$

where g_E is the external generation rate of the excess carriers defined by (9.76) and U is the net recombination rate. For the band-to-band radiative recombination, U is given by

$$U = B(np - n_0 p_0) = B\left(np - n_i^2\right). \tag{9.94}$$

In the steady-state case, the carrier generation rate is obtained by solving (9.93) and (9.94), with the result

$$g_E = U = B\left(np - n_i^2\right), \tag{9.95}$$

where $n = n_0 + \Delta n$ and $p = p_0 + \Delta p$ denote the nonequilibrium electron and hole densities, respectively.

In order to understand the kinetics of photoconduction, consider two limiting cases, namely, the low- and high-injection cases.

(i) The low-injection case ($\Delta n \ll n_0,\ \Delta p \ll p_0$). Under low-injection conditions, (9.95) becomes

$$B\Delta n\,(p_0 + n_0) = g_E = \frac{\alpha I_0 (1 - R)}{h\nu}, \tag{9.96}$$

or

$$\Delta n = \frac{g_E}{B\,(n_0 + p_0)} = \frac{\alpha I_0 (1 - R)}{Bh\nu\,(n_0 + p_0)}. \tag{9.97}$$

In (9.97) it is assumed that $\Delta n = \Delta p$ (i.e., no trapping), and the charge-neutrality condition prevails. Equation (9.97) shows that Δn is directly proportional to the light intensity I_0. In the low-injection case, I_{ph} varies linearly with Δn, and hence I_{ph} is also a linear function of the light intensity I_0.

(ii) The high-injection case ($\Delta n = \Delta p \gg n_0,\ p_0$). Under high-injection conditions, (9.95) reduces to

$$g_E = B\Delta n^2, \tag{9.98}$$

or

$$\Delta n = \left(\frac{g_E}{B}\right)^{1/2} = \left[\frac{\alpha I_0 (1 - R)}{Bh\nu}\right]^{1/2}, \tag{9.99}$$

which shows that Δn is directly proportional to the square root of the light intensity. Thus, under high-injection conditions, the photocurrent varies with the square root of light intensity when the band-to-band radiative recombination is dominant.

Another example to be given here for analyzing the kinetics of photoconduction in a semiconductor is shown in Figure 9.14 for a p-type extrinsic photoconductor with a deep acceptor center. The kinetic equation for the photoconduction process in this case is given by

$$\frac{dp}{dt} = e_p\,(N_A - p) - c_p p^2 + g_E, \tag{9.100}$$

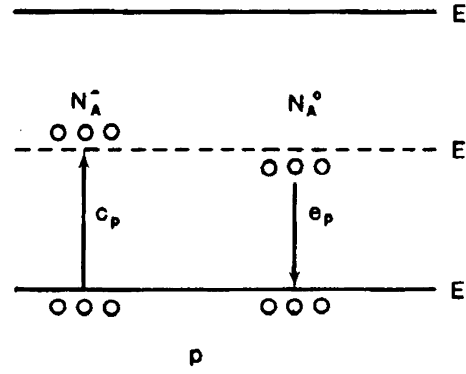


FIGURE 9.14. Kinetics of photoconduction in a p-type semiconductor with a deep acceptor center.

where e_p and c_p denote the emission and capture rates of holes, respectively, as shown in the figure. The first term on the right-hand side of (9.100) is the rate of spontaneous generation from the neutral acceptor centers. The second term gives the rate of recombination of free holes and ionized acceptor centers. For the steady-state low-injection case, solving (9.100) yields

$$\Delta p = \frac{g_E}{(e_p + 2c_p p_0)} = g_E \tau_p, \tag{9.101}$$

where $\tau_p = 1/(e_p + 2c_p p_0)$. Equation (9.101) is obtained using the fact that $e_p(N_A - p_0) = c_p {p_0}^2$ and $\Delta p \ll p_0$ in (9.100). The results predict a linear relationship between Δp (or I_{ph}) and g_E (or I_0), providing that the hole lifetime is constant and independent of injection.

For the high-injection case, under steady-state conditions, (9.100) becomes

$$g_E = c_p \Delta p^2, \tag{9.102}$$

or

$$\Delta p = \left(\frac{g_E}{c_p}\right)^{1/2}, \tag{9.103}$$

which shows that Δp is directly proportional to the square root of the generation rate.

A typical plot of the relative photocurrent versus light intensity is illustrated in Figure 9.15, which usually consists of a linear, a sublinear, and a superlinear region as the light intensity increases from low to high. This type of behavior has been observed in a wide variety of photoconductors such as CdS and other II-VI compound semiconductors.

9.5.2. *Practical Applications of Photoconductivity*

Photoconductors made from high-resistivity semiconductors are often used to detect visible-to-IR radiation. Intrinsic photoconductors such as lead sulfide (PbS),

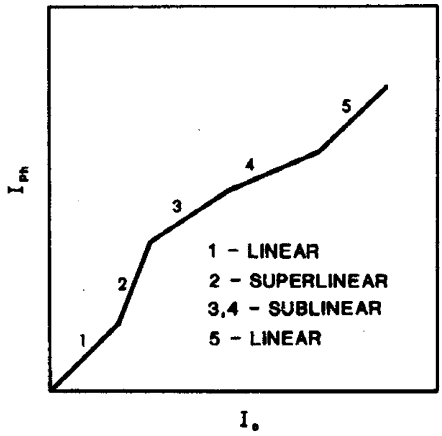


FIGURE 9.15. Photocurrent versus light intensity for a photoconductor, showing the linear, superlinear, and sublinear regions.

lead selenide (PbSe), lead telluride (PbTe), and indium antimonide (InSb), operating at 77 K, are commonly used for 3- to 5-μm middle-wavelength infrared (MWIR) detection, while wide-band-gap photoconductors such as CdS and CdTe are mainly used in the near-UV to visible spectral region. For longer-wavelength (i.e., $\lambda > 10\,\mu m$) applications, extrinsic photoconductors such as Au-, Cu-, Hg-, and Cd-doped germanium photoconductors and CdHgTe photoconductors are widely used. Figure 9.16 shows a detectivity versus wavelength plot for various PC and PV infrared detectors.[8] The detection principle for these IR detectors is based on the optical excitation of holes from the acceptor-impurity centers into the valence band. In order to suppress the competing thermal excitation, extrinsic photoconductors are generally operated in the temperature range between 77 and 4.2 K. Since optical absorption coefficients for extrinsic photoconduction are usually very small (typically in the range of 1 to 10 cm^{-1}), to obtain high quantum efficiency the thickness of the photoconductor along the direction of incident photons is usually several millimeters or centimeters.

In an intrinsic photoconductor, the conduction process is due to band-to-band excitation. Consequently, the absorption coefficients are usually very large (in the range of 10^3 to 10^5 cm^{-1}), and hence the thickness of the photoconductor in the direction of incident light may vary from a couple of micrometers to a few tens of micrometers. Practical long-wavelength infrared (LWIR) photodetectors operating in the 8- to 14-μm spectral range have been developed from II-IV-VI compound semiconductors such as $Hg_{1-x}Cd_xTe$ and $Pb_{1-x}Sn_xTe$ material systems. $Hg_{1-x}Cd_xTe$ IR detectors with very high detectivity [i.e., $D^* \geq 10^{11}$ cm · $Hz^{1/2}$/W] operating at 77 K have been developed for $10.6 = \mu m$ detection. The energy band gap of $Hg_{1-x}Cd_xTe$ can be varied by changing the mole fraction ratio x (i.e., E_g may vary from 1.4 eV to less than 0.1 eV as x varies from 1 to 0). As a result, the $Hg_{1-x}Cd_xTe$ IR detectors can be tailored to the desired wavelength by varying the mole fraction x in this material. The main drawback of $Hg_{1-x}Cd_xTe$ for LWIR detection applications is that long-term stability and composition uniformity across the wafer need to be improved.

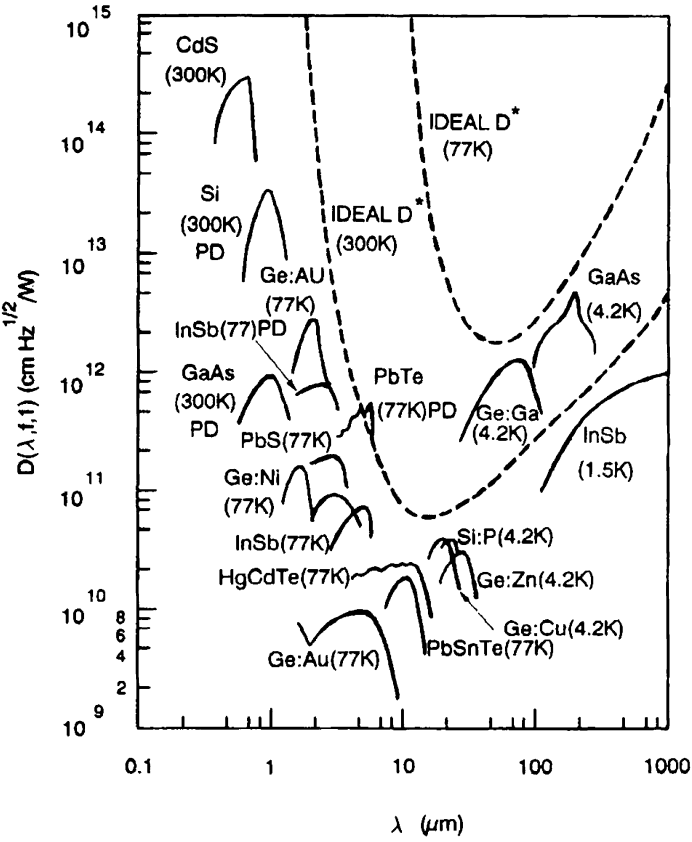


FIGURE 9.16. Spectral dependence of detectivity for some important photoconductive and photovoltaic infrared detectors. After Sze,[8] by permission.

In addition to practical applications of photoconductors described above, the fundamental physical parameters related to the recombination mechanisms of excess carriers can be determined by studying the steady-state and transient photoconductivity effects in a semiconductor. For example, from the study of photoconductance versus light intensity, one can deduce basic information concerning the recombination and trapping mechanisms in a semiconductor. The minority carrier lifetime or diffusion length in a semiconductor can be determined by using either the steady-state or transient photoconductivity method. This will be discussed later in Section 9.7.

9.6. The Photovoltaic (Dember) Effect

The internal photovoltaic (PV) or the Dember effect in a semiconductor is discussed in this section. Figure 9.12 shows the incident photons with energies greater than

the band gap energy (i.e., $h\nu \geq E_g$) impinging on a p-type semiconductor specimen. If the sample thickness is much larger than the inverse absorption coefficient, then a concentration gradient of the photoinjected excess carriers is established in the direction of incident photons. This will cause electron and hole currents to flow by diffusion along the direction of the incident light. The total diffusion current is equal to zero if the mobilities of electrons and holes are equal in the semiconductor. In general, the electron and hole mobilities are not equal in semiconductors. As a result, an unbalanced electron and hole diffusion current will create an internal electric field along the direction of incident light. The polarity of this internal electric field tends to assist hole diffusion and retard electron diffusion. As a result, an internal electric field is established under this condition. This internal electric field is usually referred to as the Dember field. To derive an expression for the Dember field, the components of electron and hole current densities (i.e., J_{ny} and J_{py}) along the direction of incident photons can be written as

$$J_{ny} = qn\mu_n \mathcal{E}_y + qD_n \frac{\partial n}{\partial y}, \tag{9.104}$$

$$J_{py} = qp\mu_p \mathcal{E}_y - qD_p \frac{\partial p}{\partial y}, \tag{9.105}$$

where $n = n_0 + \Delta n$, $p = p_0 + \Delta p$, $\Delta n = \Delta p$; μ_n/D_n and μ_p/D_p are related by the Einstein relations, which are given by

$$D_n = \left(\frac{k_B T}{q}\right)\mu_n \quad \text{and} \quad D_p = \left(\frac{k_B T}{q}\right)\mu_p. \tag{9.106}$$

The total current density in the y-direction is equal to the sum of J_{ny} and J_{py}. From (9.104) and (9.105), one obtains an expression for J_y as

$$J_y = J_{ny} + J_{py} = q(bn + p)\mu_p \mathcal{E}_y + (b-1)qD_p \frac{\partial \Delta n}{\partial y}. \tag{9.107}$$

Under open-circuit conditions, $J_y = 0$, and the Dember field in the y-direction is given by

$$\mathcal{E}_y = -\left(\frac{k_B T}{q}\right)\frac{(b-1)}{(bn+p)}\frac{\partial \Delta n}{\partial y}, \tag{9.108}$$

where $b = \mu_n/\mu_p$ is the electron-to-hole mobility ratio; the expression for $\mathcal{E}_y$ given by (9.108) is known as the Dember electric field. It is noted that the Dember field vanishes if $b = 1$. The Dember field developed inside the sample will enhance the diffusion of holes and retard the diffusion of electrons. The resulting photovoltage (or the Dember voltage) between the front side (i.e., illuminated side) and the back side of the sample is obtained by integrating (9.108) from $y = 0$ to $y = d$. For the

small-injection case (i.e., $\Delta n \ll n_0$), the Dember voltage is given by

$$\begin{aligned} V_{\mathrm{d}} &= \int_0^d -\mathcal{E}_y \, \mathrm{d}y \\ &= \left(\frac{k_{\mathrm{B}}T}{q}\right) \int_{\Delta n_0}^{\Delta n_{\mathrm{d}}} \frac{-(b-1)}{(bn_0 + p_0)} \, \mathrm{d}\Delta n \\ &= \left(\frac{k_{\mathrm{B}}T}{q}\right) \frac{(b-1)}{(bn_0 + p_0)} (\Delta n_0 - \Delta n_{\mathrm{d}}), \end{aligned} \tag{9.109}$$

where Δn_0 is the excess electron density at $y = 0$ and Δn_{d} is the excess electron density at $y = d$. For a thick sample, $\alpha d \gg 1$ and $\Delta n_{\mathrm{d}} \ll \Delta n_0$, (9.109) becomes

$$V_{\mathrm{d}} = \left(\frac{k_{\mathrm{B}}T}{q}\right) \frac{(b-1)}{(bn_0 + p_0)} \Delta n_0. \tag{9.110}$$

The excess electron density Δn_0 at $y = 0$ can be related to the incident light intensity using the result obtained in (9.90), which is given by

$$\Delta n_0 = \frac{\alpha L_{\mathrm{n}} I_0 (1 - R)\tau_{\mathrm{n}}}{h\nu(\alpha L_{\mathrm{n}} + 1)(L_{\mathrm{n}} + s\tau_{\mathrm{n}})}. \tag{9.111}$$

Now substituting (9.111) into (9.110) yields

$$V_{\mathrm{d}} = \left(\frac{k_{\mathrm{B}}T}{q}\right) \frac{\alpha L_{\mathrm{n}} I_0 (1 - R)\tau_{\mathrm{n}}(b-1)}{h\nu(\alpha L_{\mathrm{n}} + 1)(L_{\mathrm{n}} + s\tau_{\mathrm{n}})(bn_0 + p_0)}. \tag{9.112}$$

Equation (9.112) shows that under low injection the Dember voltage varies linearly with the light intensity I_0. In contrast to the PC effect, the PV effect requires no external applied voltage, and hence can be used to generate electrical power using the PV effect in a semiconductor. Devices using this internal PV effect are known as solar cells or PV devices. Typical PV devices are fabricated using a p-n junction or a Schottky barrier structure. Details of the PV devices and their operation principles will be discussed in detail in Chapter 12.

9.7. The Photomagnetoelectric Effect

The photomagnetoelectric (PME) effect refers to the voltage (or current) developed in a semiconductor specimen as a result of the interaction of an applied magnetic field with the diffusion current produced by the photogenerated excess carriers. A PME open-circuit voltage (or short-circuit current) is developed in the x-direction when a magnetic field is applied in the z-direction and the incident light is in the y-direction of the specimen, as illustrated in Figure 9.17.

The development of a PME field in a semiconductor under the influence of a magnetic field and incident light may be explained as follows: Electron–hole pairs generated near the surface of a semiconductor specimen by the incident photons are diffused into the specimen in the direction of incident light. The time-invariant

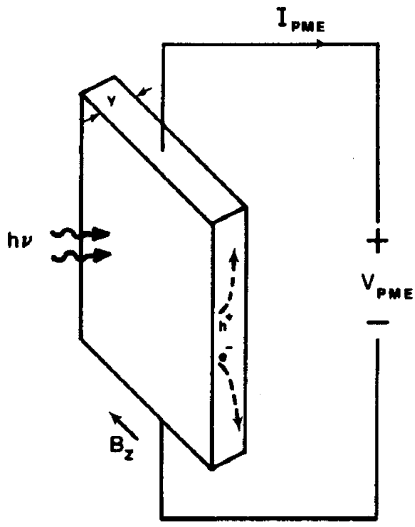


FIGURE 9.17. Schematic diagram showing the photomagnetoelectric (PME) effect in a semiconductor.

magnetic flux density B along the z-direction deflecting holes in the positive x-direction and electrons in the negative x-direction results in a net PME short-circuit current flowing in the positive x-direction. Under open-circuit conditions, a PME voltage is developed in the x-direction.

In order to derive the expression for the PME open-circuit voltage or PME short-circuit current, consider a semi-infinite semiconductor slab. The following equations hold for the rectangular slab shown in Figure 9.17. For a small magnetic field (i.e., $\mu B \ll 1$) and the small-injection case, the Hall angles are given by

$$\tan \theta_{\mathrm{p}} \approx \theta_{\mathrm{p}} = \mu_{\mathrm{p}} B \tag{9.113}$$

$$\tan \theta_{\mathrm{n}} \approx \theta_{\mathrm{n}} = -\mu_{\mathrm{n}} B \tag{9.114}$$

$$\theta = \theta_{\mathrm{p}} - \theta_{\mathrm{n}} = \mu_{\mathrm{p}}(1 + b)B, \tag{9.115}$$

where $\Delta n = \Delta p \ll n_0$ or p_0, and $b = \mu_{\mathrm{n}}/\mu_{\mathrm{p}}$ is the electron and hole mobility ratio.

The hole and electron current density can be best described by vector equations with $\hat{\imath}, \hat{\jmath}, \hat{k}$ denoting the unit vectors along the x-, y-, and z-axes, respectively

$$\mathbf{J}_{\mathrm{p}} \cong J_{\mathrm{py}}\hat{\mathbf{j}} + \theta_{\mathrm{p}} J_{\mathrm{py}}\hat{\mathbf{j}} \times \hat{\mathbf{k}}, \tag{9.116}$$

$$\mathbf{J}_{\mathrm{n}} \cong J_{\mathrm{ny}}\hat{\mathbf{j}} + \theta_{\mathrm{n}} J_{\mathrm{ny}}\hat{\mathbf{j}} \times \hat{\mathbf{k}}. \tag{9.117}$$

Here J_{py} and J_{ny} are given respectively by

$$J_{\mathrm{py}} = q\left(\mu_{\mathrm{p}} p \mathcal{E}_y - D_{\mathrm{p}} \frac{\partial \Delta p}{\partial y}\right), \tag{9.118}$$

$$J_{\mathrm{ny}} = q\left(\mu_{\mathrm{n}} n \mathcal{E}_y + D_{\mathrm{n}} \frac{\partial \Delta n}{\partial y}\right). \tag{9.119}$$

Thus, the total current density is given by

$$\mathbf{J} = \mathbf{J}_{\mathrm{p}} + \mathbf{J}_{\mathrm{n}} = (J_{\mathrm{px}} + J_{\mathrm{nx}})\hat{\mathbf{i}} + (J_{\mathrm{py}} + J_{\mathrm{ny}})\hat{\mathbf{j}}. \tag{9.120}$$

Since the total current density in the y-direction is zero, equation (9.120) reduces to

$$J = J_x = (J_{\mathrm{px}} + J_{\mathrm{nx}}) = [q(p\mu_{\mathrm{p}} + n\mu_{\mathrm{n}})\mathcal{E}_x + \theta_{\mathrm{p}} J_{\mathrm{py}} + \theta_{\mathrm{n}} J_{\mathrm{ny}}]. \tag{9.121}$$

Solving (9.118) through (9.120) yields

$$J_{\mathrm{py}} = -J_{\mathrm{ny}} = -qD\left(\frac{\partial \Delta n}{\partial y}\right), \tag{9.122}$$

where $D = D_{\mathrm{n}}(n + p)/(bn + p)$ is the effective diffusion coefficient. Substituting (9.122) into (9.121), one obtains

$$J_x = J_{\mathrm{nx}} + J_{\mathrm{px}} = q\mu_{\mathrm{p}}(p + nb)\mathcal{E}_x - q(b + 1)\mu_{\mathrm{p}} BD\left(\frac{\partial \Delta n}{\partial y}\right). \tag{9.123}$$

To derive the PME electric field or PME open-circuit voltage along the x-direction, it is usually not sufficient to assume that $J_x = J_{\mathrm{nx}} + J_{\mathrm{px}} = 0$. Such a solution would lead to J_x being a function of z, which is incorrect, because for a constant magnetic field the electric field must be irrotational. With $\partial \mathcal{E}_z/\partial x = 0$, it follows that $\partial \mathcal{E}_x/\partial z = 0$. Thus, the correct boundary condition is given by

$$\int_0^d (J_{\mathrm{nx}} + J_{\mathrm{px}})\, \mathrm{d}y = 0. \tag{9.124}$$

For the small-injection case, with $\Delta n = \Delta p \ll n_0$ or p_0, the PME electric field in the x-direction becomes

$$\mathcal{E}_x = \frac{(b + 1)BD}{d(n_0 b + p_0)} \int_0^d \frac{\partial \Delta n}{\partial y}\, \mathrm{d}y = -\frac{BD(b + 1)}{d(n_0 b + p_0)}(\Delta n_0 - \Delta n_{\mathrm{d}}) \tag{9.125}$$

The PME short-circuit current can be obtained by setting $\mathcal{E}_x = 0$ in (9.123) and then integrating the equation from $y = 0$ to $y = d$. The result is

$$\begin{aligned} I_{\mathrm{PME}} &= -W \int_0^d q(b + 1)\mu_{\mathrm{p}} BD\left(\frac{\partial \Delta n}{\partial y}\right) \mathrm{d}y \\ &= qW(b + 1)\mu_{\mathrm{p}} BD(\Delta n_0 - \Delta n_{\mathrm{d}}). \end{aligned} \tag{9.126}$$

For $\alpha d \gg 1$ and $d \gg L_{\mathrm{n}}$, one may assume that $\Delta n_{\mathrm{d}} \approx 0$. Substituting (9.111) into (9.126), one obtains

$$I_{\mathrm{PME}} = \frac{qW(b + 1)\mu_{\mathrm{p}} I_0(1 - R)BL_{\mathrm{n}}^2}{h\nu(L_{\mathrm{n}} + s\tau_{\mathrm{n}})} \cdot \frac{\alpha L_{\mathrm{n}}}{(1 + \alpha L_{\mathrm{n}})}. \tag{9.127}$$

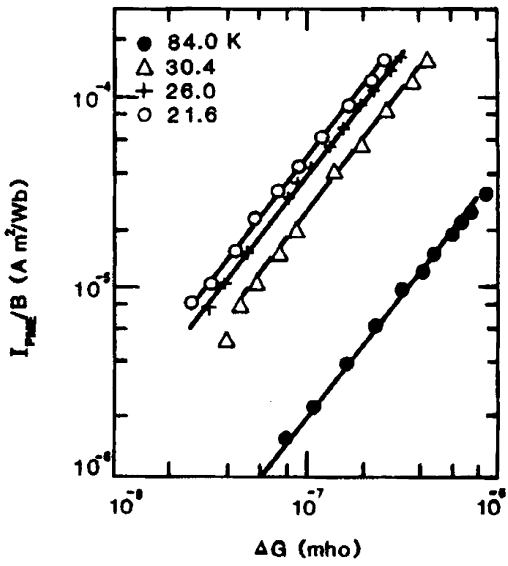


FIGURE 9.18. PME short-circuit current versus photoconductance for an Au-doped silicon specimen with $N_{Au} = 5 \times 10^{16}$ cm^{-3}. After Agraz and Li,[9] by permission.

In (9.127), the quantum yield is assumed equal to 1. Thus, at low magnetic fields, the PME short-circuit current is directly proportional to the magnetic flux density B. For $\alpha L_n \gg 1$, I_{PME} becomes independent of the wavelength.

The ratio of the PME short-circuit current given by (9.127) to the photocurrent given by (9.92) is given by

$$\frac{(I_{PME}/B)}{(I_{ph}/\mathcal{E}_x)} = \frac{\alpha L_n^2}{(1 + \alpha L_n + \frac{s\tau_n}{L_n})\tau_n}. \tag{9.128}$$

If αL_n is much larger than 1 and $s\tau_n/L_n$, then (9.128) reduces to

$$\frac{(I_{PME}/B)}{(I_{ph}/\mathcal{E}_x)} = \frac{L_n}{\tau_n} = \left(\frac{D_n}{\tau_n}\right)^{1/2}, \tag{9.129}$$

where $L_n = (D_n\tau_n)^{1/2}$ is the electron diffusion length. Equation (9.129) provides a direct means for determining the minority carrier lifetime from the PME and PC effect measurements. The electron diffusion constant can be determined from the electron mobility data using the Einstein relation given by (9.106).

The PME effect has been used in determining the minority carrier lifetimes in a semiconductor. This method is particularly attractive for semiconductors with very short carrier lifetimes for which the transient photoconductivity decay method fails. Studies of the PME effect have been reported in various semiconductors such as Ge, Si, InSb, and GaAs. Figure 9.18 shows the PME short-circuit current versus photoconductance for an Au-doped silicon sample measured at different temperatures. The nonlinear relationship between I_{PME} and ΔG is due to the effect of minority carrier trapping in the Au-doped silicon sample.

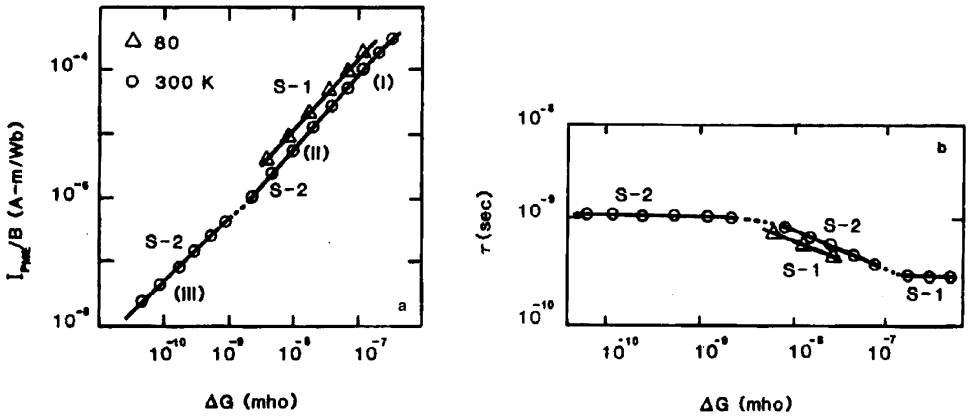


FIGURE 9.19. (a) PME short-circuit current and (b) lifetime versus photoconductance for a Cr-doped n-type GaAs sample. After Huang and Li,[10] by permission.

Figure 9.19a shows the PME short-circuit current versus photoconductance over a wide range of light intensity for two Cr-doped n-type GaAs samples. The dependence of lifetime on photoconductance is also illustrated in Figure 9.19b. In the low- and high-injection regimes, I_{PME} varies linearly with ΔG, while in the intermediate-injection range a nonlinear relationship exists between I_{PME} and ΔG as a result of the trapping effect in the sample.

Problems

9.1. Consider an n-type silicon specimen. The sample is 0.2 cm thick, 1 cm wide, and 2 cm long. Monochromatic light with wavelength $\lambda = 0.9\ \mu\text{m}$ and intensity $I_0 = 5 \times 10^{-4}\ \text{W/cm}^2$ is impinging on the sample. Assuming that the equilibrium electron density is $n_0 = 10^{16}\ \text{cm}^{-3}$ and the absorption coefficient α is equal to $320\ \text{cm}^{-1}$ at $\lambda = 0.9\ \mu\text{m}$, find:
   (a) The number of incident photons per second on this sample.
   (b) The depth at which the light intensity I_0 is 10% of its value at the surface.
   (c) The number of electron–hole pairs generated per second in this sample, assuming a quantum yield of $\eta = 0.8$.
   (d) The photoconductance ΔG, assuming an electron diffusion length $L_{\text{n}} = 50\ \mu\text{m}$ and that the surface recombination at the illuminated surface is zero ($s = 0$).

9.2. Repeat Problem 9.1, (a) through (d), assuming that the wavelength of the incident photons is $\lambda = 0.63\ \mu\text{m}$ and the absorption coefficient of silicon is $\alpha = 3 \times 10^3\ \text{cm}^{-1}$ at $\lambda = 0.63\ \mu\text{m}$.

9.3. Show that under high-injection conditions (i.e., $\Delta p = \Delta n \gg n_0$ or p_0) and assuming $\tau_{\text{n}} = \tau_{\text{p}} = \tau_{\text{h}}$ (where τ_{h} is the carrier lifetime at high injection as defined by the Shockley–Read–Hall model), the following relations prevail:

(a) $\Delta G = qW\mu_p(1+b)(D\tau_n)^{1/2}\Delta n_0 l$,
(b) $I_{PME} = qW\mu_p(1+b)DB\Delta n_0$,
(c) $V_{PME} = I_{PME}/\Delta G = (D/\tau_n)^{1/2}Bl$,
where $D = 2D_n/(1+b)$ is the ambipolar diffusion constant and Δn_0 is the excess electron density generated at the illuminated surface.

9.4. An n-type CdS photoconductor, which has a dark resistivity of $10^8\ \Omega$ cm, is illuminated by a He–Ne laser ($\lambda = 0.6328\ \mu$m). If the power output of the laser beam is 0.5 mW/cm^2, find:
(a) The incident photon flux density per second.
(b) The volume generation rate given $\alpha = 4 \times 10^3$ cm^{-1} at $\lambda = 0.6328\ \mu$m and $R = 0$.
(c) The photogenerated electron density given $\tau_n = 10^{-3}$ s.
(d) The photoconductivity. Given $\mu_n = 400$ cm^2/V $\cdot$ s.

9.5. Using (9.92), calculate the photocurrent for an InSb photoconductor for $s = 0$, 100, and 10,000 cm/s, respectively, given $\tau_n = 10\ \mu$s, $L_n = 10\ \mu$m, $R = 0.35$, $w/l = 1$, $I_0 = 1$ mW/cm^2, $V = 100$ mV, $\mu_p = 500$ cm^2/V $\cdot$ s, $b = 100$, $\lambda = 4\ \mu$m, $hv = hc/\lambda$.

9.6. Using (9.112), calculate the Dember voltages for the InSb photoconductor given in Problem 9.5 with $\alpha = 10^4$ cm^{-1} and $n_0 = 5 \times 10^{16}$ cm^{-3} assume that $p_0 \approx 0$, quantum yield $\eta = 0.8$, surface recombination velocities $s = 0$, 100, and 10^4 cm/s, and $T = 300$ K.

9.7. If the intensity of incident photons on an n-type semiconductor is such that $\Delta p \gg n_0$ (i.e., high-injection case), and the band-to-band radiative recombination is dominant, show that the photocurrent is directly proportional to the square root of the light intensity.

9.8. Suppose the energy band gap versus alloy composition x for an $Al_xGa_{1-x}As$ ternary compound semiconductor is given by

$$E_g = \begin{cases} 1.424 + 1.247x & \text{for } 0 \le x \le 0.45, \\ 1.900 + 0.125x + 0.143x^2 & \text{for } 0.45 \le x \le 1.0. \end{cases}$$

(a) Plot E_g versus x for $0 \le x \le 1.0$. Note that Al_xGa_{1-x} As is a direct band gap material for $x \le 0.45$, and becomes an indirect band gap material for $x > 0.45$.
(b) Plot the optical absorption coefficient (α) versus wavelength λ for $Al_xGa_{1-x}As$ materials with $x = 0.2$, 0.4, 0.6, and 0.8.
(c) If a He–Ne laser beam with $\lambda = 0.6328\ \mu$m and intensity $I_0 = 0.5$ mW/cm^2 is illuminated on an $Al_xGa_{1-x}As$ specimen 10 μm thick, 100 μm wide, and 500 μm long, what is the total volume generation rate for this sample? Assume that the optical absorption coefficient at this wavelength is $\alpha = 5 \times 10^4$ cm^{-1}.

9.9. Consider an intrinsic photoconductor. If the wavelength of incident photons is $\lambda = 0.5\ \mu$m, the optical absorption coefficient at this wavelength is $\alpha = 10^4$ cm^{-1}, the reflection coefficient is $R = 0.3$, the excess electron lifetime $\tau_n = 10\ \mu$s, and the photon flux density is 10^{14} cm^{-2}-s^{-1}, calculate the excess

electron density generated by the incident photons described in this problem. Assume uniform absorption in this thin film photoconductor. If the electron mobility μ_n is equal to 1,500 $cm^2/V \cdot s$, hole mobility $\mu_p = 500$ $cm^2/V \cdot s$, and assuming $\tau_n = \tau_p$, what is the photosensitivity factor for this intrinsic photoconductor?

References

1. T. S. Moss, *Optical Properties of Semiconductors*, Academic Press, New York (1959).
2. I. Kudman and T. Seidel, *J. Appl. Phys*. **33**, 771 (1962).
3. G. G. Macfarlane, T. P. Mclean, J. E. Quarrington, and V. Roberts, *J. Phys. Chem. Solids* **8**, 390 (1959).
4. W. G. Spitzer, M. Gershenzon, C. J. Frosch, and D. F. Gibbs, *J. Phys. Chem. Solids* **11**, 339 (1959).
5. T. S. Moss and T. D. Hawkins, *Infrared Phys*. **1**, 111 (1962).
6. W. C. Dash and R. Newman, *Phys. Rev*. **99**, 1151 (1955).
7. D. L. Greenaway and G. Harbeke, *Optical Properties and Band Structure of Semiconductors*, Pergamon Press, New York (1968).
8. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley-Interscience, New York (1981).
9. A. G. Agraz and S. S. Li, *Phys. Rev*. **2**, 1947 (1970).
10. C. I. Huang and S. S. Li, *J. Appl. Phys*. **44**, 4214 (1973).

Bibliography

R. H. Bube, *Photoconductivity of Solids*, Wiley, New York (1960).
R. H. Bube, *Electronic Properties of Crystalline Solids*, Academic Press, New York (1974).
O. Madelung, *Physics of III–V Compounds*, Wiley, New York (1964).
T. Moss, *Optical Properties of Semiconductors*, Butterworths, London (1959).
A. Rose, *Concepts of Photoconductivity and Allied Problems*, Wiley, New York (1963).
S. M. Ryvkin, *Photoelectric Effects in Semiconductors*, Consultants Bureau, New York (1968).
R. K. Willardson and A. C. Beer, *Semiconductors and Semimetals*, vol. 3, Academic Press, New York (1967).

10
Metal–Semiconductor Contacts

10.1. Introduction

In this chapter, the basic device physics, the electrical and transport properties, and the formation and characterization of various metal–semiconductor contacts are presented. It is well known that the quality of metal–semiconductor contacts plays an important role in the performance of various semiconductor devices and integrated circuits. For example, good ohmic contacts are essential for achieving excellent performance of a semiconductor device, while Schottky (i.e., rectifying) contacts can be used for a wide variety of device applications. In addition to different device and circuit applications, Schottky contacts can also be used as test vehicles for investigating the physical and electrical properties of a semiconductor material and its surfaces. For example, a Schottky diode can be used to study bulk defects and interface properties of a metal–semiconductor system. Therefore, it is essential to obtain a better understanding of the fundamental physical and electrical properties of the metal–semiconductor systems so that technologies for preparing good ohmic and Schottky contacts can be developed for a wide variety of device applications.

Two types of metal–semiconductor contacts are commonly used in the fabrication of semiconductor devices and integrated circuits. They are the Schottky and ohmic contacts. A Schottky barrier contact exhibits an asymmetrical current–voltage (I–V) characteristic when the polarity of a bias voltage applied to the metal–semiconductor contacts is changed. The ohmic contact, on the other hand, shows a linear I–V characteristic regardless of the polarity of the external bias voltage. A good ohmic contact is referred to the case in which the voltage drop across a metal–semiconductor contact is negligible compared to that of the bulk semiconductor material.

The Schottky barrier diode is actually a variation of the point-contact diode in which the metal–semiconductor junction is a surface rather than a point contact. In fact, a large contact area between the metal and the semiconductor in a Schottky barrier diode provides some advantages over the point-contact diode. Lower forward resistance and lower noise generation are the most important

advantages of the Schottky barrier diode. The applications of a Schottky barrier diode are similar to those of the point-contact diode. The low noise level generated by Schottky diodes makes them especially suitable for uses in microwave receivers, detectors, and mixers. The Schottky barrier diode is sometimes called the hot electron or hot carrier diode because the electrons flowing from the semiconductor to the metal have a higher energy level than electrons in the metal. The effect is the same as it would be if the metals were heated to a higher temperature than normal.

Section 10.2 describes the metal work function and the Schottky effect at a metal–vacuum interface. Thermionic emission theory, used to describe carrier transport in a metal–semiconductor contact, is presented in Section 10.3. In Section 10.4, the energy band diagram, the spatial distributions of the space charge, potential, and electric field across the depletion layer of a Schottky barrier diode are derived. Section 10.5 presents the diffusion and thermionic emission models for carrier transport in a Schottky barrier diode. Section 10.6 describes the *I–V* characteristics and fabrication schemes for a metal-Si and metal-GaAs Schottky barrier diode. Section 10.7 describes three common methods for determining the barrier height of a Schottky diode. Methods for the effective barrier height enhancement of a metal–semiconductor Schottky contact are discussed in Section 10.8. In Section 10.9, applications of Schottky barrier diodes for photodetectors, microwave mixers, clamped transistors, metal-gate field-effect transistors (MESFETs), and solar cells are discussed. Finally, some conventional and novel approaches for forming ohmic contacts on semiconductors are presented in Section 10.10.

10.2. Metal Work Function and Schottky Effect

The schematic energy band diagram under equilibrium conditions for a metal in free space is shown in Figure 10.1. The energy difference between the vacuum level and the Fermi level is known as the work function of a metal. The work function, ϕ_m, is defined as the minimum kinetic energy required for an electron to escape from the metal surface (or the Fermi level) into free space at $T = 0$ K. The probability for an electron to escape from the metal surface into the vacuum depends on the velocity of electrons perpendicular to the metal surface. The minimum kinetic energy required for an electron to escape from the metal surface into vacuum is

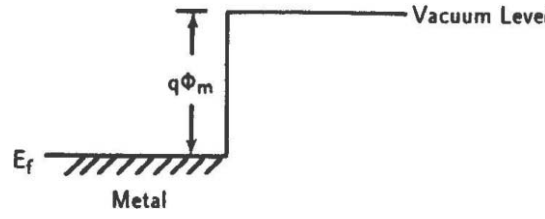


FIGURE 10.1. Energy band diagram at a metal–vacuum interface: ϕ_m is the metal work function and E_f is the Fermi level.

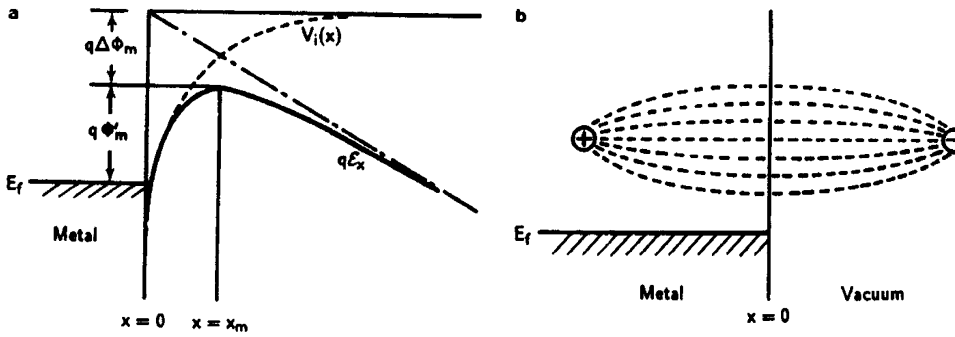


FIGURE 10.2. Schottky (or image-lowering) effect at the metal–vacuum interface in the presence of an applied electric field: (a) energy band diagram, showing the applied field (g_x), the image potential $V_i(x)$, and the image-lowering potential $\Delta\phi_m$; (b) the induced image charge (positive) inside the metal.

given by

$$\frac{1}{2}\left(m_0 v_1^2\right) \geq q\phi_m \,, \tag{10.1}$$

where v_1 is the electron velocity normal to the metal surface, and m_0 is the free electron mass. The Schottky effect, or image-lowering effect, occurs when an external electric field is applied to the metal surface. To understand the Schottky effect, consider the energy band diagram shown in Figure 10.2a. When an electric field is applied to the metal surface, electrons that escape from the metal surface will experience two external forces: the image force that arises from the Coulomb attractive force as a result of the positive image charges induced inside the metal by the escaping electrons, and the Lorentz force due to the applied electric field. The positive image charges create a Coulomb attractive force, which tends to pull the escaping electrons back into the metal. The image force can be expressed by

$$F_i = \frac{q^2}{16\pi\varepsilon_0 x^2} \,, \tag{10.2}$$

where x is the distance from the metal surface. The potential energy associated with this image force is given by

$$V_i(x) = -\int_{\infty}^{x} F_i \, dx = -\frac{q^2}{16\pi\varepsilon_0 x} \,. \tag{10.3}$$

The potential energy due to the applied electric field can be written as

$$V_a(x) = -q\mathcal{E}x. \tag{10.4}$$

The total potential energy of the electron is equal to the sum of (10.3) and (10.4),

namely,

$$V(x) = V_{\mathrm{i}}(x) + V_{\mathrm{a}}(x) = -\frac{q^2}{16\pi\varepsilon_0 x} - q\mathcal{E}x. \tag{10.5}$$

The distance at which the maximum potential energy occurs is obtained by differentiating (10.5) with respect to x and then setting the result equal to 0, which yields

$$x_{\mathrm{m}} = \sqrt{\frac{q}{16\pi\varepsilon_0\mathcal{E}}}. \tag{10.6}$$

Substituting (10.6) into (10.5), one obtains the maximum potential energy for the electron, which is

$$V_{\mathrm{m}}(x_{\mathrm{m}}) = -q\mathcal{E}\sqrt{\frac{q}{4\pi\varepsilon_0\mathcal{E}}} = -2q\mathcal{E}x_{\mathrm{m}} = -q\Delta\phi_{\mathrm{m}}. \tag{10.7}$$

As shown in Figure 10.2, the effect of the image force and the applied electric field is to lower the work function of a metal. Therefore, the effective metal work function under the applied electric field can be obtained from Figure 10.2, and the result is

$$q\phi'_{\mathrm{m}} = q\phi_{\mathrm{m}} + V_{\mathrm{m}} = q\phi_{\mathrm{m}} - q\Delta\phi_{\mathrm{m}} = q\phi_{\mathrm{m}} - q\sqrt{\frac{q\mathcal{E}}{4\pi\varepsilon_0}}\,, \tag{10.8}$$

where $q\Delta\phi_{\mathrm{m}} = -V_{\mathrm{m}}$ is the image-lowering potential energy. To see the effect of the image-lowering effect, one can consider two electric field strengths. If the applied electric field $\mathcal{E}$ is equal to 10^5 V/cm, then x_{m} is equal to 60 Å and $q\Delta\phi_{\mathrm{m}} = 0.12$ eV. On the other hand, if $\mathcal{E} = 10^7$ V/cm, then $x_{\mathrm{m}} = 6$ Å and $q\Delta\phi_{\mathrm{m}} = 1.2$ eV. Therefore, it is obvious that the effective metal work function is greatly reduced at high electric fields as a result of the image-lowering effect. Figure 10.3 shows the image-lowering potential versus square root of the applied electric field with

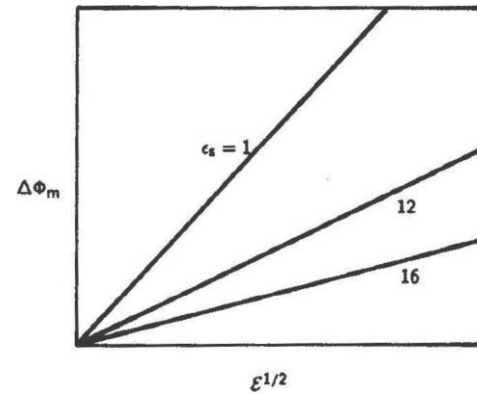


FIGURE 10.3. Image-lowering potential versus square root of the applied electric field with dielectric constant ε_{s} as parameter.

TABLE 10.1. Metal work functions for a clean metal surface in vacuum.

Metal(s)	Work function (eV)	Metal (s)	Work function (eV)	Metal	Work Function (eV)
Ti, Al, Ta, Ag	4.33	Sn	4.4	Ir	5.3
Au, Pd	5.10	W, Mo, Sb	4.63	T1	3.9
Pt	5.65	Ga, Cd	4.28	In	4.2
Cr, Hg	4.5	Ni	5.15	Zn	4.4
Mg	3.65	Rh	5.05	Fe	4.45
Cu	4.65	In	4.2	Mn	4.15
Si	4.85	Se	5.9	Co	5.0

dielectric constant ε_s as a parameter. Table 10.1 lists the work function data for some metals.

10.3. Thermionic Emission Theory

Thermionic emission usually refers to the emission of electrons from a hot metal surface. If the metal is used as a cathode, and all the emitted electrons from the metal surface are collected at the anode of a vacuum diode, then the cathode is in a saturation emission condition. The emitted current density is then called the saturation current density J_s, and the equation that relates J_s to the cathode temperature and the work function of a metal is known as the Richardson equation.

The Richardson equation is derived using the geometry of metal surface as shown in Figure 10.4. The surface is assumed infinite in the x–y plane and the electron emission is normal to the metal surface along the z-axis. The free electron

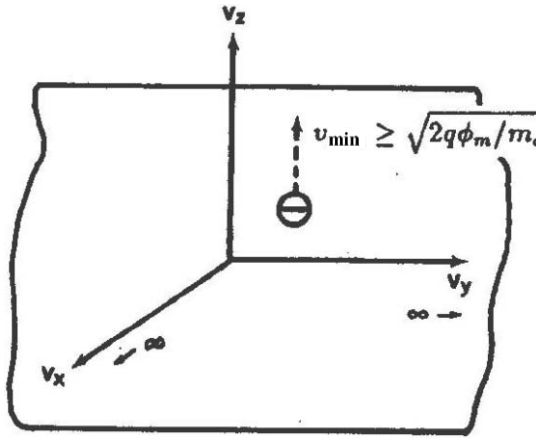


FIGURE 10.4. Thermionic emission of electrons from a metal surface.

density in the metal with velocity between (v_x, v_y, v_z) and $(v_x + \mathrm{d}v_x, v_y + \mathrm{d}v_y, v_z + \mathrm{d}v_z)$ is given by

$$\mathrm{d}n = 2\left(\frac{1}{2\pi}\right)^3 f(k)\,\mathrm{d}^3k = \left(\frac{2m_0^3}{h^3}\right) f(v)\,\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z, \tag{10.9}$$

where m_0 is the free electron mass and $\hbar k = m_0 v$. Using Maxwell–Boltzmann statistics, the electron distribution function $f(v)$ is given by

$$f(v) = \exp\left[-\frac{m_0(v_x^2 + v_y^2 + v_z^2)}{2k_\mathrm{B}T}\right]. \tag{10.10}$$

Now substituting (10.10) into (10.9) one obtains the thermionic emission current density in the z-direction, which is given by

$$\begin{aligned} J_\mathrm{s} &= \int q v_z\,\mathrm{d}n = \left(\frac{2qm_0^3}{h^3}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{m_0 v_x^2}{2k_\mathrm{B}T}\right)\mathrm{d}v_x \\ &\quad\times\int_{-\infty}^{\infty}\exp\left(-\frac{m_0 v_y^2}{2k_\mathrm{B}T}\right)\mathrm{d}v_y\int_{v_{z\mathrm{m}}}^{\infty} v_z\exp\left(-\frac{m_0 v_z^2}{2k_\mathrm{B}T}\right)\mathrm{d}v_z \\ &= A_0 T^2\exp\left(-\frac{q\phi_\mathrm{m}}{k_\mathrm{B}T}\right), \end{aligned} \tag{10.11}$$

where $A_0 = 4\pi q m_0 k_\mathrm{B}^2/h^3$ is the Richardson constant, which is equal to 120 A/(cm$^2\cdot$K^2) for electrons in free space. In (10.11), it is noted that only electrons with kinetic energies greater than the metal work function (i.e., $1/2(m_0 v_{z\mathrm{m}}^2) \geq q\phi_\mathrm{m}$) can escape from the metal surface along the z-direction.

It is noted from (10.11) that both the Richardson constant A_0 and the metal work function ϕ_m can be determined from the plot of $\ln(J_\mathrm{s}/T^2)$ versus $1/T$, as illustrated in Figure 10.5. The intercept of this plot with the ordinate yields A_0, while the slope gives the value of the metal work function ϕ_m.

The image-lowering effect should be considered in deriving the thermionic emission current density when an electric field is applied to the metal surface. In

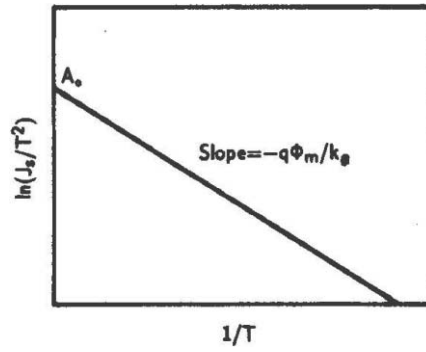


FIGURE 10.5. Plot of $\ln(J_\mathrm{s}/T^2)$ versus $1/T$ using (10.11). The Richardson constant A_0 and the metal work function ϕ_m can be determined from this plot.

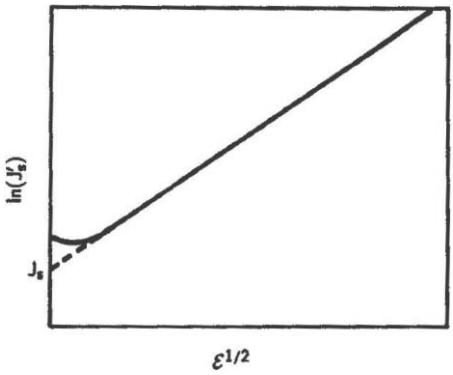


FIGURE 10.6. Plot of ln J_s' versus $\mathcal{E}^{1/2}$ for a thoriated tungsten metal, assuming $J_s = 1$ A/cm^2 at $T = 1873$ K.

this case, ϕ_m in the exponent of (10.11) is replaced by an effective metal work function ϕ_m' given by (10.8). By replacing ϕ_m' for ϕ_m in (10.11), one obtains an expression for the effective thermionic current density, which is given by

$$\begin{aligned} J_s' &= A_0 T^2 \exp\left(-\frac{q\phi_m'}{k_B T}\right) \\ &= A_0 T^2 \exp\left(-\frac{q\phi_m}{k_B T}\right) \exp\left[\left(\frac{q}{2k_B T}\right)\left(\frac{q\mathcal{E}}{\pi\varepsilon_0}\right)^{1/2}\right] \\ &= J_s \exp\left(\frac{4.39\mathcal{E}^{1/2}}{T}\right). \end{aligned} \tag{10.12}$$

It is noted that the exponential term in (10.12) is due to the image-lowering effect. In general, a plot of $\ln(J_s')$ versus $\mathcal{E}^{1/2}$ yields a straight line over a wide range of the electric field. However, deviation from linearity is expected at very low electric fields. Figure 10.6 shows the plot of $\ln(J_s')$ versus $\mathcal{E}^{1/2}$ for thoriated tungsten metal. Using (10.11) and assuming that $J_s = 1$ A/cm^2, $T = 1873$ K, and $A_0 = 120$ A/(cm$^2\cdot$ K^2), one obtains a value of $\phi_m = 3.2$ eV for thoriated tungsten metal.

10.4. Ideal Schottky Contact

According to the Schottky–Mott model, the barrier height of an ideal metal/n-type semiconductor Schottky contact is equal to the difference between the metal work function ϕ_m and the electron affinity χ_s of a semiconductor, which can be written as

$$\phi_{Bn} = \phi_m - \chi_s. \tag{10.13}$$

Figure 10.7 shows the schematic energy band diagrams for a metal/n-type semiconductor system before and after contact and under various conditions. Figures 10.7a, b, and c denote the cases for $\phi_m > \phi_s$, and Figures 10.7d, e, and

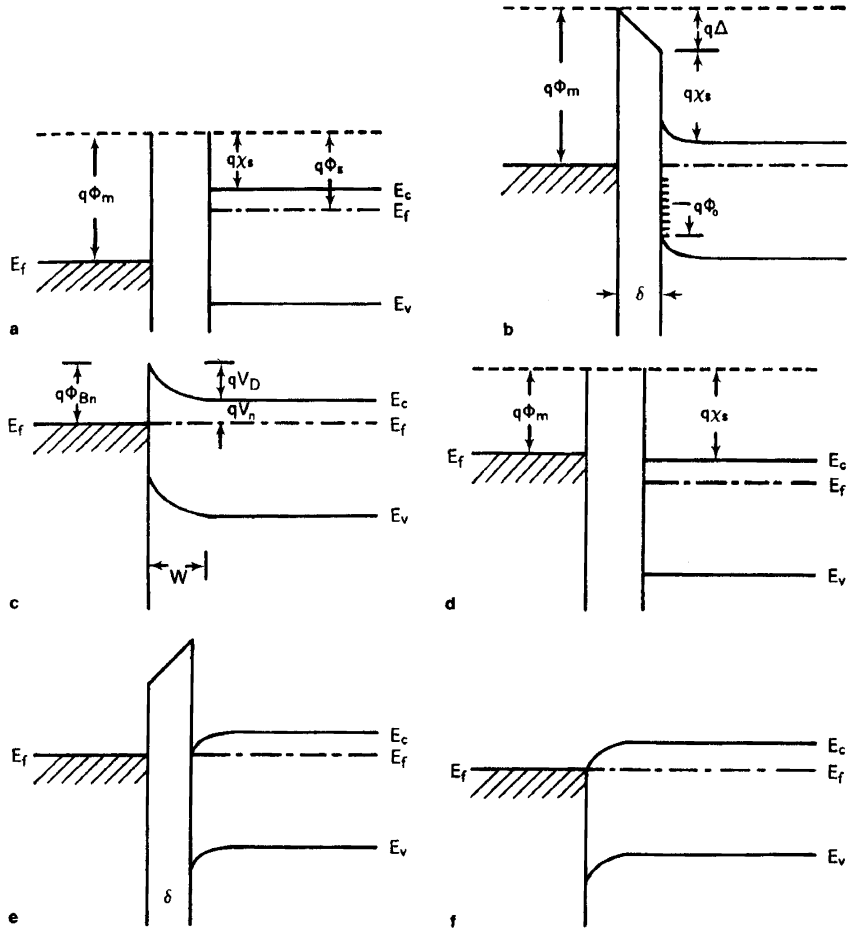


FIGURE 10.7. Energy band diagrams for an ideal metal/n-type semiconductor contact. (a) to (c) $\phi_m > \phi_s$: (a) before contact, (b) in contact with a small air gap and interface states, (c) in intimate contact (rectifying contact), with no interface states; (d) to (f) $\phi_m < \phi_s$, (d) before contact, (e) in contact with a small air gap, and (f) in intimate contact (ohmic contact).

f denote the cases with $\phi_m < \phi_s$. Figures 10.7a and d are before the contact, and Figures 10.7b and e are after the contact, assuming that a thin insulating interfacial layer (e.g., 20–30 Å) exists between the metal and semiconductor. Figures 10.7c and f pertain to intimate contact without the insulating interfacial layer. From Figure 10.7c it is seen that for $\phi_m > \phi_s$, there exists a potential barrier for electrons to cross from the metal to the semiconductor, and the metal–semiconductor contact exhibits a rectifying behavior. However, an ohmic contact is obtained if $\phi_m < \phi_s$, as shown in Figure 10.7f. For a metal/p-type semiconductor contact, the opposite behavior results. It should be noted that the measured barrier heights for most of

the metal/n-type semiconductor contacts do not always follow the simple prediction given by (10.13), owing to the fact that it does not consider the interface state density and the image-lowering effect. In fact, for many III-V compound semiconductors, because of the high surface state density and Fermi-level pinning at the interface states, the barrier height for the Schottky contacts formed on III-V semiconductor materials was found to be independent of the metals used. A detailed explanation of this result will be given in Section 10.7.

Similarly, the barrier height for an ideal metal/p-type semiconductor Schottky contact can be expressed by

$$\phi_{\mathrm{Bp}} = \frac{E_{\mathrm{g}}}{q} - (\phi_{\mathrm{m}} - \chi_{\mathrm{s}}) = \frac{E_{\mathrm{g}}}{q} - \phi_{\mathrm{Bn}}, \tag{10.14}$$

where E_{g} is the energy band gap and q is the electronic charge. Equation (10.14) shows that for a given metal–semiconductor system, the sum of barrier heights for a metal on n- and p-type semiconductor contacts is equal to the band gap energy of the semiconductor (i.e., $q\phi_{\mathrm{Bn}} + q\phi_{\mathrm{Bp}} = E_{\mathrm{g}}$). As shown in Figure 10.7c, the potential difference, $q(\phi_{\mathrm{m}} - \chi_{\mathrm{s}} - V_{\mathrm{n}})$, known as the contact potential or the diffusion potential V_{D}, can be expressed by

$$V_{\mathrm{D}} = \phi_{\mathrm{m}} - \Phi_{\mathrm{s}} = \phi_{\mathrm{Bn}} - V_{\mathrm{n}}, \tag{10.15}$$

where ϕ_{Bn} is the barrier height and $V_{\mathrm{n}} = (E_{\mathrm{c}} - E_{\mathrm{f}})/q = (k_{\mathrm{B}}T/q)\ln(N_{\mathrm{c}}/N_{\mathrm{D}})$ is the Fermi (or chemical) potential of an n-type semiconductor.

Equation (10.15) shows that the contact (or diffusion) potential for an ideal metal/n-type Schottky barrier diode is equal to the difference between the metal work function and the semiconductor work function, or the difference between the Schottky barrier height and the Fermi potential of an n-type semiconductor.

To find the spatial distributions of potential and electric fields, the depletion layer width, and the junction capacitance of a Schottky diode, one needs to solve the Poisson equation in the space-charge region using proper boundary conditions. The one-dimensional (1-D) Poisson equation in the depletion region of a Schottky diode is given by

$$\frac{\mathrm{d}^2 V(x)}{\mathrm{d}x^2} = -\frac{\rho}{\varepsilon_0 \varepsilon_{\mathrm{s}}}, \tag{10.16}$$

where ε_{s} is the dielectric constant of the semiconductor and ε_0 is the permittivity of free space. The charge density for $0 \leq x \leq W$ is given by

$$\rho = q[N_{\mathrm{D}} - n(x)], \tag{10.17}$$

where $n(x)$ is the electron density in the space-charge region, which is equal to $n_0 \exp(-qV_{\mathrm{D}}/k_{\mathrm{B}}T)$ at the edge of the depletion layer (i.e., at $x = W$). It is noted that $n(x)$ decreases exponentially with distance from the depletion layer edge (at $x = W$) into the space-charge region.

Using a one-sided abrupt junction approximation and assuming that $n(x) = 0$ for $0 < x < W$, one can obtain the spatial distribution of the electric field by

integrating (10.16) once, with the result

$$\mathcal{E}(x) = -\frac{\mathrm{d}V(x)}{\mathrm{d}x} = \left(\frac{qN_\mathrm{D}}{\varepsilon_0\varepsilon_\mathrm{s}}\right)x + C_1, \tag{10.18}$$

where C_1 is a constant to be determined by the boundary conditions.

The potential distribution can be obtained by integrating (10.18) once more, which yields

$$V(x) = -\left(\frac{qN_\mathrm{D}}{2\varepsilon_0\varepsilon_\mathrm{s}}\right)x^2 - C_1x + C_2, \tag{10.19}$$

where C_2 is another constant of integration. The constants C_1 and C_2 can be determined using the following boundary conditions:

$$\begin{aligned} V(0) &= -\phi_\mathrm{Bn} \quad \text{at } x = 0, \\ \mathcal{E}(x) &= -\frac{\mathrm{d}V(x)}{\mathrm{d}x} = 0 \qquad \text{at } x = W. \end{aligned} \tag{10.20}$$

Solving (10.18), (10.19), and (10.20), one obtains

$$C_1 = -\frac{qN_\mathrm{D}W}{\varepsilon_0\varepsilon_\mathrm{s}}, \quad C_2 = -\phi_\mathrm{Bn}. \tag{10.21}$$

Now substituting C_1 and C_2 given by (10.20) and (10.21) into (10.18) and (10.19), one obtains the spatial distributions of the electric field and potential inside the depletion region, which are given respectively by

$$\mathcal{E}(x) = \left(\frac{qN_\mathrm{D}}{\varepsilon_0\varepsilon_\mathrm{s}}\right)(x - W), \tag{10.22}$$

$$V(x) = -\left(\frac{qN_\mathrm{D}}{\varepsilon_0\varepsilon_\mathrm{s}}\right)\left(\frac{x^2}{2} - Wx\right) - \phi_\mathrm{Bn}. \tag{10.23}$$

The depletion layer width W can be expressed in terms of N_D, V_D, and V_a across the barrier. From Figure 10.8a and (10.23) one obtains the potential at $x = W$ as

$$V(W) = (V_\mathrm{D} - V_\mathrm{a}) - \phi_\mathrm{Bn} = \left(\frac{qN_\mathrm{D}W^2}{2\varepsilon_0\varepsilon_\mathrm{s}}\right) - \phi_\mathrm{Bn}. \tag{10.24}$$

From (10.24), the depletion layer width W is given by

$$W = \sqrt{\frac{2\varepsilon_0\varepsilon_\mathrm{s}(V_\mathrm{D} - V_\mathrm{a})}{qN_\mathrm{D}}}. \tag{10.25}$$

It is seen from (10.25) that the depletion layer width is directly proportional to the square root of the applied voltage (V_a), and is inversely proportional to the square root of the dopant density of the semiconductor. Furthermore, (10.25) shows that the depletion layer width decreases with the square root of the forward-bias voltage (i.e., for $V_\mathrm{a} \geq 0$), and increases with the square root of the reverse-bias voltage (i.e., for $V_\mathrm{a} < 0$).

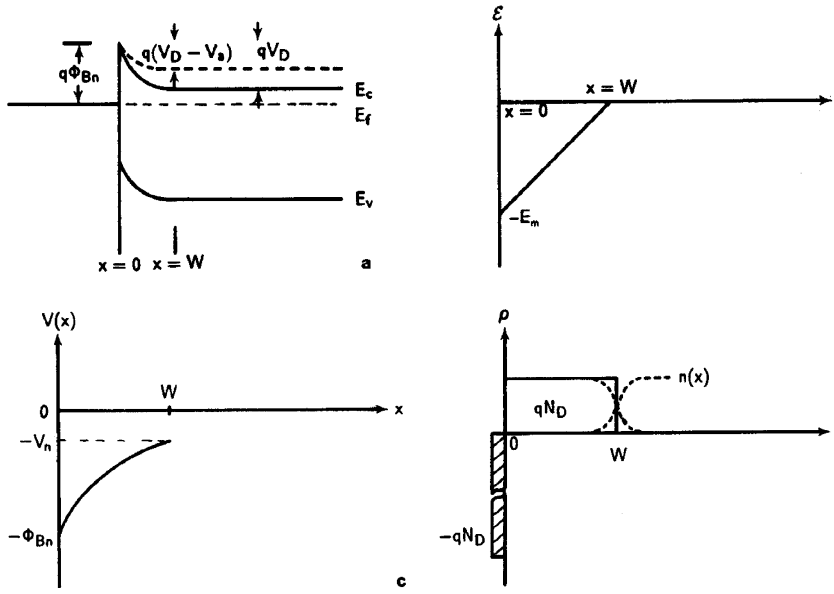


FIGURE 10.8. (a) Energy band diagram, (b) electric field, (c) potential distribution, and (d) space-charge distribution for a metal/n-type semiconductor Schottky barrier diode.

To find the depletion layer capacitance, it is noted in Figure 10.8d that the space charge per unit area, Q_s, in the depletion region is given by

$$Q_s = qN_D W = \sqrt{2qN_D\varepsilon_0\varepsilon_s(V_D - V_a)}. \tag{10.26}$$

The depletion layer capacitance per unit area can be obtained by differentiating (10.26) with respect to the applied voltage V_a, which yields

$$C_d = \frac{dQ_s}{dV_a} = \sqrt{\frac{qN_D\varepsilon_0\varepsilon_s}{2(V_D - V_a)}}. \tag{10.27}$$

Equation (10.27) shows that the depletion layer capacitance is inversely proportional to the square root of the applied voltage. Figure 10.8a shows the energy band diagram for a metal/n-type semiconductor Schottky barrier diode in thermal equilibrium (solid line) and under forward-bias conditions (dashed line). Figure 10.8b illustrates the spatial dependence of the electric field in the depletion region. From (10.22), the maximum electric field, which occurs at $x = 0$, is given by

$$\mathcal{E}_m = -\frac{qN_D W}{\varepsilon_0\varepsilon_s}. \tag{10.28}$$

The spatial distributions of the potential and the space charge in the depletion region are shown in Figures 10.8c and d, respectively. In Figure 10.8d the dashed line denotes the actual charge distribution, which shows that at $x = W$ the free electron density n_0 decreases exponentially with distance as it spreads into the

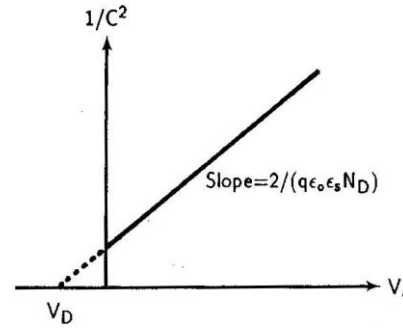


FIGURE 10.9. Square of the inverse capacitance versus the applied voltage for a metal/n-type semiconductor Schottky barrier contact.

depletion region. The solid line is the abrupt junction approximation that was used in the present derivation. The above analysis is valid only for an ideal Schottky diode in which both the surface states and the image-lowering effect are neglected. Figure 10.9 shows a plot of $1/C_d^2$ versus the applied bias voltage V_a. A linear relation is obtained if N_D is constant throughout the depletion region, and N_D can be determined from the slope of this plot, while the intercept at the horizontal axis yields V_D. From the measured V_D, the value of barrier height ϕ_{Bn} can be calculated from (10.15).

10.5. Current Flow in a Schottky Diode

A metal–semiconductor Schottky barrier diode is a majority-carrier device, because the current flow in such a device is due to the majority carriers (e.g., electrons in an n-type semiconductor). This is in contrast to a p-n junction diode, in which both the majority and minority carriers participate in the current conduction. To illustrate the current flow in a Schottky diode, the energy band diagrams and current components for an ideal metal/n-type semiconductor Schottky barrier diode under zero-bias, forward-bias, and reverse-bias conditions are shown in Figures 10.10a,

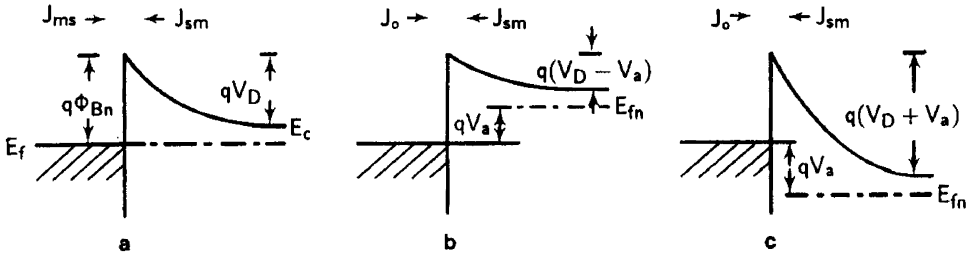


FIGURE 10.10. Energy band diagrams and current components for a Schottky barrier diode under (a) zero bias, (b) forward bias, and (c) reverse bias. Here J_{sm} denotes the current flow from semiconductor to metal, J_{ms} is the current density from metal to semiconductor, and $J_0 = J_{ms}$ is the saturation current density.

b, and c, respectively. The potential barrier for electrons moving from the semiconductor side to the metal side is designated as V_D, while the potential barrier for electrons moving from the metal side to the semiconductor side is defined as ϕ_{Bn}.

If a forward-bias voltage V_a is applied to the Schottky diode, then the potential barrier on the semiconductor side of the diode is reduced to $V_D - V_a$, as shown in Figure 10.10b. It is noted that the barrier height remains relatively unaffected by the applied bias voltage or the doping density of the semiconductor. Thus, the current flow from the semiconductor to the metal increases dramatically under forward-bias conditions, while the current flow from the metal to the semiconductor remains essentially the same. Under forward-bias conditions, the net current flow is controlled by the electron current flow from the semiconductor to the metal, as shown in Figure 10.10b. Under reverse-bias conditions, the potential barrier on the semiconductor side increases to $V_D + V_a$, and the current flow from the semiconductor to the metal becomes negligibly small compared to the current flow from the metal to the semiconductor. Thus, the net current flow under reverse-bias conditions is controlled by the thermionic emission from the metal to the semiconductor, as shown in Figure 10.10c.

The carrier transport and current flow in a Schottky barrier diode can be analyzed using the thermionic emission, the diffusion, or the combined thermionic–diffusion model. The current–voltage (I–V) equation derived from these models may be used to predict the current versus temperature or voltage behavior in a Schottky barrier diode. The simple thermionic emission model developed by Bethe and the diffusion model developed by Schottky are the most widely used physical models for predicting the I–V characteristics of a Schottky barrier diode. In this section, the current density equations are derived from both the thermionic emission and diffusion models. In addition, the current density expression obtained from the combined thermionic–diffusion model developed by Sze and Crowell[1] is also given. Finally, the tunneling phenomenon in a highly doped Schottky contact will also be described.

10.5.1. The Thermionic Emission Model

The thermionic emission model described in Section 10.3 for electron emission from a hot metal surface into free space can be easily modified for a metal–semiconductor system. The current flow from semiconductor to metal in a Schottky diode is determined mainly by the barrier potential ($V_D - V_a$) under a forward-bias condition. To overcome this potential barrier, the minimum kinetic energy of electrons in the semiconductor side along the x-direction is given by

$$\frac{1}{2}\left(m_n^* v_{xm}^2\right) \geq q(V_D - V_a). \tag{10.29}$$

Therefore, the electron current density component flowing from semiconductor to metal side, J_{sm}, can be obtained by modifying (10.11) for the

metal–semiconductor contacts, yielding

$$\begin{aligned} J_{\mathrm{sm}} &= \left(\frac{2qm^{*^3}}{h^3}\right)\int_{-\infty}^{\infty}\exp\left[\frac{-m^*v_z^2}{2k_{\mathrm{B}}T}\right]\mathrm{d}v_z\int_{-\infty}^{\infty}\exp\left[\frac{-m^*v_{\mathrm{y}}^2}{2k_{\mathrm{B}}T}\right]\mathrm{d}v_y \\ &\quad\times\int_{v_{\mathrm{xm}}}^{\infty}v_x\exp\left[\frac{-m^*v_x^2}{2k_{\mathrm{B}}T}\right]\mathrm{d}v_x \\ &= A^*T^2\exp\left(-\frac{q\phi_{\mathrm{Bn}}}{k_{\mathrm{B}}T}\right)\exp\left(\frac{qV_{\mathrm{a}}}{k_{\mathrm{B}}T}\right) \\ &= J_0\exp\left(\frac{qV_{\mathrm{a}}}{k_{\mathrm{B}}T}\right), \end{aligned} \tag{10.30}$$

where

$$J_0 = A^*T^2\exp\left(-\frac{q\phi_{\mathrm{Bn}}}{k_{\mathrm{B}}T}\right) \tag{10.31}$$

is the saturation current density. In (10.31), $A^* = 4\pi m_{\mathrm{n}}^* q k_{\mathrm{B}}^2/h^3$ is the effective Richardson constant, m_{n}^* is the electron effective mass, and ϕ_{Bn} is the barrier height. The current flow from metal to semiconductor side can be obtained from (10.30) by using the fact that in thermal equilibrium, $V_{\mathrm{a}} = 0$ and

$$J_{\mathrm{ms}} = -J_{\mathrm{sm}} = -J_0. \tag{10.32}$$

Thus, the total current flow under forward-bias conditions is equal to the sum of (10.30) and (10.32), which reads

$$J = J_{\mathrm{sm}} + J_{\mathrm{ms}} = J_0\left[\exp\left(\frac{qV_{\mathrm{a}}}{k_{\mathrm{B}}T}\right) - 1\right]. \tag{10.33}$$

Equation (10.33) is the well-known Schottky diode equation, which predicts an exponential dependence of the current density on both the temperature and applied bias voltage. Since the saturation current density J_0 depends exponentially on the barrier height, a large barrier height is needed in order to reduce the value of J_0 in a Schottky diode. Methods of increasing the effective barrier height of a Schottky barrier diode will be discussed in Section 10.8.

10.5.2. Image-Lowering Effect

As in the case of a metal–vacuum interface, the image-lowering effect also exists in the metal–semiconductor interface, as shown in Figure 10.11. Taking into account the image-lowering effect in (10.31), the saturation current density can be expressed as

$$\begin{aligned} J_0 &= A^*T^2\exp\left[-\frac{q(\phi_{\mathrm{Bn}} - \Delta\phi_{\mathrm{m}})}{k_{\mathrm{B}}T}\right] \\ &= A^*T^2\exp\left(-\frac{q\phi_{\mathrm{Bn}}}{k_{\mathrm{B}}T}\right)\exp\left(\frac{q^3\mathcal{E}_{\mathrm{m}}}{4\pi\varepsilon_0\varepsilon_{\mathrm{s}}k_{\mathrm{B}}^2T^2}\right)^{1/2}. \end{aligned} \tag{10.34}$$

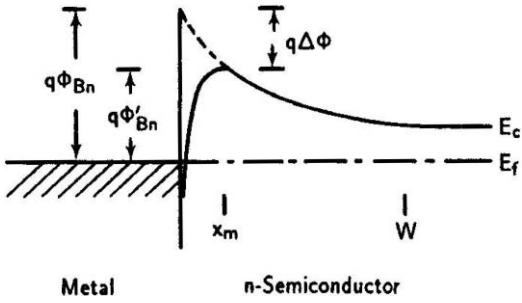


FIGURE 10.11. Energy band diagram for a metal/n-type semiconductor Schottky barrier diode showing the image-lowering effect; $q\,\Delta\phi$ is the image-lowering potential.

In the depletion region, the maximum field strength $\mathcal{E}_{\rm m}$ at the metal–semiconductor interface can be obtained from solving (10.25) and (10.28), with result

$$\mathcal{E}_{\rm m} = \sqrt{\frac{2qN_{\rm D}(V_{\rm D} - V_{\rm a})}{\varepsilon_0 \varepsilon_{\rm s}}}. \tag{10.35}$$

As shown in Figure 10.8b, the maximum electric field occurs at $x = 0$, and the field decreases linearly with distance from the metal–semiconductor interface (i.e., $x = 0$) to the edge of the depletion layer (i.e., $x = W$) in the bulk semiconductor. From (10.34) and (10.35) it is noted that $\ln(J_0)$ is directly proportional to $(\mathcal{E}_{\rm m})^{1/2}$, or $(V_{\rm D} - V_{\rm a})^{1/4}$ when the image-lowering-effect is considered. This current–voltage (I–V) behavior has indeed been observed in many metal–semiconductor Schottky barrier diodes.

10.5.3. The Diffusion Model

The Schottky diffusion model is based on the assumption that the barrier height is greater than a few $k_{\rm B}T$ and that the semiconductor is lightly doped so that the depletion layer width is larger than the carrier diffusion length. Based on this model, both the drift and diffusion current components are considered in the depletion region, and the electron current density $J_{\rm n}$ can be written as

$$\begin{aligned} J_{\rm n} &= qn(x)\mu_{\rm n}\mathcal{E}_x + qD_{\rm n}\frac{{\rm d}n(x)}{{\rm d}x} \\ &= qD_{\rm n}\left[\left(\frac{qn(x)}{k_{\rm B}T}\right)\left(-\frac{{\rm d}V(x)}{{\rm d}x}\right) + \frac{{\rm d}n(x)}{{\rm d}x}\right]. \end{aligned} \tag{10.36}$$

It is noted that the Einstein relation $\mu_{\rm n} = (q/k_{\rm B}T)D_{\rm n}$ and $\mathcal{E}_x = -{\rm d}V(x)/{\rm d}x$ were used in (10.36). Since the total current density $J_{\rm n}$ in the depletion region is constant and independent of x, one can multiply both sides of (10.36) by $\exp[-qV(x)/k_{\rm B}T]$ and then integrate the equation over the entire depletion region from $x = 0$ to

$x = W$, which yields

$$J_n \int_0^W e^{-qV(x)/k_B T}\,dx = qD_n \int_0^W \left[-\left(\frac{qn(x)}{k_B T}\right)\frac{dV(x)}{dx}\, e^{-qV(x)/k_B T} + \frac{dn(x)}{dx}\, e^{-qV(x)/k_B T}\right] dx, \tag{10.37}$$

or

$$J_n \int_0^W e^{-qV(x)/k_B T}\,dx = qD_n n(x)\, e^{-qV(x)/k_B T}\Big|_0^W. \tag{10.38}$$

The boundary conditions for (10.38) at $x = 0$ and $x = W$ are given by

$$qV(0) = -q\phi_{Bn} \quad \text{and} \quad qV(W) = -q(V_n + V_a), \tag{10.39}$$

where $qV_n = E_c - E_f$ and V_a is the applied voltage. The electron densities at $x = 0$ and $x = W$ are given by

$$n(0) = N_c \exp\left\{-\frac{[E_c(0) - E_f]}{k_B T}\right\} = N_c \exp\left(-\frac{q\phi_{Bn}}{k_B T}\right), \tag{10.40a}$$

$$n(W) = N_c \exp\left(-\frac{qV_n}{k_B T}\right). \tag{10.40b}$$

Now substituting (10.39) and (10.40) into (10.38), one obtains

$$J_n = \frac{(qD_n N_c)\left[\exp(qV_a/k_B T) - 1\right]}{\int_0^W \exp\left[-qV(x)/k_B T\right] dx}. \tag{10.41}$$

The integral in the denominator of (10.41) can be carried out by substituting $V(x)$ given in (10.23) (neglecting the x^2 term) and W given in (10.25) into (10.41), and one obtains

$$\begin{aligned} J_n &= \left(\frac{q^2 D_n N_c}{k_B T}\right)\sqrt{\frac{2q(V_D - V_a)N_D}{\varepsilon_0 \varepsilon_s}}\exp\left(-\frac{q\phi_{Bn}}{k_B T}\right)\left[\exp\left(\frac{qV_a}{k_B T}\right) - 1\right] \\ &= J_0'\left[\exp\left(\frac{qV_a}{k_B T}\right) - 1\right], \end{aligned} \tag{10.42}$$

where

$$J_0' = \left(\frac{q^2 D_n N_c}{k_B T}\right)\sqrt{\frac{2q(V_D - V_a)N_D}{\varepsilon_0 \varepsilon_s}}\exp\left(-\frac{q\phi_{Bn}}{k_B T}\right) \tag{10.43}$$

is the saturation current density derived from the diffusion model.

A comparison of (10.43) and (10.31) reveals that the saturation current density derived from the thermionic emission model is more sensitive to temperature than that from the diffusion model. However, the latter shows a stronger dependence on the applied-bias voltage than the former. It is noted that the image-lowering effect

is neglected in (10.43). Both models predict the same exponential dependence of the saturation current density on the barrier height and the temperature.

Finally, a synthesis of the diffusion and thermionic emission models has been reported by Crowell and Sze.[1] The so-called thermionic-emission diffusion model uses the boundary conditions of the thermionic recombination velocity at the metal–semiconductor interface and considers the effects of electron-optical phonon scattering and quantum-mechanical reflection at the metal–semiconductor interface. The current density equation derived from the thermionic-emission diffusion model is given by[1]

$$J = \frac{qN_c v_R}{(1 + v_R/v_D)} \exp\left(-\frac{q\phi_{Bn}}{k_B T}\right)\left[\exp\left(\frac{qV_a}{k_B T}\right) - 1\right], \tag{10.44}$$

where v_R is the recombination velocity at the interface and v_D is the diffusion velocity associated with electron transport from the depletion layer edge at W to the potential energy maximum at x_m. If v_R is much larger than v_D, then the diffusion process is dominant. On the other hand, if v_D is much greater than v_R, then the preexponential factor in (10.44) is dominated by v_R and the thermionic emission current becomes the predominant current component.

Finally, it should be noted that if a metal–semiconductor Schottky contact is formed on a degenerate semiconductor, the barrier width becomes very thin, so that the flow of current through the Schottky contact is dominated by a tunneling process. In this case, the current flow in the diode is determined by the quantum-mechanical tunneling transmission coefficient, and the tunneling current density is proportional to the exponential function of the barrier height and doping density, which is given by

$$J_t \approx \exp(-q\phi_{Bn}/E_{00}), \tag{10.45}$$

where $E_{00} = (q\hbar/2)\sqrt{N_D/m^*\varepsilon_0\varepsilon_s}$. From (10.45), it is seen that the tunneling current density will increase exponentially with the square root of dopant density and decrease exponentially with increasing barrier height. Equation (10.45) may be applied to analyze the specific contact resistance for the ohmic contact on a heavily doped semiconductor. This will be discussed further in Section 10.10.

10.6. Current–Voltage Characteristics of a Si and a GaAs Schottky Diode

In this section, the current–voltage (I–V) characteristics of a Au/n-type Si Schottky diode and an Au/n-type GaAs Schottky diode are described. The experimental results for both diodes under forward-bias conditions are shown in Figures 10.12 and 10.14, respectively.[2] In a practical Schottky barrier diode the slope of the I–V curve under forward-bias conditions is usually greater than unity; a diode ideality factor "n" is incorporated in (10.33). A semiempirical formula for predicting the

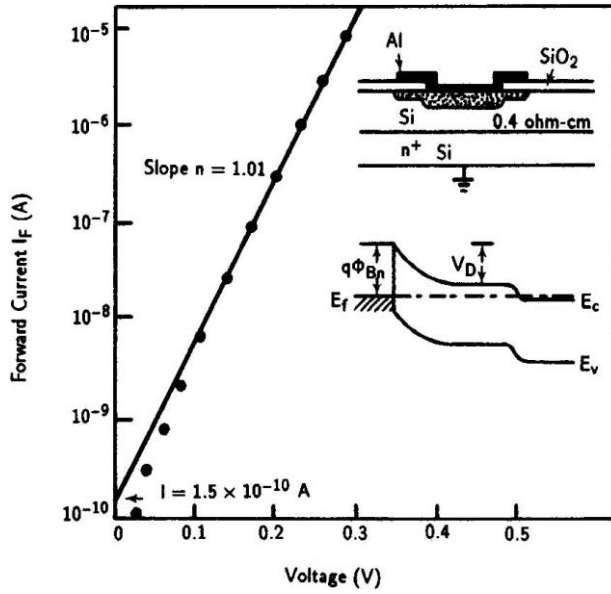


FIGURE 10.12. Forward *I–V* curve and the energy band diagram for a Al/n-type Si Schottky diode with a field-plate structure. After Yu and Mead,[2] by permission.

I–V characteristics of a practical Schottky diode is given by

$$J = J_0 \left[\exp\left(\frac{q V_a}{n k_B T} \right) - 1 \right], \tag{10.46}$$

where J_0 is the saturation current density given by (10.31). Under forward-bias conditions and for $q V_a \geq 3 k_B T$, (10.46) becomes

$$J_F \approx J_0 \exp\left(\frac{q V_a}{n k_B T} \right). \tag{10.47}$$

For an ideal metal–semiconductor Schottky diode, the diode ideality factor n is equal to 1. Deviation of n from unity may be attributed to a number of factors such as large surface leakage current, high density of bulk recombination centers in the depletion region, and high interface state density as well as high series resistance.

The metal/n-type Si Schottky barrier diode with diode ideality factor n varying from 1.01 to 1.12 has been reported in the literature. To achieve near-ideal *I–V* characteristics for the Si Schottky barrier diodes, various fabrication techniques have been developed in the past. The two most widely used techniques to achieve near-ideal Schottky contacts are the field-plate and guard-ring structures. Figure 10.12 shows an Al/n-type Si Schottky barrier diode with a field-plate structure. As shown in Figure 10.12, a field oxide (e.g., SiO_2) is grown underneath the edge of an Al Schottky contact. The Al film is overlaid on top of this field oxide to serve as a field plate. When the Schottky diode is reverse-biased, this field plate keeps

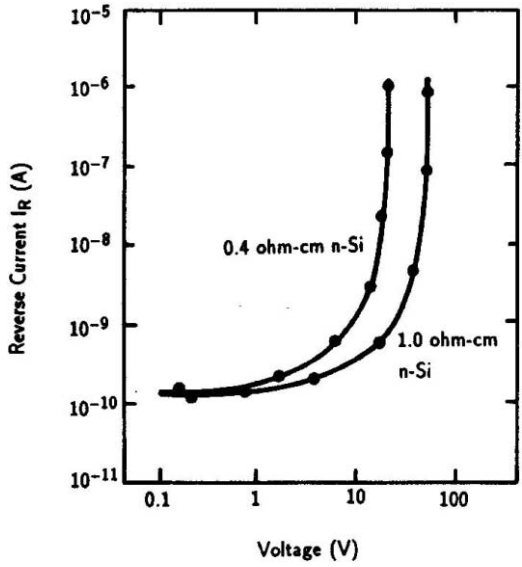


FIGURE 10.13. Reverse *I–V* curves for the A1/n-type silicon Schottky barrier diode with two different substrate resistivities. Al overlaid on SiO_2 is used to control the soft breakdown due to the edge effect. After Yu and Mead,[2] by permission.

the underlying contact surface fully depleted so that soft breakdown arising from the surface accumulation layer formed around the edge of the metal plate does not occur in this structure. As can be seen in Figure 10.12, the *I–V* characteristics for this diode closely follow the theoretical prediction given by the thermionic emission model for about six decades of current with values of n very close to unity. The intercept of the forward current at zero bias gives a barrier height of $\phi_{Bn} = 0.70$ V. The barrier height deduced from the activation energy plot of $\ln(J_F)$ versus $1/T$ at a fixed forward bias is found to be equal to 0.69 V. This value is in good agreement with the value determined using the photoemission excitation of electrons from metal into the semiconductor. The reverse *I–V* curves for an Al/n-type Si Schottky barrier diode are also displayed in Figure 10.13 for two different substrate resistivities (i.e., $\rho = 0.4$ and $1.0\ \Omega \cdot \text{cm}$). The breakdown voltages for both diodes are presumably limited by the metal edge curvature in the depletion region. Figure 10.14 shows the near-ideal forward *I–V* characteristics for three Au/n-type GaAs Schottky barrier diodes formed on the GaAs substrates with different crystal orientations.[3]

Another Schottky barrier diode structure with near-ideal *I–V* characteristics is obtained using a p-type diffused guard-ring structure on an n-type silicon substrate, as shown in Figure 10.15.[4] A p-type diffused guard-ring structure is extended in the normal planar fashion under the oxide. The PtSi Schottky barrier contact formed on the n-type silicon inside the p^+ guard ring is in electrical contact with the p-type Si substrate. The doping profile of the p^+ guard ring is tailored in such a way that the breakdown voltage of the p-n junction in the guard-ring region is higher than that of the Schottky barrier contact. In this structure, the region of maximum electric field depends on the depth and profile of the diffused junction. For an ideal linearly graded junction, the breakdown voltage is higher than that of a planar junction.

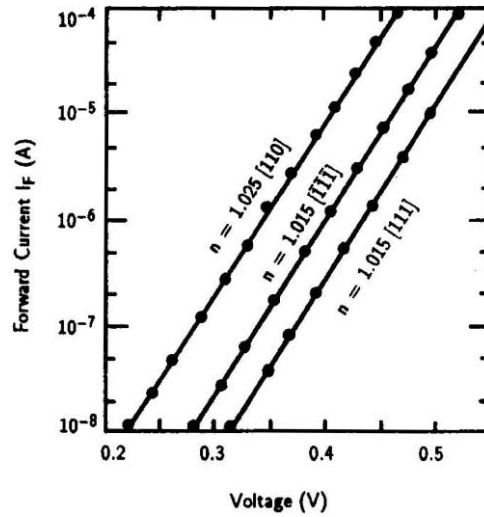


FIGURE 10.14. Forward *I–V* curves for Au/n-type GaAs diodes fabricated on different substrate orientations. After Kahng,[3] by permission.

Figure 10.16 shows the reverse *I–V* characteristics of a PtSi/n-Si Schottky diode with a diffused guard-ring structure.[4] The solid line is for the experimental data, while the dashed line is calculated from (10.34) by including the image lowering effect.

In silicon integrated circuits, aluminum and its alloy (Al–Cu) have been widely used for ohmic contacts and interconnects for silicon devices and silicon integrated circuits. In addition, aluminum is also widely used as a gate metal for silicon MOS devices and as Schottky contacts for bipolar transistor circuits. Unfortunately, the aluminum/silicon system has low eutectic temperature (577°C) and interdiffusion occurs at a relatively low temperature (i.e., approximately 400°C). As a result, large leakage current is often observed in the silicon shallow junction bipolar transistors and the n-p junction diodes when aluminum is used for interconnects and ohmic contacts. To overcome this problem, metal silicides with low resistivity and high-temperature stability are required for contacts in silicon integrated circuits (ICs). Silicide is a metal–silicon compound, which can be formed with

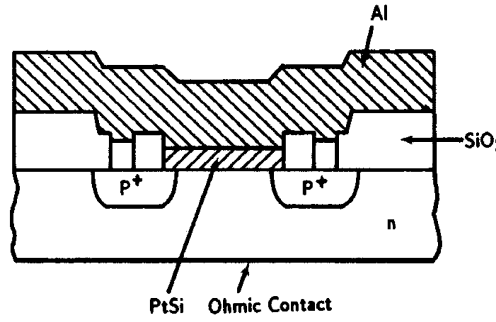


FIGURE 10.15. A PtSi-n-Si Schottky barrier diode with a diffused guard-ring structure. After Lepselter and Sze,[4] by permission.

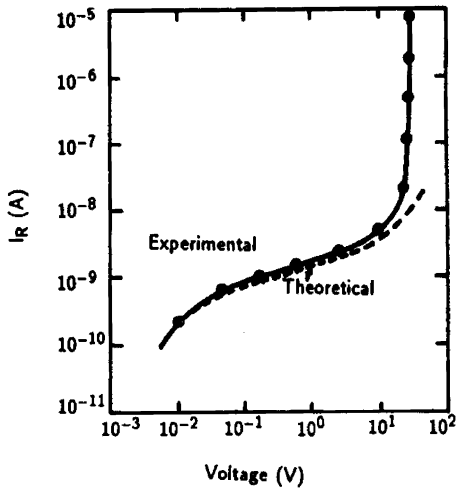


FIGURE 10.16. Comparison of theoretical and measured reverse *I–V* characteristics for a PtSi/n-type Si Schottky barrier diode shown in Figure 10.15. After Lepselter and Sze,[4] by permission.

a specific ratio of metal–silicon composition. Important silicides for silicon are those of the refractory metals such as Mo, Ti, Ta, and W and the near-noble metals such as Co, Ni, Pt, and Pd. Silicides formed from these metals have low resistivity, high eutectic temperature, good adhesive characteristics, and stability. The most stable silicides are the silicon-rich metal disilicides (e.g., $CoSi_2$, $MoSi_2$, $TiSi_2$, and WSi_2), which have eutectic temperatures ranging from 1195 to 1440°C and a resistivity of 2 to 4×10^{-5} $\Omega \cdot$cm. The reaction temperatures for these silicides may vary from 350 to 650°C. Schottky barrier diodes formed on these silicides have barrier heights varying between 0.58 and 0.67 eV on n-type silicon. The most widely used metal silicide Schottky barrier contact in bipolar circuit applications is the PtSi/n-Si system. The barrier height for a PtSi/n-Si Schottky barrier diode is around 0.90 eV, which is probably the highest barrier height (without barrier height enhancement) for a silicon Schottky barrier diode. In addition, high-quality PtSi/p-type silicon Schottky barrier diodes with low barrier height of 0.2 eV have been developed for mid-IR (3–5 μm) photodetector array applications.

In recent years, metal silicides have been widely used at the source or drain contact region of silicon MOS transistors. In processing technology, the silicide is formed by depositing the metal onto the exposed silicon area and followed by annealing to form the silicide film. The annealing occurs at temperatures well below the melting point of the silicon, but solid-state interdiffusion takes place and a silicide film is formed. For metals deposited on Si, different silicide compounds are formed under different annealing conditions. For example, in the case of Pt deposited on Si, a Pt_2Si film will form at around 300°C, and it transforms into PtSi with further annealing at 450°C. Although PtSi has been widely used in silicon ICs, the PtSi films are not very stable under high-temperature operation and hence require further processing steps. The group of refractory metal silicides of titanium (Ti), tantalum (Ta), molybdenum (Mo), and tungsten (W) has proved

stable at high-temperature operation. For example, Ti film deposited on Si forms stable $TiSi_2$ compound following a 650°C annealing. $TiSi_2$ has been widely used in VLSI device contacts. In addition, silicide films can also be formed by epitaxial growth. For example, epitaxial silicides, such as $CoSi_2$ and $NiSi_2$, which have cubic crystal structure, have been used as low-resistivity contacts and in novel high-speed device structures, such as metal-based transistors. More recently, epitaxial silicide film of $TiSi_2$ has also been reported for use in VLSI circuits and devices.

10.7. Determination of Schottky Barrier Height

Expressions of the barrier height for an ideal metal on n- and p-type semiconductor Schottky barrier diodes are given by (10.13) and (10.14), respectively. However, these expressions are valid only when the image lowering effect is negligible and the surface state density is small. However, in most III-V compound semiconductors, the surface state density is usually very high. As a result, it is necessary to include the interface state effect in the barrier height expression. As shown in Figure 10.7b, the effect of the surface states is represented by the energy level $q\phi_0$. This energy level coincides with the Fermi level at the semiconductor surface before the metal–semiconductor contact is formed. In fact, $q\phi_0$ could be considered as a demarcation level in which the surface states below it must be filled in order to satisfy the charge-neutrality condition at the surface. If the surface states become very large, then the Fermi level at the surface will be pinned at $q\phi_0$, and the barrier height for a metal–semiconductor contact becomes independent of the metal work function. This has indeed been observed in many Schottky barrier contacts formed on III-V compound semiconductors. Cowley and Sze have derived a general expression for the barrier height by taking into account the effects of image lowering and surface state density. This is given by[5]

$$\phi_{Bn} = c_2(\phi_m - \chi_s) + (1 - c_2)(E_g/q - \phi_0) - \Delta\phi = c_2\phi_m + c_3, \tag{10.48}$$

where

$$c_2 = \frac{\varepsilon_i\varepsilon_0}{\varepsilon_i\varepsilon_0 + q^2\delta D_s}. \tag{10.49}$$

Here, ε_i is the dielectric constant of the interfacial layer and δ is the thickness of this interfacial layer (see Figure 10.7b). Equation (10.48) is obtained by assuming that δ is only a few angstroms thick, and hence ε_i is roughly equal to unity. Since c_2 and c_3 are determined experimentally, one can express $q\phi_0$ and D_s in terms of these two quantities,

$$q\phi_0 = E_g - q\left(\frac{c_2\chi_s + c_3 + \Delta\phi}{1 - c_2}\right), \tag{10.50}$$

and the interface state density D_s is given by

$$D_s = \frac{(1 - c_2)\varepsilon_i\varepsilon_0}{c_2\delta q^2}. \tag{10.51}$$

It is noted that the value of $q\phi_0$ for a wide variety of metals on III-V compound semiconductor Schottky contacts was found to be about one-third of the band gap energy above the valence band edge. Therefore, the barrier height for a Schottky diode formed on n-type semiconductors with very high surface state density is roughly equal to two-thirds of the band gap energy (i.e., $q\phi_{\mathrm{Bn}} \approx (2/3)E_{\mathrm{g}}$). Measurements of barrier heights for many metal/III-V semiconductor Schottky diodes with high surface state density are found to be in good agreement with this prediction. Theoretical calculations and experimental data reveal that the surface state densities for many III-V semiconductors such as GaAs and GaN are indeed very high (e.g., $Q_{\mathrm{ss}} \geq 10^{13}$ states/cm^2) and the barrier height was found to be independent of the metal work function for Schottky diodes formed on these semiconductor materials.

If the interfacial layer is assumed only a few angstroms thick and $\varepsilon_{\mathrm{i}} = 1$, then the interface state density given by (10.51) is reduced to

$$D_{\mathrm{s}} \approx 1.1 \times 10^{13}(1 - c_2)/c_2 \quad \text{states/(cm}^2 \cdot \text{eV)}. \tag{10.52}$$

For infinite interface state density, $D_{\mathrm{s}} \to \infty$ and $c_2 \to 0$, (10.48) becomes

$$\phi_{\mathrm{Bn}} = E_g/q - \phi_0 - \Delta\phi \approx \frac{2}{3}(E_g/q). \tag{10.53}$$

If the interface state density is negligible and only the image-lowering effect is considered, then the barrier height is given by

$$\phi_{\mathrm{Bn}} = (\phi_{\mathrm{m}} - \chi_{\mathrm{s}}) - \Delta\phi, \tag{10.54}$$

which reduces to the ideal Schottky barrier height given by (10.13) when the image-lowering effect is neglected.

Experimental results reveal that values of c_2 for Si, GaAs, and GaP are equal to 0.27, 0.09, and 0.27 eV, respectively. The calculated values of $q\phi_0$ for Si, GaAs, and GaP are found to be 0.30, 0.54, and 0.67 eV, respectively. Figure 10.17 shows the experimental results of the barrier height versus metal work function for n-type Si, GaAs, GaP, and CdS Schottky contacts.[5] The results clearly show that for GaAs Schottky contacts, the barrier height is nearly independent of the metal work function. This is due to the very high interface state density for GaAs crystal, and the barrier height is determined by (10.53).

The three most commonly used methods for determining the barrier height of a Schottky barrier diode are discussed next. These are (i) current–voltage (*I–V*), (ii) capacitance–voltage (*C–V*), and (iii) photoemission (*I–E*) methods, which are now discussed.

(i) The current–voltage method. A semiempirical formula for the current density of a practical Schottky barrier diode can be expressed as

$$J = C \exp\left(-\frac{q\phi_{\mathrm{Bn}}}{k_{\mathrm{B}}T}\right)\left[\exp\left(\frac{qV_{\mathrm{a}}}{nk_{\mathrm{B}}T}\right) - 1\right], \tag{10.55}$$

where C is a preexponential factor; its value depends on the model employed (i.e., thermionic emission or diffusion model). Typical plots of $\ln(J)$ versus applied

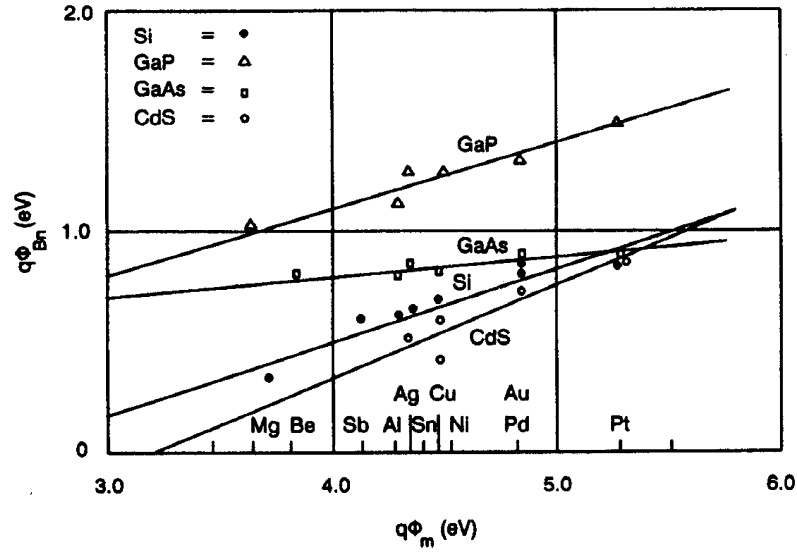


FIGURE 10.17. Schottky barrier heights versus metal work function for metal–Si GaAs, GaP, and CdS Schottky contacts. After Cowley and Sze,[5] by permission.

voltage V_a and inverse temperature $1/T$ for a Schottky barrier diode are shown in Figures 10.18a and b, respectively. The barrier height can be determined either from the saturation current density J_0, as shown in Figure 10.18a, or from the $\ln(J)$ versus $1/T$ plot at a fixed-bias voltage, as shown in Figure 10.18b. To increase the accuracy of the barrier height determined from the $\ln(J_F/T^2)$ versus $1/T$ plot at a fixed forward-bias voltage, it is important to choose a bias voltage in which the diode ideality factor is nearly equal at different temperatures (i.e., the slope of

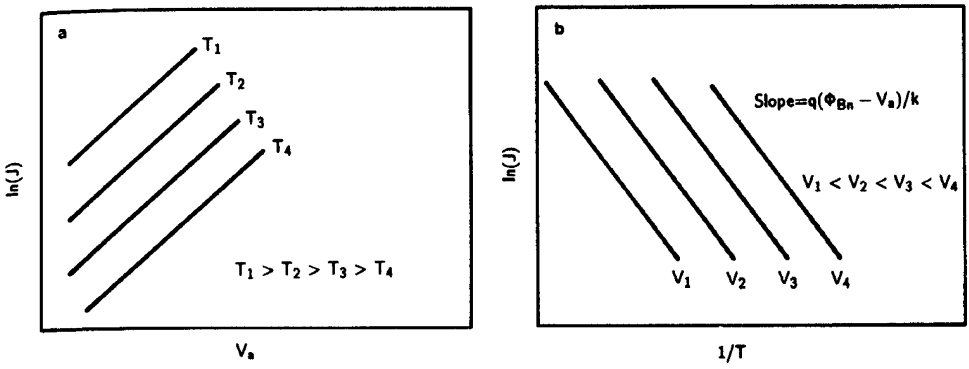


FIGURE 10.18. (a) $\ln J$ versus V_a for a Schottky barrier diode at four different temperatures; (b) $\ln J$ versus $1/T$ for four different forward-bias voltages.

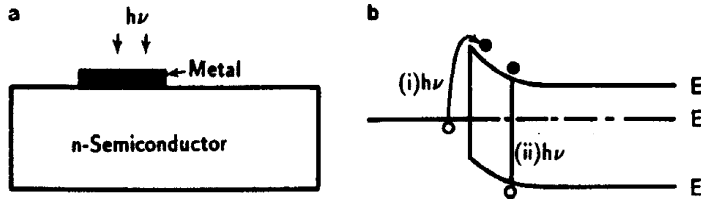


FIGURE 10.19. (a) Illumination from the top surface of a Schottky photodiode. (b) Photoexcitation of electrons: (i) from the metal into the semiconductor with $q\phi_{Bn} < h\nu < E_g$, (ii) inside the semiconductor with $h\nu > E_g$.

$\ln(J_F)$ vs. V_F plots at T_1, T_2, T_3, and T_4 should be identical), as shown in Figure 10.18a.

(ii) The photoemission method. The barrier height of a Schottky diode can be determined by measuring the photocurrent versus wavelength of the incident photons near the fundamental absorption edge, as illustrated in Figure 10.19a. When photons with energies falling between the barrier height and the band gap energy of the semiconductor (i.e., $q\phi_{Bn} < h\nu < E_g$) impinge on the Schottky contact, electrons are excited from the metal and injected into the semiconductor. This is illustrated in process (i) of Figure 10.19b. If the energy of the incident photons exceeds the band gap energy of the semiconductor (i.e., $h\nu > E_g$), then direct band-to-band excitation occurs and electron–hole pairs are generated in the semiconductor. This is illustrated in process (ii) of Figure 10.19b. When process (ii) becomes dominant, a sharp increase in photoresponse near the absorption edge is observed, and intrinsic photoconduction becomes the dominant process.

In the photoemission method, the energies of the incident photons are limited between the barrier height and the band gap energy of the semiconductor. According to Fowler's theory, the photocurrent of a Schottky barrier diode produced by photogenerated electrons in the metal [i.e., process (i)] is given by[6]

$$I_{ph} = C(h\nu - q\phi_{Bn})^2, \tag{10.56}$$

which is valid for $(h\nu - q\phi_{Bn}) \geq 3k_BT$ and $q\phi_{Bn} < h\nu < E_g$. From (10.56), it is seen that the photocurrent is directly proportional to the square of the photon energy. Therefore, a plot of the square root of the photocurrent versus photon energy should yield a straight line. Extrapolation of this straight line to the intercept of the horizontal axis yields the barrier height ϕ_{Bn}. Figure 10.20 shows values of the barrier heights determined by the photoemission method for a W/Si and a W/GaAs Schottky diode.

(iii) The capacitance–voltage method. Another method of determining the barrier height of a Schottky diode uses the capacitance versus voltage (C–V) measurement. From (10.27), it is seen that for a uniformly doped semiconductor, a plot of C_d^{-2} versus V should yield a straight line, and its intercept with the voltage axis is equal to the diffusion potential V_D. This is illustrated in Figure 10.21 for a W/n-Si and

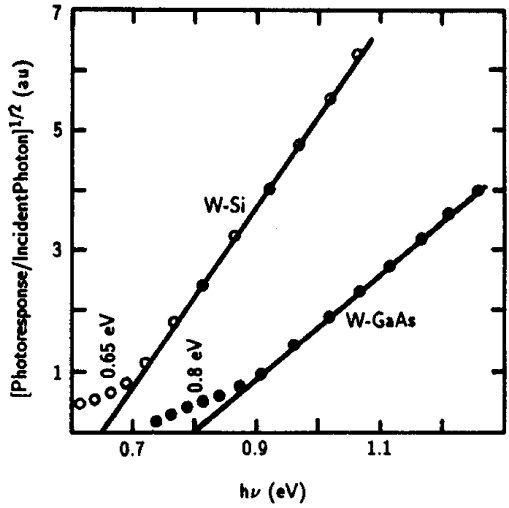


FIGURE 10.20. Square root of photocurrent (in arbitrary units) versus photon energy for a W–Si and a W–GaAs Schottky diode. The intercept of the curves with the horizontal axis yields the barrier height. After Crowell et al.,[7] by permission.

a W/n-GaAs Schottky barrier diode.[7] The diffusion potential determined by the C–V measurements is directly related to the barrier height by the expression

$$\phi_{\mathrm{Bn}} = V_{\mathrm{D}} + V_{\mathrm{n}} - \Delta\phi + k_{\mathrm{B}}T/q, \tag{10.57}$$

where

$$\Delta\phi = \sqrt{\frac{q\mathcal{E}_{\mathrm{m}}}{4\pi\varepsilon_0\varepsilon_{\mathrm{s}}}}, \tag{10.58}$$

and

$$V_{\mathrm{n}} = \left(\frac{k_{\mathrm{B}}T}{q}\right)\ln\left(\frac{N_{\mathrm{c}}}{N_{\mathrm{D}}}\right). \tag{10.59}$$

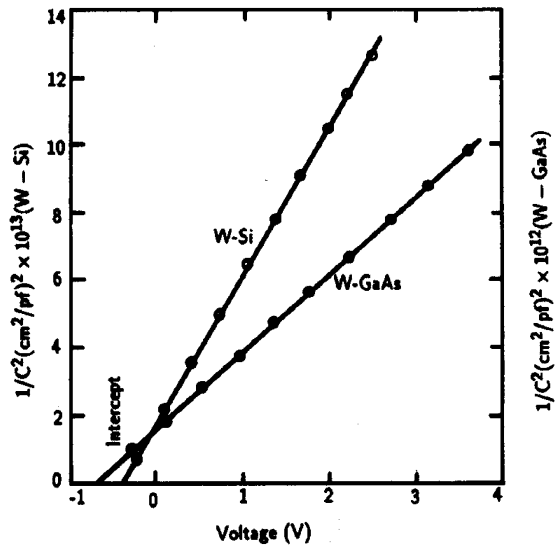


FIGURE 10.21. $1/C^2$ versus applied voltage for a W–Si and a W–GaAs Schottky diode. After Crowell et al.,[7] by permission.

TABLE 10.2. Barrier heights, ϕ_{Bn} (eV), at 300 K for some metal/n-type semiconductor Schottky diodes.

Metal	Si	Ge	GaAs	GaP	InP	GaN	InSb
Al	0.55–0.77	0.48	0.80	1.05	–	0.6, 0.8	–
Ag	0.56–0.79	–	0.88	1.20	0.54	–	0.18
Au	0.76–0.81	0.45	0.90	1.30	0.49	0.94	0.17
Cu	0.69–0.79	0.48	0.82	1.20	–	–	–
Mo	0.68	–	–	–	–	–	–
Ni	0.67–0.70	–	–	1.27	–	1.13	–
Pd	0.71	–	–	–	–	0.93	–
Pt	0.90	–	0.86	1.45	–	0.8–1.60	–
PtSi	0.85	–	–	–	–	–	–
W	0.66	0.48	0.71–0.80	–	–	–	–
Ti	0.60	–	0.82	–	–	1.12	–
Ti/Ai	–	–	–	–	–	Ohmic	–

is the depth of the Fermi level below the condution band edge. Thus, knowing V_D, V_n, and $\Delta\phi$, the barrier height ϕ_{Bn} can be determined from the C–V measurement. Table 10.2 lists values of the barrier heights for some metal/n-type semiconductor Schottky contacts determined using the three methods described above. Table 10.3 lists values of the barrier heights for some metal/p-type semiconductor Schottky contacts.

The Schottky barrier energy (ϕ_B) for Al, Ni, Pd, Co, Au, and Ag contacts on chemically etched ⟨100⟩ surfaces of both n-and p-type InP was measured and the metallurgical behavior of the contact structures was studied using Auger-electron spectroscopy (AES) by E. Hokelek and G. Y. Robinson.[8] In their study, they found two distinct Fermi-level pinning positions located at $E_f = E_c - 0.50$ eV and $E_f = E_c - 0.40$ eV and correlated them to the metallurgical state of the contact structures. Their findings strongly suggested that the Schottky barrier formation on InP is controlled by the chemical reaction between the contact metal and the InP substrate, with the degree of chemical reactivity appearing to determine the Fermi-level pinning position at the interface. No simple linear relationship could be found between the measured Schottky barrier height on InP and the work function or the electron negativities of the contact metals. Thus, the results could not be explained in terms of the traditional Schottky[9] and Bardeen theories.[10] Table 10.4

TABLE 10.3. Barrier heights, ϕ_{Bp} (eV), at 300 K for some metal/p-type semiconductor Schottky barrier diodes.

Metal	Si	Ge	GaAs	InP	GaN
Al	0.58	0.48	–	0.92	Ohmic
Ag	0.54	0.50	0.63	0.81	–
Au	0.34	0.30	0.42	0.81	0.57
Ti	0.61	0.48	–	–	0.65
Hf	0.54	–	0.68	–	–
Ni	0.51	–	–	0.90	0.50
Pt	0.20	–	–	–	0.50

TABLE 10.4. Schottky barrier heights for metal/p-type InP Schottky diodes calculated from the photoemission, *C–V*, and *I–V* measurements.[8]

Metal	$\phi_{Bp}(I-E)$ (meV)	$\phi_{Bp}(I-V)$ (meV)	$\phi_{Bp}(C-V)$ (meV)	$\Delta\phi$ (meV)	ϕ_m (eV)	χ(eV)
Al	915 ± 25	889 ± 11	1118 ± 68	229 ± 79	4.17	1.5
Ni	895 ± 5	897 ± 5	1140 ± 10	243 ± 15	5.10	1.8
Co	780 ± 0	803 ± 8	865 ± 23	62 ± 31	4.97	1.8
Ag	810 ± 10	788 ± 27	862 ± 4	74 ± 31	4.41	1.9
Au	810 ± 10	794 ± 4	930 ± 10	136 ± 14	5.10	2.4
Pd	810 ± 15	823 ± 3	895 ± 16	72 ± 19	5.17	2.2

lists the Schottky barrier heights for several metal/p-type InP Schottky barrier diodes determined using the photoemission, *C–V*, and *I–V* measurements.

Recent study of Schottky barrier contacts on n- and p-type GaN have been reported by Rickert et al.[11] using X-ray photoemission techniques to determine the barrier heights of Au, Al, Ni, Ti, Pt, and Pd on n- and p-type GaN Schottky contacts. Two different behaviors were observed for the six metals studied. For Au, Ti, and Pt, the surface Fermi-level position lies about 0.5 eV higher in the band gap for n-type than for the p-type GaN. For Ni, Al, and Pd, the surface Fermi-level position is independent of doping, but varies from one metal to the other. Results for Ni, Pd, and Al fit a modified Schottky–Mott theory, while Au, Ti, and Pt show a more complex behavior. Table 10.5 lists the Schottky barrier heights for six metals with contact to n- and p-type GaN determined using the X-ray photoemission method.[11]

It should be noted that for metals (e.g., Ni, Al, and Pd) that exhibit a single Fermi-level pinning position, the sum of ϕ_{Bn} and ϕ_{Bp} values shown in Table 10.5 is very close to the GaN band gap ($E_g = 3.4$ eV). In this respect, Ni, Al, and Pd could be considered to follow the Schottky–Mott theory.

10.8. Enhancement of Effective Barrier Height

As discussed in the previous section, the barrier height of an ideal metal–semiconductor Schottky diode is equal to the difference between the metal work

TABLE 10.5. The schottky barrier heights and Fermi-level positions for metals on n- and p-type GaN as determined by the X-ray photoemission method.[11]

Metal	ϕ_m(eV)	χ_m(eV)	n-GaN: position of E_F (above E_v)	Barrier height (eV)	
				ϕ_{Bn}	ϕ_{Bp}
Al	4.28	1.61	2.6 ± 01	0.8 ± 01	2.5 ± 01
Ti	4.33	1.54	2.8 ± 01	0.6 ± 01	2.3 ± 01
Au	5.10	2.54	2.5 ± 01	0.9 ± 01	1.9 ± 01
Pd	5.12	2.20	1.5 ± 01	1.9 ± 01	1.5 ± 01
Ni	5.15	1.91	2.0 ± 01	1.4 ± 01	1.9 ± 01
Pt	5.65	2.28	1.8 ± 01	1.6 ± 01	1.4 ± 01

function and the electron affinity of the semiconductor. In reality, however, the surface state density of a semiconductor plays an important role in determining the effective barrier height of a Schottky diode. Since only a limited number of metals are suitable for forming Schottky contacts on the semiconductor, it is important to explore alternative methods for enhancing the effective barrier height of a Schottky diode.

It is noted that the effective barrier height of a Schottky diode can be strongly affected by the electric field distribution near the metal–semiconductor interface. Therefore, the barrier height of a Schottky contact can be modified by altering the built-in electric field distribution (e.g., through creating a concentration gradient near the semiconductor surface) in a thin region below the metal–semiconductor interface. Evidence of such dependence has indeed been observed in various Schottky barrier contacts. In fact, the barrier height will decrease if a heavily doped n^+ or p^+ layer is grown on the n or p semiconductor to form a metal-n^+/n or metal-p^+/p structure. This technique is widely used for making good ohmic contacts in various semiconductor devices and integrated circuits. On the other hand, if a thin surface layer of opposite dopant to the substrate is deposited onto it to form a metal-p^+/n or metal–n^+/p structure, the effective barrier height can be significantly enhanced by using such a structure.

In this section, three different barrier height enhancement methods are described. In the first approach, the effective barrier height of a Schottky diode can be enhanced by depositing a very thin epilayer of opposite dopant on the semiconductor substrate. In such a structure, the barrier height of a metal/p^+-n or a metal/n^+-p Schottky barrier contact is controlled by the thickness and dopant density of the thin epilayer grown on top of the semiconductor substrate. This thin surface layer can be deposited using low-energy ion implantation, molecular beam epitaxy (MBE), or a metal-organic chemical vapor deposition (MOCVD) technique. Theoretical and experimental results for the metal/p^+-n and metal/n^+-p silicon Schottky barrier diodes are discussed next.

Figure 10.22a shows the cross-sectional view and Figure 10.22b the energy band diagram of a metal/p^+-n GaAs Schottky barrier diode. It is noted that the p^+-n junction shown in Figure 10.22a is an abrupt junction structure, and the thickness (W_p) of the p region is treated as an adjustable parameter. As long as this p layer remains very thin, the entire p layer will be fully depleted even at zero-bias conditions. The potential distribution in such a structure can be evaluated using the depletion approximation. However, if the p layer becomes too thick, then it will be partially depleted. As a result, a quasineutral p region will exist, and the structure becomes a conventional metal/p-type Schottky barrier diode in series with a p-n junction diode. Such a structure will be avoided in the present analysis. Therefore, it is important to keep in mind that the metal/n^+-p or metal/p^+-n structure will work as a Schottky diode only if the thin n^+ or p^+ surface layer remains fully depleted.

To analyze the barrier height enhancement in a metal/p^+-n or metal/n^+-p Schottky barrier structure, the abrupt junction approximation will be used. The basic device parameters are defined as follows: χ_s is the electron affinity of the

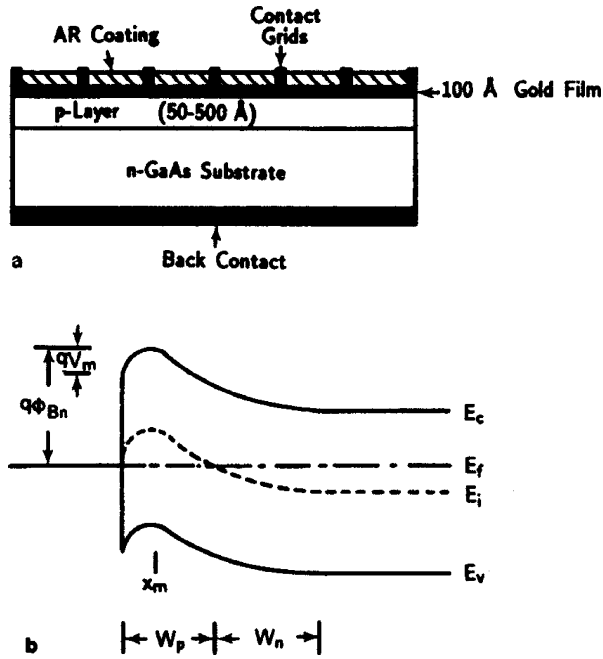


FIGURE 10.22. Schematic diagram of a metal/p-n Schottky barrier diode: (a) cross-sectional view, and (b) energy band diagram showing barrier height enhancement of qV_m.

semiconductor; ϕ_m, ϕ_p, and ϕ_n are the work functions of the metal, p-semiconductor, and n-semiconductor, respectively. If a voltage V is applied to the metal contact, then a potential maximum V_m will appear in front of the metal contact. In this case, $\phi'_{Bn} = \phi_m - \chi_s + V_m(V)$ is the barrier height seen by electrons in the metal, and $\phi'_{Bn} = \phi_m - \phi_n - V + V_m(V)$ is the barrier height seen by electrons on the n-type semiconductor side. If W_n is the width of the space-charge region that extends into the n region and $x = 0$ at the metal contact, then using the depletion approximation, Poisson's equation can be written as

$$\frac{d^2V(x)}{dx^2} = \begin{cases} \dfrac{qN_a}{\varepsilon_0\varepsilon_s} & \text{for } 0 < x < W_p, \qquad (10.60) \\ -\dfrac{qN_d}{\varepsilon_0\varepsilon_s} & \text{for } W_p < x < W_p + W_n, \qquad (10.61) \end{cases}$$

with boundary conditions given by

$$V(x) = \begin{cases} 0 & \text{at } x = 0, \\ V(x) = \phi_m - \phi_n + V_n & \text{at } x = W_p + W_n. \end{cases} \qquad (10.62)$$

It is noted that $V(x)$ and $\frac{dV(x)}{dx}$ are continuous at $x = W_p$, and

$$\left.\frac{dV(x)}{dx}\right|_{x=0} = \frac{q(N_d W_n - N_a W_p)}{\varepsilon_0 \varepsilon_s},$$
$$\left.\frac{dV(x)}{dx}\right|_{x=W_p+W_n} = 0. \tag{10.63}$$

It can be shown that the solution of $V(x)$ for $0 \le x \le W_p$ is given by

$$V(x) = V_1(x) = \left(\frac{qN_a}{\varepsilon_0 \varepsilon_s}\right)\left(\frac{x^2}{2} - xW_p\right) + \left(\frac{qN_d}{\varepsilon_0 \varepsilon_s}\right) W_n x. \tag{10.64}$$

And the solution of $V(x)$ for $W_p < x < W_p + W_n$ is given by

$$V(x) = V_2(x) = -\left(\frac{qN_d}{\varepsilon_0 \varepsilon_s}\right)\left[\frac{x^2}{2} - x(W_p + W_n)\right] - \frac{q(N_d + N_a)W_p^2}{\varepsilon_0 \varepsilon_s}. \tag{10.65}$$

The width of the n region can be obtained using the second boundary condition (10.62), namely,

$$\phi_m - \phi_n + V_n = \frac{1}{2}\left[\frac{qN_d(W_n + W_p)^2}{\varepsilon_0 \varepsilon_s}\right] - \frac{1}{2}\left[\frac{q(N_d + N_a)W_p^2}{\varepsilon_0 \varepsilon_s}\right]. \tag{10.66}$$

If $N_d W_n \ll N_a W_p$, then a potential maximum exists inside the space-charge region of the semiconductor and in front of the metal contact. The position of this potential maximum, x_m, can be determined by setting $dV(x)/dx = 0$ at $x = x_m$, which yields

$$\frac{qN_a(x_m - W_p)}{\varepsilon_0 \varepsilon_s} + \frac{qN_d W_n}{\varepsilon_0 \varepsilon_s} = 0, \tag{10.67}$$

or

$$x_m = W_p - \left(\frac{N_d}{N_a}\right) W_n. \tag{10.68}$$

Note that V_m can be obtained by substituting x_m given by (10.68) into (10.64), which yields

$$V_m = -\Delta\phi = \left(\frac{q}{2\varepsilon_0 \varepsilon_s N_a}\right)(N_a W_p - N_d W_n)^2. \tag{10.69}$$

Therefore, the effective barrier height for the metal/p^+-n Schottky barrier diode shown in Figure 10.22a is given by

$$\phi'_{Bn} = \phi_m - \chi_s + V_m, \tag{10.70}$$

where V_m is given by (10.69), and W_p is the thickness of the p layer. The depletion layer width in the n region, W_n, can be calculated using the expression

$$W_n = -W_p + (W_p^2 + C)^{1/2}, \tag{10.71}$$

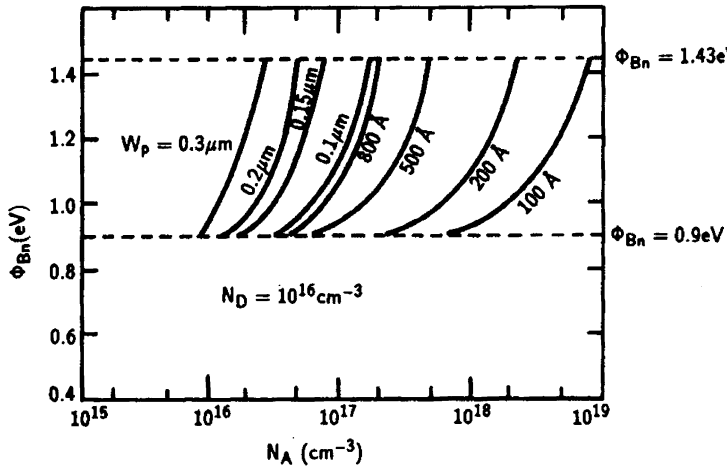


FIGURE 10.23. Calculated barrier heights of a Au/p-n GaAs Schottky barrier diode versus dopant density, N_A, of the p-layer for different p-layer thicknesses. Note: N_d of the n-substrate is fixed at 10^{16} cm^{-3}. After Li,[12] by permission.

where

$$C = \left(\frac{N_a}{N_d}\right) W_p^2 + \frac{2\varepsilon_0\varepsilon_s(\phi_m - \phi_n)}{qN_d} \tag{10.72}$$

and

$$\phi_n = \chi_s + V_n = \chi_s + \left(\frac{k_B T}{q}\right) \ln\left(\frac{N_c}{N_d}\right). \tag{10.73}$$

From the results discussed above, the barrier height enhancement for a metal/p^+-n Schottky barrier diode can be calculated using (10.69) through (10.73). Figures 10.23 and 10.24 show the theoretical calculations of barrier height enhancement versus dopant density of the p^+ and n^+ layers for a Au/p^+-n GaAs Schottky barrier diode with p-layer thickness as parameter.[12] Figure 10.25a shows a plot of the effective barrier height for a Ti/n^+-p silicon Schottky barrier diode, and Figure 10.25b shows the forward *I–V* characteristics for several Ti/n^+-p silicon Schottky diodes with different phosphorus implant doses in the n^+-implanted layer.[13] The barrier height was found to increase from $\phi_{Bo} = 0.60$ eV for a conventional Ti/p-silicon Schottky diode to $\phi_{Bp} = 0.93$ eV for a Ti/n^+-p silicon Schottky diode fabricated using a phosphorus implant dose of 1.2×10^{12} cm^{-2} on the p-type silicon substrate. The results show that a significant (> 50%) increase in barrier height was obtained using this approach. In principle, an effective barrier height equal to the band gap energy of the semiconductor can be achieved using the structure described in this section, provided that the thickness and dopant density of the thin surface layer are properly chosen for such a Schottky barrier structure. For silicon Schottky barrier diodes, incorporation of such a thin surface layer can be achieved using epitaxial growth or an ion implantation technique, while for

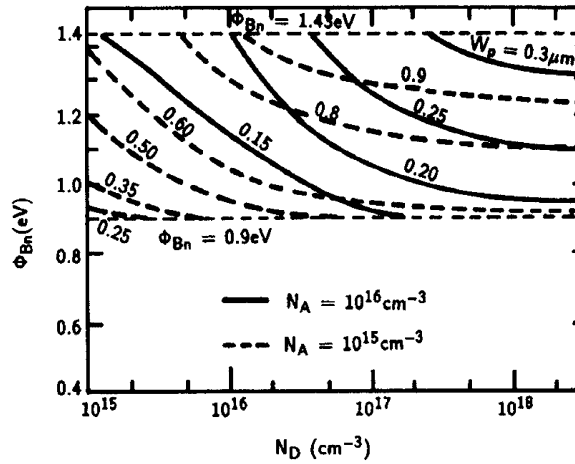


FIGURE 10.24. Calculated barrier heights of a Au/p-n Schottky barrier diode versus dopant density, N_d, of the n-substrate for different p-layer thicknesses and for $N_A = 5 \times 10^{16}$ cm^{-3}. After Li,[12] by permission.

Schottky barrier diodes formed on III-V compound semiconductors the thin epi-layer can be deposited using either molecular beam epitaxy (MBE), atomic layer epitaxy (ALE), or the metal-organic chemical vapor deposition (MOCVD) growth technique. Layer thickness from a few tens of Å to a few hundreds or thousands of Å can be readily deposited onto the GaAs or InP substrates using either the MBE

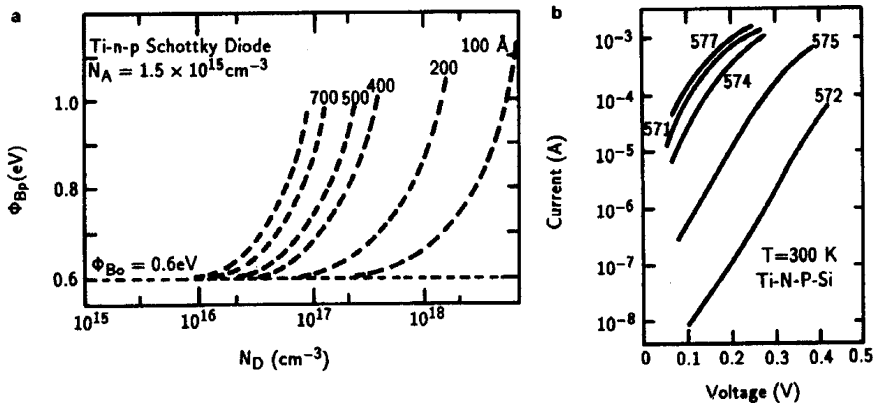


FIGURE 10.25. (a) Calculated barrier heights versus density of the phosphorous implanted layer for different layer thicknesses of a Ti/n-p silicon Schottky barrier diode. Solid dots denote experimental data. (b) Forward-biased *I–V* curves for a controlled Ti/p-silicon Schottky diode (577) and four other Ti/n-p silicon Schottky barrier diodes with different implant doses. After Li et al.,[13] by permission, © IEEE–1980.

or MOCVD growth technique. Therefore, the metal/p^+-n or metal/n^+-p Schottky barrier structure described in this section can be considered as a viable approach for enhancing the effective barrier height of a conventional metal–semiconductor Schottky diode.

Another barrier height enhancement technique using the band gap engineering approach on a small-band-gap semiconductor such as $In_{0.53}Ga_{0.47}As$ has been reported recently.[14] The technique involves the growth of n periods of thin graded superlattices consisting of a larger-band-gap material and a smaller-band-gap epilayer of variable thickness to create a larger-band-gap surface layer so that the effective barrier height can be enhanced in such a Schottky barrier diode. For example, a high-quality $In_{0.53}Ga_{0.47}As$ Schottky barrier diode has been fabricated using a novel graded superlattice structure consisting of 10 periods of n-$In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ graded superlattice deposited on top of the n-type $In_{0.53}Ga_{0.47}As$ epilayer grown by the MBE technique on the InP substrate. The result shows a barrier height enhancement of 0.41 eV (i.e., from $\phi_{B0} = 0.3$ eV to $\phi_{Bn} = 0.71$ eV), and near-ideal I–V and C–V characteristics are obtained for this novel Schottky diode. Figure 10.26a shows the energy band diagram of this n-$In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ graded superlattice structure formed on an InGaAs Schottky diode, and Figure 10.26b shows the dimensions of this graded InAlAs/InGaAs superlattice layer structure and doping densities in each region. The composition of InAlAs/InGaAs and total thickness of each period remain the same in the superlattice layer. The graded composition is achieved by changing the thickness ratio of the InAlAs/InGaAs superlattice in each period (i.e., each period is 60 Å and the thickness ratio of InAlAs/InGaAs superlattice varied from 55/5, 50/10, . . . ,30/30, . . . ,5/55 Å from the top to the bottom layers). Figure 10.26c shows the reverse leakage current for a Schottky barrier diode formed on the graded superlattice structure shown in Figure 10.26b. It is noted that very low leakage current was obtained in this InAlAs/InGaAs superlattice Schottky diode.

The third method of enhancing the effective barrier height in a Schottky barrier diode is using the metal–insulator–semiconductor (MIS) structure. In this structure, a very thin insulating layer with thickness of 1 to 3 nm is inserted between the metal and the semiconductor Schottky contact, which results in an MIS Schottky barrier structure (see Figures 12.10a and b). The MIS structure can increase the effective barrier height by $\Delta\phi_B = \delta\chi^{1/2}$, where δ is the thickness of the insulating layer and χ is the mean incremental barrier height. In an MIS Schottky diode, the current conduction is due to the majority carriers tunneling through the thin insulating layer. This tunneling current can be described by

$$J_t = A^*T^2 \exp(-q\phi_{Bn}/k_BT) \exp(-\delta\chi^{1/2}) \exp(qV_a/nk_BT). \tag{10.74}$$

In (10.74), it is noted that the dark current of an MIS Schottky diode can be reduced sharply by the incorporation of a thin insulating layer between the metal/semiconductor contact. As a result, the MIS structure has been widely used in the fabrication of Schottky barrier solar cells to increase the open-circuit voltage and the conversion efficiency. This will be discussed further in Section 12.2.4.

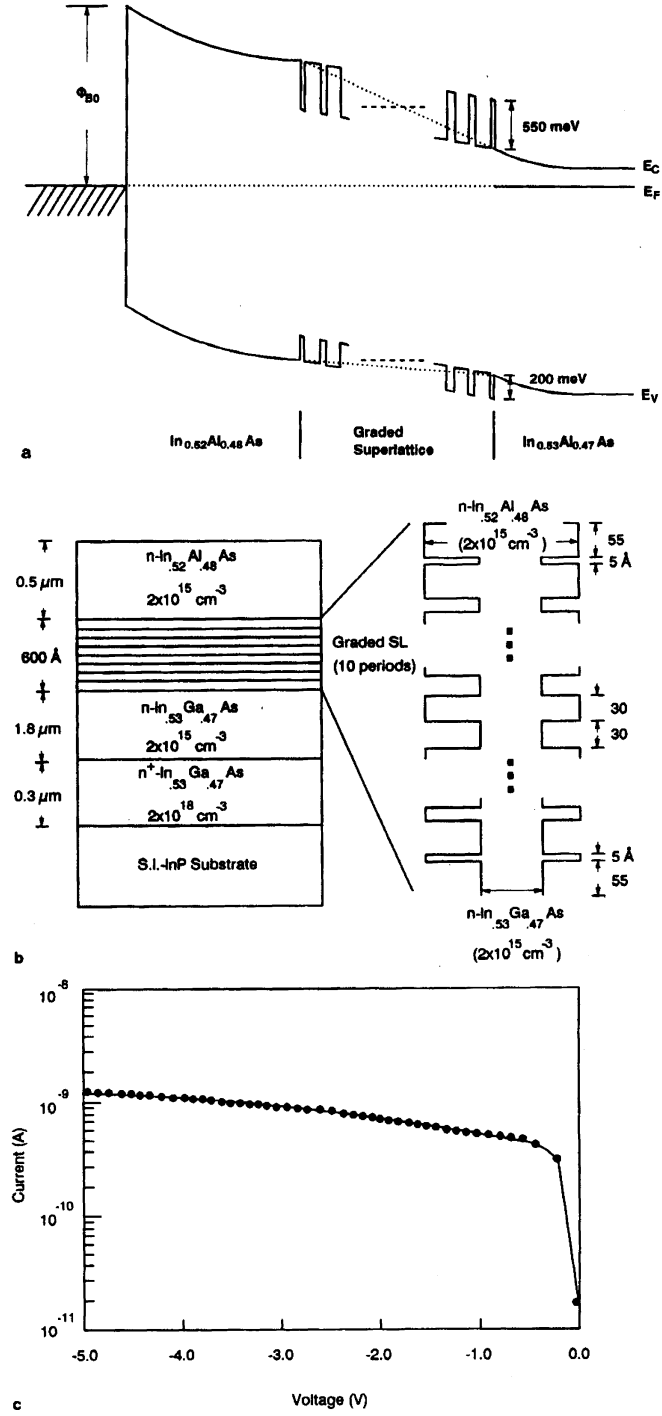


FIGURE 10.26. (a) Energy diagram of a Au/n-InAlAs/n-InGaAs Schottky barrier diode with a 600 Å graded superlattice of InAlAs/InGaAs for barrier height enhancement, (b) dimensions and dopant densities of the structure shown, and (c) the reverse *I–V* characteristics for this Schottky barrier diode. After Lee et al.,[14] by permission.

10.9. Applications of Schottky Diodes

Schottky diodes have been widely used for a wide variety of applications such as solar cells, photodetectors, Schottky-clamped transistors, metal gate field-effect transistors (MESFETs), modulation-doped field-effect transistors (MODFETS or HEMTs), microwave mixers, RF attenuators, rectifiers, varactors, Zener diodes, Schottky transistor logic (STL) gate arrays, and various integrated circuits. For example, the exact logarithmic relationship displayed by the *I–V* curve of a Schottky diode under forward-bias conditions over several decades of current change enables it to be used in logarithmic converter circuits. A metal–semiconductor Schottky diode can also be used as a variable capacitor in parametric circuits for frequency multiplication. The Schottky barrier solar cell has the potential for use as a low-cost photovoltaic power conversion device for large-scale terrestrial power generation. High-speed Schottky barrier photodiodes covering a broad wavelength range from ultraviolet to visible and into the mid-IR spectral regime have been reported using different metal–semiconductor contacts. In this section, some practical applications of Schottky barrier diodes are described.

10.9.1. Photodetectors and Solar Cells

A Schottky barrier diode can be used as a high-speed photodetector for low-level light detection or as a solar cell for conversion of solar energy into electricity. To reduce absorption loss in the metal contact of a Schottky barrier photodiode, it is a common practice to use either a thin metal film (100 Å or less) or a grating-type (metal grids) structure for the Schottky contact. The reflection loss on a semiconductor surface is minimized by using an antireflection (AR) coating an the front side of a Schottky barrier photodiode, as illustrated in Figure 10.27a. For a grating-type Schottky barrier photodiode shown in Figure 10.27b, the pattern of metal-grating structure for the Schottky contact can be defined and produced using the photolithography technique. Selection of the metal-grid spacing is determined by the operating bias voltage and the substrate doping concentration of the diode to ensure that spacing between the metal grids is fully depleted under operating conditions. For example, in the case of a Au/n-Si Schottky diode with a doping density of $N_D = 10^{14}$ cm^{-3}, a spacing of around 10 μm between the metal grids is adequate for creating a fully depleted region between the metal grids of such a Schottky contact. A photodetector using a semitransparent Schottky contact or a grating-type Schottky contact structure has shown excellent quantum efficiency and high responsivity.

In general, there are three detection modes that are commonly used in a Schottky barrier photodiode; these are illustrated in Figures 10.28a, b, and c. The operation of each of these detection modes depends greatly on the incident photon energies, the applied bias voltage, and the breakdown voltage of the photodiode. These are discussed as follows.

(i) $q\phi_{Bn} < h\nu < E_g$ *and* $V_a \ll V_B$. In this detection mode, electrons are excited from the metal and injected into the semiconductor, as illustrated in Figure 10.28a. In this case, the Schottky barrier photodiode may be used for a wide variety of applications, which include (1) IR detector, (2) as a test structure to determine the barrier height by the photoemission (*I–E*) technique, and (3) as a test device for studying the bulk defects and interface states in a semiconductor, and hot electron transport in a metal film. The reason a Schottky diode can be used for long-wavelength infrared (LWIR) detection is that the barrier height for most Schottky diodes is smaller than the band gap energy of the semiconductor. As a result, photons with energy equal to the barrier height absorbed inside the metal film of a Schottky diode usually fall in the infrared regime. Since the barrier height for an IR Schottky barrier photodiode is usually small, the reverse leakage current in such a device is expected to be very large at room temperature. Therefore, in order to reduce the reverse leakage current, an IR Schottky barrier photodetector is usually operated at cryogenic temperatures (e.g., $T < 77$ K). For example, PtSi/p-Si Schottky barrier photodiode (with barrier height $\phi_{Bp} = 0.2$ eV) arrays integrated with CCD (charge-coupled device) arrays have been developed for 3- to 5-μm IR image-sensor array applications. Extending the detection wavelength to 10 μm is possible if the operating temperature for the low-barrier ($\approx$ 0.1 eV) Schottky barrier photodiode is lowered to 4.2 K.

(ii) $h\nu \geq E_g$ *and* $V_a \ll V_B$. As shown in Figure 10.28b, in this detection mode, the electron–hole pairs are generated inside the depletion region of the semiconductor, and the Schottky diode is operating as a high-speed photodetector. Since the Schottky diode is a majority carrier device, its response speed is limited mainly by the RC time constant and the carrier transit time across the depletion region of the detector. The grating-type Au/n-Si Schottky barrier photodiode shown in Figure 10.27b has a responsivity of 0.63 A/W at 0.9 μm and a bandwidth of 1 GHz. A Au/n-GaAs Schottky barrier photodiode has achieved a response speed of less than 100 ps at 0.85 μm. In fact, a high-speed GaAs Schottky barrier photodiode with a 3-dB bandwidth greater than 100 GHz has been reported recently. Further description of high-speed photodetectors using Schottky barrier structures will be given in Chapter 12.

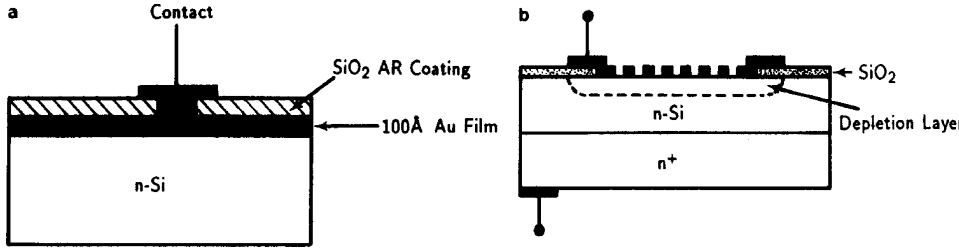


FIGURE 10.27. Schematic diagrams of (a) conventional Schottky barrier photodiode and (b) a grating-type Schottky barrier photodiode.

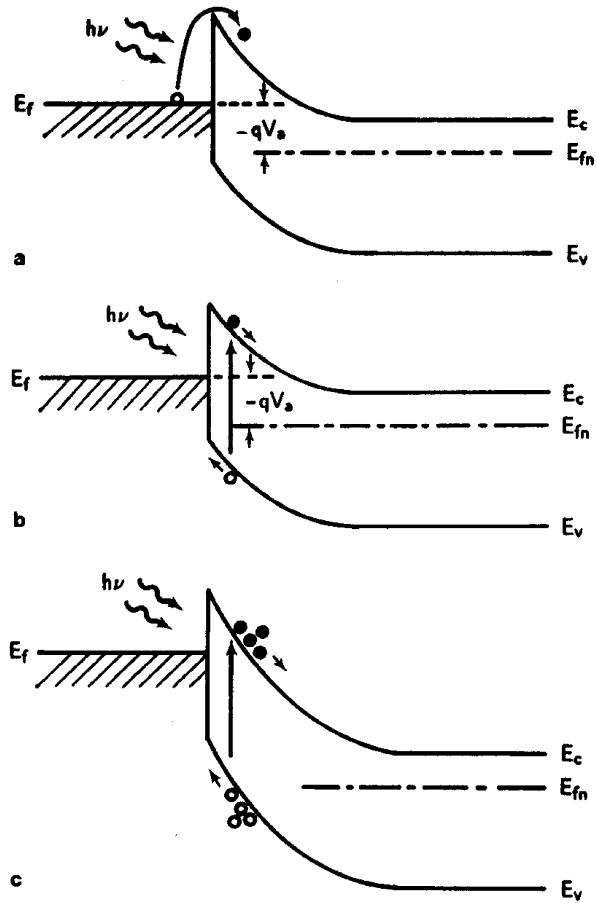


FIGURE 10.28. Different detection modes of a Schottky barrier photodiode: (a) $q\phi_{Bn} \leq h\nu \leq E_g$, (b) $h\nu \geq E_g$, $V_a \ll V_B$, (c) $h\nu \geq E_g$, $V_a \approx V_B$.

(iii) $h\nu \geq E_g$ *and* $V_a = V_B$. In this mode of operation, the Schottky barrier photodiode is in the avalanche mode of detection; this is shown in Figure 10.28c. When a Schottky diode is operating in the avalanche regime ($V_a = V_B$), an internal current gain is obtained. Thus, a Schottky barrier avalanche photodiode (APD) can provide both high-speed and high-sensitivity detection. A diffused guard-ring structure is usually employed in a Schottky barrier APD to eliminate the possible edge breakdown effect.

A Schottky barrier photodiode can also be used as an efficient ultraviolet (UV) photon detector. For example, in the UV regime the absorption coefficient for most semiconductors is greater than 10^5 cm^{-1}, which corresponds to an effective absorption depth of 0.1 μm or less. Thus, by using a thin metal film and an AR coating film simultaneously on the Schottky barrier structure, efficient

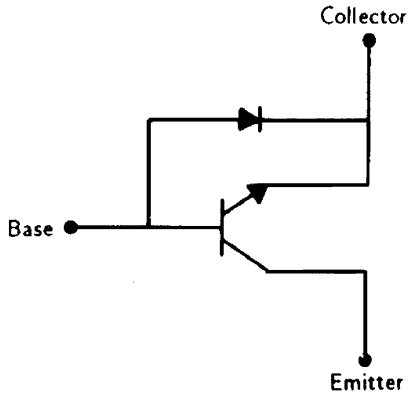


FIGURE 10.29. A Schottky-clamped bipolar transistor. The Schottky diode is connected between the base and collector of a bipolar junction transistor.

collection of photons in the UV spectrum can be achieved. For example, both Ag/ZnS and Au/ZnSe Schottky barrier photodiodes have been developed for UV light detection. More recently, Schottky barrier photodiodes formed on GaN and SiC wide-band-gap materials have shown superior performance characteristics in the UV and solar blind spectral regions. This will be discussed further in Chapter 12.

Finally, Schottky barrier solar cells have also been developed for photovoltaic conversion of sunlight into electricity. The Schottky barrier solar cell is easy to fabricate, and has the potential for low-cost and large-scale production. However, because of the low barrier height, Schottky barrier solar cells in general have lower open-circuit voltage and lower conversion efficiency than most p-n junction solar cells. This will be discussed further in Chapter 12.

10.9.2. Schottky-Clamped Transistors

In the saturated switching process of a conventional n-p-n or p-n-p bipolar junction transistor (BJT), the turnoff speed is limited by the minority carrier storage time in the collector region of the transistor. For a Si switching transistor, the conventional method of reducing the storage time is to shorten the minority carrier lifetime by doping the Si transistors with gold impurity. However, since current gain is also proportional to the minority carrier lifetime, the gold-doped silicon transistor will also have a lower current gain. Therefore, it is generally not desirable to dope silicon transistors with gold impurity. To overcome this problem a Schottky barrier diode is usually connected to the base–collector junction of the transistor. As shown in Figure 10.29, the minority carrier storage problem can be virtually eliminated if a Schottky barrier diode is connected between the base–collector junction of the transistor to form a Schottky-clamped transistor. The switching time constant is drastically shortened in this transistor, since the minority carrier storage in the collector is greatly reduced by the Schottky diode connecting in parallel with the base–collector junction. In the saturation region, the collector junction of

the transistor is slightly forward-biased instead of reverse-biased. If the forward voltage drop in the Schottky diode is much smaller than the base–collector voltage of the transistor, then most of the excess base current will flow through the Schottky diode. Therefore, the minority carriers are not stored in the collector. Furthermore, the saturation time is greatly reduced when compared to a transistor without a Schottky diode connecting to the base–collector junction. A switching time of less than 1 ns for a Schottky-clamped silicon BJT has been reported. Recently, low-power and high-speed Schottky-clamped transistor logic (STL) gate arrays have been developed for computer and other custom IC applications. In an STL gate array, two Schottky barrier diodes with different barrier heights are used. For example, in the Si STL gate array, one Schottky diode (e.g., PtSi/n-Si) with large barrier height is connected to the base–collector junction, while another Schottky diode with low barrier height (e.g., TiW/p-Si) is connected in the collector region of the transistor. The Si STL gate arrays with propagation delay times of less than 1 ns and voltage swings of a few hundred millivolts or less have been achieved. The STL gate arrays fabricated from III-V compound semiconductors such as GaAs, InP, and InGaAs with propagation delay times of a few tens of picoseconds have also been reported.

10.9.3. Microwave Mixers

High-frequency applications of Schottky barrier diodes deal with low-level signal detection and mixing at microwave frequencies. It has been shown that burnout resistance and noise performance of a Schottky diode is usually superior to that of a point-contact mixer diode.

The frequency response of a Schottky diode is generally superior to that of a p-n junction diode, since it is limited by the RC time constant of the Schottky diode rather than by the minority carrier lifetime as in the case of a p-n junction diode. Using an n-n^+ epitaxial wafer for Schottky contact fabrication, both junction capacitance (e.g., $C < 0.1$ pF at $V_a = 0$ and $N_d = 10^{17}$ cm^{-3}) and series resistance of the diode can be reduced to a very low value. This is a direct result of using a very thin n-type epitaxial layer on an n^+-silicon substrate, and the resistance drop across the n^+ region is negligibly small. A point-contact Schottky diode with barrier contact area 5- to 10-μm in diameter has been reported. Figure 10.30 shows the geometry of a microwave Schottky barrier diode mixer using a microstrip line configuration. This Schottky diode is capable of excellent mixer performance, either as a discrete component in a waveguide or as a balanced mixer in a microstrip line at carrier frequencies as high as several tens of GHz.

The RC time constant of a Schottky barrier diode can be calculated using an n-n^+ structure with a barrier contact of radius r, which is large compared to the epilayer thickness. The series resistance of a Schottky diode can be calculated using the expression

$$R_s = \frac{\rho d}{\pi r^2} = \left(\frac{1}{qN_D\mu_n}\right)\left(\frac{d}{\pi r^2}\right). \tag{10.75}$$

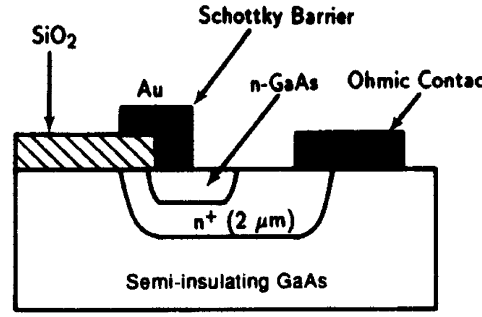


FIGURE 10.30. Cross-sectional view of a Au/n-type GaAs microwave mixer using a Schottky barrier structure.

The depletion layer capacitance can be calculated using the expression

$$C_{\mathrm{d}} = (\pi r^2)\left(\frac{qN_{\mathrm{D}}\varepsilon_0\varepsilon_{\mathrm{s}}}{2V_{\mathrm{D}}}\right)^{1/2}, \tag{10.76}$$

where N_{D} is the doping density of the n-epitaxial layer and V_{D} is the diffusion potential. Thus, the RC time constant for such a Schottky diode is given by

$$R_{\mathrm{s}}C_{\mathrm{d}} = \left(\frac{d}{\mu_{\mathrm{n}}}\right)\left(\frac{\varepsilon_0\varepsilon_{\mathrm{s}}}{2qN_{\mathrm{D}}V_{\mathrm{D}}}\right)^{1/2}. \tag{10.77}$$

To achieve high-frequency performance in a Schottky diode, the RC time constant of the diode must be kept as small as possible. This requires the use of a very thin epitaxial layer with high carrier mobility and doping density. Other important applications of the Schottky barrier structure include metal gate (i.e., Schottky gate) field-effect transistors (MESFETs) and modulation-doped field-effect transistors (MODFETs) formed on III-V compound semiconductors such as GaAs/AlGaAs, InGaAs/AlGaAs, InGaAs/InAlAs, and GaN/InGaN, and the metal–semiconductor IMPATT diodes for microwave power generation.

10.10. Ohmic Contacts in Semiconductors

Formation of good ohmic contact between metal and semiconductor is an extremely important process for fabricating high-performance semiconductor devices and integrated circuits. Table 10.6 gives a list of metals that are used in forming the ohmic contacts on a wide variety of semiconductors. Good ohmic contacts are necessary in order to effectively extract electric current and power from a semiconductor device. In general, an ohmic contact is referred to a noninjecting contact in which the current–voltage (I–V) relationship under both the reverse- and forward-bias conditions is linear and symmetrical. However, in reality, a contact is considered ohmic if the voltage drop across the metal–semiconductor interface is small compared to the voltage drop across the bulk semiconductor.

Ohmic contacts can be characterized in terms of the specific contact resistance R_{c}, which is defined as the reciprocal of the derivative of the current density with

TABLE 10.6. Metals for forming ohmic contacts in various semiconductors.

Semiconductor	Metals
Ge (N)	Ag-Al-Sb, Al, Au, Bi, Ai-Au-P, Sb, Sn, Pb-Sn
Ge (P)	Ag-Al, Au, Cu, Ga, Ga-In, In, Al-Pd, Ni, Pt, Sn
Si (N)	Ag-Al, Al-Au, Au, Ni, Pt, Cu, In, Ge-Sn, Au-Sb, Al-Cu,
Si (P)	Ag, Al, Al-Au, Au, Ni, Pt, Sn, In, Pb, Ga, Ge, Al-Cu
GaAs (N)	Au-Ge (88%, 12%)-Ni, Ag-In (95%, 5%)-Ge, Ag-Sn
GaAs (P)	Au-Zn (84%, 16%), Ag-In-Zn, Ag-Zn
GaP (N)	Ag-Te-Ni, Al, Au-Si, Au-Sn, In-Sn
GaP (P)	Au-In, Au-Zn, Ga, In-Zn, Zn, Ag-Zn
GaAsP (N)	Au-Sn
GaAsP (P)	Au-Zn
GaAlAs (N)	Au-Ge-Ni
GaAlAs (P)	Au-Zn
InAs (N)	Au-Ge, Au-Sn-Ni, Sn
InGaAs (N)	Au-Ge, Ni
InGaAs (P)	Au-Zn, Ni
InP (N)	Au-Ge, In, Ni, Sn
InSb (N)	Au-Sn, Au-In, Ni, Sn
InSb (P)	Au-Ge
CdS (N)	Ag, Al, Au, Au-In, Ga, In, Ga-In
CdTe (N)	In
CdTe (P)	Au, In-Ni, Pt, Rh
ZnSe (N)	In, In-Ga, Pt, In-Hg
SiC (N)	W, Ni
SiC (P)	Al-Si, Si, Ni
GaN (P)	Pd-Ni, Au-Ni
GaN (N)	Ti-Al, Ti-Al-Ni-Au

respect to the applied voltage. An expression for the specific contact resistance evaluated at zero bias is given by

$$R_{\mathrm{c}} = \left(\frac{\mathrm{d}J}{\mathrm{d}V}\right)^{-1}\Bigg|_{V=0}. \tag{10.78}$$

The specific contact resistance defined by (10.78) is an important figure of merit for evaluating ohmic contacts. In general, the current conduction in the ohmic contact region of a metal/moderately doped n-type semiconductor contact is usually dominated by the thermionic emission process. Therefore, the expression of R_{c} can be derived directly from (10.33) and (10.78), which yields

$$R_{\mathrm{c}} = \left(\frac{k_{\mathrm{B}}}{qA^{*}T}\right)\exp\left(\frac{q\phi_{\mathrm{Bn}}}{k_{\mathrm{B}}T}\right). \tag{10.79}$$

Equation (10.79) shows that in order to achieve a small specific contact resistance, the barrier height of the metal–semiconductor contact should be as small as possible. A smaller decrease in the barrier height will result in a very large reduction in the specific contact resistance in the Schottky contact.

For the ohmic contact on a heavily doped semiconductor, the field-emission process (i.e., tunneling) dominates the current transport, and hence the specific contact resistance R_c can be expressed by[15]

$$R_c \approx \exp\left[\left(\frac{2\phi_{Bn}}{\hbar}\right)\sqrt{\frac{\varepsilon_0\varepsilon_s m^*}{N_D}}\right], \tag{10.80}$$

which shows that in the tunneling process, R_c depends strongly on the doping density and varies exponentially with $(\phi_{Bn}/N_D^{1/2})$. The specific contact resistance for n-type GaAs may vary between 10^{-4} and $10^{-7}\ \Omega\cdot\text{cm}^2$, while values of specific contact resistance for GaN ranging from 10^{-4} to $10^{-8}\Omega\cdot\text{cm}^2$ have been reported. Figures 10.31a and b show the energy band diagrams for a low-barrier-height contact and an ohmic contact on a heavily doped n^{++}/n-type semiconductor, respectively. Figure 10.31c shows the *I–V* characteristics of a metal/n^{++}-n-type semiconductor ohmic contact. It is noted that the specific contact resistance calculated from (10.80) agrees well with the measured value for the MBE doped ohmic contacts. However, alloyed contacts usually exhibit a linear dependence of $\ln(R_c)$ on $N_D^{-1/2}$, and the simple formula given by (10.80) cannot adequately explain the *I–V* behavior of the alloyed ohmic contacts.

Formation of ohmic contacts can be achieved in a number of ways. These include (1) choosing a metal with a lower work function than that of an n-type semiconductor (i.e., $\phi_m < \phi_s$) such that the potential barrier between the metal and the semiconductor is small enough for the thermionic emission electrons to tunnel through both directions of the metal–semiconductor contact; (2) deposition of a thin and heavily doped epilayer of the same doping type as the substrate to form an n^{++}/n or p^{++}/p high–low junction structure on the semiconductor surface. This will reduce the barrier width of the metal–semiconductor contact such that current flow can be achieved by quantum-mechanical tunneling through the thin barrier with low contact resistance; (3) using a graded heterojunction approach by using a small-band-gap material for ohmic contact (e.g., form an n^+-InAs/n-GaAs or n^+Ge/n-GaAs heterojunction structure using the MBE technique); (4) using a nonalloyed short-period superlattice (SPS), composed of GaN and narrow-band-gap InN, sandwiched between the GaN channel and the InN cap layer to form ohmic contacts on GaN; and (5) increasing the density of recombination centers at the semiconductor surface (e.g., by surface roughening) so that the surface will serve as an infinite sink for the majority carriers at the contact.

Techniques for forming ohmic contacts on the semiconductor devices include alloying, electroplating, thermal or E-beam evaporation, sputtering, ion implantation, and MBE techniques. In principle, an ohmic contact is formed if the potential barrier between the metal and semiconductor contact is very small. It is seen in Figures 10.7c and f that in order to obtain good ohmic contact on an n-type semiconductor, the metal work function should be less than that of the n-type semiconductor. Conversely, to form ohmic contact on a p-type semiconductor, the metal work function must be larger than that of the p-type semiconductor. Unfortunately, for covalent semiconductors such as Ge, Si, and GaAs, formation of ohmic contacts does not always follow the simple rule cited above. For example,

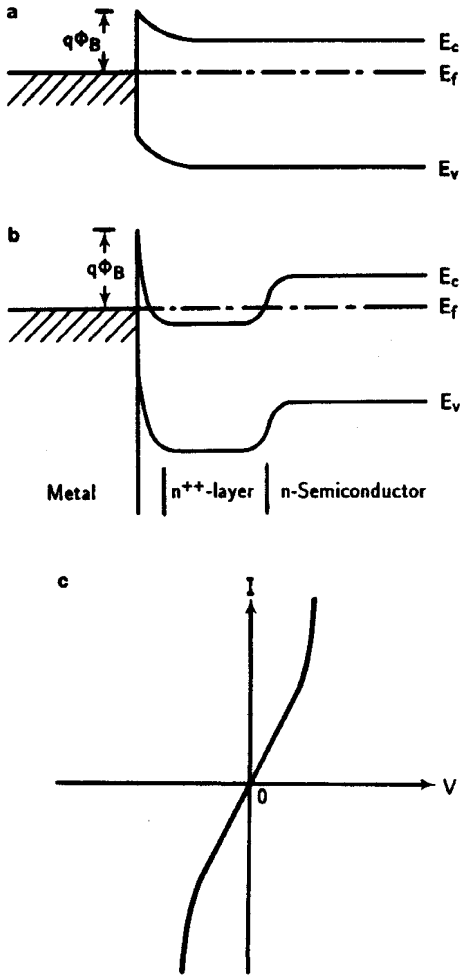


FIGURE 10.31. (a) Energy band diagram for a low-barrier Schottky contact, (b) a metal–n^{++}–n–semiconductor ohmic contact, and (c) the *I–V* curve for the ohmic contact structure shown in (b).

the barrier height of an n-GaAs Schottky barrier contact tends to remain nearly constant regardless of the metals used. This is because the surface state density for GaAs is usually very high (e.g., $\geq 10^{13}$ $cm^{-2} \cdot eV^{-1}$). As a result, the Fermi level at the interface is pinned at the level where the peak interface state density occurs. In this case, the barrier height of the Schottky contact is determined by the Fermi-level pinning at the surface and is independent of the metal work function.

The most widely used technique for forming ohmic contact on a semiconductor is by first growing a heavily doped thin surface layer to form an n^{++}/n or p^{++}/p high–low junction structure on an n- or p-type semiconductor substrate before making ohmic contact. By using such a structure, the barrier width between the metal and the heavily doped n^{++} or p^{++} layer can be greatly reduced, and hence quantum-mechanical tunneling of charge carriers through such a thin barrier becomes possible. In this case, the barrier becomes essentially transparent to

the charge carriers, and the specific contact resistance is usually very small. The heavily doped layer can be readily grown using alloying, thermal diffusion, ion implantation, or epitaxial growth with a suitable dopant impurity on the growth layer. This approach of forming ohmic contact has been quite successful for both the silicon and GaAs IC technologies. However, it is not as successful for many large band gap semiconductors such as CdS, AlN, and SiC materials because these crystals have a tendency to compensate the foreign dopant impurities. This is due to the large density of native defects created by the nonstoichiometric crystal structures in these crystals. Finally, it should be pointed out that good ohmic contacts on p-type III-V compound semiconductors are in general more difficult to achieve than their n-type counterparts. This is due to the fact that a p-type III-V compound semiconductor surface (e.g., p-type AlGaAs) is much easier to oxidize than an n-type surface during metallization or simple exposure to the air.

Another approach that has also been employed for forming ohmic contacts on a semiconductor is introducing a large density of recombination centers at the semiconductor surface before making ohmic contact. The recombination centers at the semiconductor surface may be created by damaging or straining the surface using mechanical lapping and polishing, or by introducing impurities at the semiconductor surface region to create large recombination centers. Plating of metals to such a damaged surface usually results in adequate ohmic contact.

The procedures for forming ohmic contacts on a semiconductor are discussed next. Various contact-forming techniques, such as alloying, thermal or E-beam evaporation, sputtering, ion implantation, plating, and liquid regrowth have been employed. The alloying process can be achieved by first placing the metal pellets or a thin metal foil on the semiconductor surface and then heating the specimen to the eutectic temperature of metal such that a small portion of the semiconductor is dissolved with the metal. The contact is then cooled, and semiconductor regrowth takes place. This regrowth results in the incorporation of some metal and residual impurities into the metal–semiconductor interface. In this method, impurities in the metal form a heavily doped interfacial layer of the same dopant type as the bulk semiconductor, and hence good ohmic contact on the semiconductor can be obtained. However, many large-band-gap semiconductors have contact problems because they cannot be doped heavily enough to form good ohmic contact. Another technique for making ohmic contact is to wet the semiconductor with a metal. In order to promote wetting and good ohmic contact, both the metal and semiconductor surface must be ultraclean. A flux is often used during alloying in order to remove any residual impurities or oxide film so that surface wetting can be enhanced. Care must be taken to identify the difference in thermal expansion coefficients between the semiconductor and metal so that residual stress does not develop in the semiconductor upon cooling from the alloying temperature to room temperature. Liquid regrowth can also be used to form a high–low junction (i.e., n^+/n or p^+/p) on a semiconductor substrate before forming ohmic contact. This technique has been used to form ohmic contact on GaAs Gunn-effect devices.

Ohmic contacts on semiconductors can also be achieved using electrodeless plating. This method involves the use of a metallic salt (such as $AuCl_3$), which is reduced to a metal at the semiconductor surface by a chemical reducing agent

present in the plating solution. Solutions of nickel, gold, and platinum are the most widely used metals for electrodeless plating in semiconductors. It has been found that a nickel film adheres better on a mechanically lapped semiconductor surface than on a chemically polished semiconductor surface.

In the traditional silicon IC technology, the most commonly used technique for depositing metal films on silicon devices in IC chips uses thermal or electron-beam (E-beam) evaporation. The E-beam evaporation is particularly attractive for integrated circuits in which complex contact patterns are formed using a photolithography technique. The metal contact is deposited using a heated filament, an electron-beam evaporation, or sputtering. The heated boat or filament technique is the simplest method, but suffers from the disadvantage of being contaminated by the heated container (e.g., tungsten basket or graphite boat). A shutter mechanism is often used to block the initial vapor that may be contaminated with the more volatile impurities. If multiple boats or filaments are used, two or more metals may be evaporated simultaneously with independent control of each evaporating source. In the E-beam system, the E-beam gun melts the evaporant alone, which serves as its own crucible. This eliminates contamination from the crucible and gives purer deposited metal films. In the sputtering system, positive gas ions bombard the source (cathode), emitting metal atoms. These ejected atoms traverse the vacuum chamber and are deposited onto the semiconductor substrate. The sputtering system has the advantage that the polarity of the system may be reversed so that sputtering may occur from the substrate to a remote anode, thereby cleaning the substrate surface. The back-sputtering approach is particularly useful for removing any residual thin oxide films or other impurities on the substrate surface while it is in the vacuum, and can be followed immediately by depositing metal contact on the substrate.

The standard procedure for making ohmic contacts in silicon IC fabrication is usually accomplished by a two-step process. In general, ion implantation is first employed to create a thin, heavily doped layer of the same dopant type as that of the bulk material to form a p^+/p or n^+/n high–low junction, and is then followed by deposition of metal contacts using either thermal or E-beam evaporation. Finally, thermal annealing is performed on the implanted region to achieve good ohmic contact. In this procedure, careful cleaning of the semiconductor surface is essential to ensure good ohmic contact prior to metal deposition.

The wire bonding in semiconductor devices is usually carried out using either the thermal compression or ultrasonic wire bonder. In thermal compression bonding, both heat and pressure are applied simultaneously to the ball bonder (using 1-mil gold or aluminum wires) and the contact pad. In ultrasonic bonding, a combination of pressure and 60 kHz ultrasonic vibrations is employed. The ultrasonic vibration gives rise to a scrubbing action that breaks up any thin-surface insulating film, and hence intimate contacts between the metal and semiconductor can be made. The advantage of using ultrasonic bonding is that heating is not required, and any previous bondings will not be affected.

In conventional silicon IC technologies, interconnects are incorporated after front-end processing. The front-end processing refers to the sequence of fabrication steps, typically at very high temperatures (700–1100°C), that form the

MOS transistors in the active regions, the pockets of thick isolation in the field regions that separate adjacent transistors, and the silicidation of the transistor terminals for low-resistance contacts. The back-end processing, which refers to the interconnection of transistors, is subsequently formed by contacting the transistor terminals and then vertically stacking layers of metal wires and vias encased in the dielectric materials. Back-end processing temperatures typically do not exceed 450°C to avoid melting of metals and to control stress. Integration success in the conventional silicon IC technology is largely attributed to the processes that maintain excellent planarity after fabricating each via and wire level. The state-of-the-art 0.25-μm back-end processing uses the integration of conventional Al metallization and Al alloy (Al/Cu) wires and W vias (plugs or studs), which serves as a blocking barrier to the Al for diffusing into silicon transistors and oxides.

In 1999, IBM Corp. introduced a new chip-manufacturing technique to the mainstream to boost the performance of the Power PC microprocessor by more than 30% using a combination of the SOI (silicon-on-insulator) technology and the copper-interconnect wiring scheme. The high-performance copper-interconnects technique offers lower contact resistance to the transistors, lower wiring parasitic capacitance, and significantly higher electromigration resistance over the standard Al or Al alloy interconnnects. It improves overall chip performance including higher speed, higher packing density, and lower power consumption for the microprocessor. In the silicon IC industry, the method of embedding metal structures in dielectrics (known as the Damascene process) is widely used for metal wire interconnects. The copper interconnects are usually deposited using electrochemical plating instead of the physical vapor deposition (PVD) process. Since the PVD process could not fill the Damascene features, and is more expensive than the electroplating technique, tungsten (W) via (plug) technology has matured to the point where void-free and untapered vias with aggressive aspect ratios exceeding 3:1 are routinely formed, thus enabling increases in wiring density and reduction of capacitive parasitics to under- and overlying wires. Advances in lithography alignment have also enabled borderless vias to be formed, thereby permitting even further improvement in wiring density. In addition, Damascence tungsten has been adapted as planar local interconnects for strapping source/drain and gate contacts. Although this process is more difficult to control, successful implementation of W-local interconnects can reduce the cell size of SRAMs used as microprocessor cache memories by 20–30%.

Problems

10.1. The saturation current density for a thoriated tungsten metal is 1 A/cm^2 at 1873 K, and the work function computed from (10.11) for this metal is 3.2 eV. Assume that $A_0 = 120$ A/cm$^2 \cdot$ K^2.
(a) Plot $\ln(J_s')$ versus $\mathcal{E}^{1/2}$ for $T = 1873$ K and for $10^4 < \mathcal{E}^{1/2} < 10^7$ V/cm.
(b) Repeat (a) for $T = 873$ K and 1500 K.

10.2. (a) Draw the energy band diagram for an ideal metal/p-type semiconductor Schottky barrier diode and show that the barrier height for a metal/p-type semiconductor is given by (10.14).
(b) Plot the energy band diagrams for an ideal metal/p-type semiconductor Schottky barrier diode for
(i) $\phi_m > \phi_s$.
(ii) $\phi_m < \phi_s$.
(iii) Explain which of the above cases would yield an ohmic or a Schottky contact.

10.3. Using (10.25), plot the depletion layer width versus reverse-bias voltage ($V_R = 0$ to 20 V) for a Au/n-type silicon Schottky barrier diode for $N_D = 10^{14}$, 10^{16}, and 10^{18}cm^{-3}, given $\varepsilon_s = 11.7$, $V_D = \phi_{Bn} - (k_B T/q)\ln(N_c/N_D)$, and $q\phi_{Bn} = 0.81$ eV at $T = 300$ K.

10.4. Taking into account the image-lowering effect, using (10.34), plot the saturation current density versus reverse-bias voltage for a Au/n-Si and Pt/n-Si Schottky diode. Assume that $q\phi_{Bn} = 0.81$ eV for a Au/n-Si Schottky diode, $q\phi_{Bn} = 0.90$ eV for a Pt/n-Si Schottky diode, and $A^* = 110\ \text{A/cm}^2 \cdot \text{K}^2$.

10.5. Derive (10.30) and (10.42), and compare the results with the current density equation derived from the thermionic–diffusion model by C. R. Crowell and S. M. Sze.[1]

10.6. Design a Au/n-type GaAs Schottky barrier photodiode for detecting a 20-GHz modulated optical signal with center wavelength at 0.84 μm. Show the Schottky barrier structure, and calculate the thickness of the AR coating layer (e.g., SiO_2), the diode area, and the RC time constant of this photodiode. If the incident photosignal has a power intensity of 2 mW/cm^2, what is the responsivity of this photodiode? (*Hint*: Choose your own design parameters.)

10.7. If the barrier height of a TiW/p-type Si Schottky barrier diode is equal to 0.55 eV, use (10.65) and (10.66) and (69) through (73) to design a TiW/n^+-p Schottky barrier diode structure to enhance the effective barrier height to 0.90 eV. Assuming that the p substrate has a dopant density of 1×10^{16} cm^{-3}, calculate the required dopant density and thickness of the n^+ surface layer.

10.8. Assuming that the diode ideality factor n for a Schottky barrier diode is defined by

$$n = \left(\frac{q}{k_B T}\right)\frac{\partial V}{\partial(\ln J)},$$

(a) Show that

$$n = \left\{1 + \left(\frac{\partial \Delta\phi}{\partial V}\right) + \left(\frac{k_B T}{q}\right)\left[\frac{\partial(\ln A^*)}{\partial V}\right]\right\}^{-1}.$$

(b) What are the possible physical mechanisms that may cause the n value to deviate from unity?

10.9. Using the general expressions of the barrier height for a Schottky barrier diode given by (10.48) and (10.49) to (10.51), with $D_s \approx 1.1 \times 10^{13}(1 - c_2)/c_2$ states/cm^2-eV, where D_s is the interface state density,
   (a) What is the barrier height as D_s approaches infinity? Explain the Fermi-level pinning effect under this condition.
   (b) If $D_s \to 0$, what is the value of c_2 and the expression for ϕ_{Bn}?
   (c) If the values of c_2, c_3, and χ_s for Si, GaAs, and GaP Schottky contacts are given by

$c_2 = 0.27, 0.09,$ and 0.27 for Si, GaAs, and GaP, respectively,
$c_3 = -0.66, -0.61,$ and -0.07V for Si, GaAs, and GaP, respectively,
$\chi_s = 4.05, 4.07,$ and 4.0 eV for Si, GaAs, and GaP, respectively,

   calculate the values of D_s, $q\phi_0$, and $q\phi_{Bn}$ for the above Schottky diodes.

10.10. Using (10.69) to (10.73), calculate the barrier height enhancement for a Au/p-n GaAs Schottky barrier diode for the following cases:
   (a) $N_d = 10^{16}$ cm^{-3}, plot ϕ_{Bn} versus $N_a(10^{15}$ to 5×10^{18} cm$^{-3})$ for $W_p = 0.3, 0.2, 0.1, 0.05,$ and $0.02\,\mu$m; $q\phi_{Bn} = 0.9$ eV for a Au/n-GaAs Schottky contact.
   (b) $N_a = 2 \times 10^{17}$ cm^{-3}, calculate and plot ϕ_{Bn} versus N_d and W_n for the values given in (a).

References

1. C. R. Crowell and S. M. Sze, *Solid-State Electron.* **8**, 979 (1966).
2. A. Y. C. Yu and C. A. Mead, *Solid-State Electron.* **13**, 97 (1970).
3. D. Kahng, *Bell Syst. Tech. J.* **43**, 215 (1964).
4. M. P. Lepselter and S. M. Sze, *Bell Syst. Tech. J.* **47**, 195 (1968).
5. A. M. Cowley and S. M. Sze, *J. Appl. Phys.* **36**, 3212 (1965).
6. R. H. Fowler, *Phys. Rev.* **38**, 45 (1931).
7. C. R. Crowell, J. C. Sarace, and S. M. Sze, *Trans. Metall. Soc. AIME* **233**, 478 (1965).
8. E. Hokelek and G. Y. Robinson, *J. Appl. Phys.* **54**(9), 5199 (1983).
9. W. Schottky, *Naturwissenschaften* **26**, 843 (1938).
10. J. Bardeen, *Phys. Rev.*, **71**,171 (1947).
11. K. A. Rickert, A. B. Ellis, J. K. Kim, J. Lee, F. J. Himpsel, F. Dwikusuma, and T. F. Kuech, *J. Appl. Phys.* **92**, 6671 (2002).
12. S. S. Li, *Solid-State Electron.* **21**, 435–438 (1977).
13. S. S. Li, C. S. Kim, and K. L. Wang, *IEEE Trans. Electron Devices* **ED-27**, 1310–1312 (1980).
14. D. H. Lee, S. S. Li, N. J. Sauer, and T. Y. Chang, *Appl. Phys. Lett.* **54**(19), 1863 (1989).
15. F. A. Padovani and R. Stratton, *Solid-State Electron.* **9**, 695 (1966).

Bibliography

N. Braslau, *Thin Solid Films* **104**, 391 (1983).
P. Chattopadhyay and A. N. Daw, *Solid-State Electron.* **28**, 831 (1985).
M. Heiblum, M. I. Nathan, and C. A. Chang, *Solid-State Electron.* **25**, 185 (1982).
H. K. Henisch, *Semiconductor Contacts: An Approach to Ideas and Models*, Clarendon Press, Oxford (1984).
V. G. Keramidas, *Inst. Phys. Conf. Ser.* No. 45, 396 (1979).
R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, Wiley, New York (1977).
D. A. Neamen, *Semiconductor Physics and Devices*, McGraw-Hill Publishing, New York (2003).
E. H. Rhoderick, *Metal–Semiconductor Contacts*, Clarendon Press, Oxford (1978).
E. H. Rhoderick and R. H. William, *Metal–Semiconductor Contacts*, Vol. 19, 2nd ed., Oxford University Press, Oxford (1988).
V. L. Rideout, *Solid-State Electron.* **18**, 541 (1975).
B. L. Sharma, in: *Semiconductors and Semimetals*, Vol. 15, Academic Press, New York (1981), p. 1.
B. L. Sharma, *Metal–Semiconductor Schottky Barrier Junctions and Their Applications*, Plenum Press, New York (1984).
W. E. Spicer, I. Lindau, P. Skeath, C. Y. Su, and P. W. Chyre, *Phys. Rev. Lett.* **44**, 420 (1980).
W. E. Spicer, P. W. Chyre, C. M. Garner, I. Lindau, and P. Pianetta, *Surf. Sci.* **86**, 763 (1979).
S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1982).
A. Van der Ziel, *Solid State Physical Electronics*, 3rd ed., Prentice Hall, New York (1976).
C. W. Wilmsen, *Physics and Chemistry of III-V Compound Semiconductor Interfaces*, Plenum Press, New York (1985).

11
p-n Junction Diodes

11.1. Introduction

In this chapter, the basic device physics, the ideal static and dynamic characteristics, the operation principles, and practical applications of p-n junctions will be described. Unlike a Schottky diode (a majority carrier device), a p-n junction diode is known as a minority carrier device since the current conduction is controlled by the diffusion of minority carriers (i.e., electrons in the p region and holes in the n region) in a p-n junction diode.

A p-n junction diode can be fabricated by doping the semiconductor material with opposite doping impurities (i.e., acceptor or donor impurities) to form the p and n regions of the diode. If a p-n junction is formed on the same semiconductor it is referred to as a p-n homojunction diode. On the other hand, if a p-n junction is formed using two semiconductor materials of different band gaps and with opposite doping impurities, then it is referred to as a p-n heterojunction diode. Both the p-n homo- and heterojunction diodes are discussed in this chapter. The p-n junction plays an important role as the basic device structure for fabricating a wide variety of electronic and photonic devices. For example, p-n junction structures have been used in fabricating switching diodes, diode rectifiers, solar cells, light emitting diodes (LEDs), laser diodes (LDs), photodetectors, bipolar junction transistors (BJTs), heterojunction bipolar transistors (HBTs), junction field-effect transistors (JFETs), metal–semiconductor field-effect transistors (MESFETs), high-electron mobility transistors (HEMTs), tunnel diodes, multiquantum well (MQW) and superlattice (SL) devices. The p-n heterojunctions can be formed from a wide variety of elemental and compound semiconductors such as n-Si/p-SiGe, n-ZnSe/p-GaAs, p-AlGaAs/n-GaAs, p-Ge/n-GaAs, n-InGaAs/n-InP, p-InAlAs/n-InGaAs, p-GaN/n-InGaN, and p-AlGaN/n-InGaN semiconductor heterojunction devices.

The p-n junction theory serves as a foundation for the interpretation of device physics in various semiconductor devices. The basic device theory used in predicting the current–voltage (I–V) characteristics in a p-n junction diode was first developed by Shockley,[1] and later extended by Sah, Noyce, and Shockley,[2] and Moll.[3] Derivation of charge carrier distribution, built-in potential, electric field,

and the potential distribution in the junction space charge region of a p-n junction under equilibrium and applied bias conditions are given in Sections 11.2 and 11.3. The minority carrier distributions and current flow in a p-n junction are derived using the continuity equations presented in Chapter 6. The current–voltage (I–V) and capacitance–voltage (C–V) characteristics under forward- and reverse-bias conditions are described in Section 11.4. The minority carrier storage and transient behavior in a p-n junction are discussed in Section 11.5. Section 11.6 presents the junction breakdown phenomena in a p-n junction under large reverse-bias conditions. Finally, the basic device theory and general characteristics of a p-n heterojunction diode are discussed in Section 11.7.

11.2. Equilibrium Properties of a p-n Junction Diode

A p-n junction diode is formed when an opposite doping impurity (i.e., donor or acceptor impurity) is introduced into a region of the semiconductor using the alloying, thermal diffusion, ion-implantation, or epitaxial growth technique. For example, a silicon p-n junction diode can be formed when a p-type doping impurity such as boron (B), aluminum (Al), or gallium (Ga) is introduced into an n-type silicon substrate via the thermal diffusion or ion-implantation process. On the other hand, a silicon n-p junction diode is formed when an n-type doping impurity such as a phosphorus (P) or arsenic (As) impurity is introduced into a p-type silicon substrate. The n-type doping impurity is called a donor impurity since it will contribute an extra electron to the silicon lattice, while the p-type doping impurity is called an acceptor impurity since it will give an extra hole to the silicon lattice. For III-V compound semiconductors such as GaAs, InP, InGaAs, and AlGaAs, a p-n junction can be formed in these material systems using different growth techniques such as liquid-phase epitaxy (LPE), vapor-phase epitaxy (VPE), metal-organic chemical vapor deposition (MOCVD), and molecular beam epitaxy (MBE).

Figures 11.1a and b show the energy band diagrams of a p-n junction under thermal equilibrium conditions before and after the intimate contacts. It is noted that the Fermi level is constant across the entire region of the p-n junction under thermal equilibrium conditions. Figure 11.2a shows the charge distribution in the p- and n-quasineutral regions as well as in the depletion region of the junction. In general, depending on the doping impurity profile across the junction, a diffused p-n junction may be approximated by either a step- (or abrupt-) junction or a linear-graded junction. As shown in Figure 11.3a, the impurity profile for a step junction changes abruptly across the metallurgical junction of the diode, while the impurity profile for a linear-graded junction varies linearly with distance across the junction, as illustrated in Figure 11.3b.

The static properties of an abrupt p-n junction and a linear-graded p-n junction diode are discussed next. The carrier distribution, built-in potential, electric field, and potential profile in the junction space-charge region of a p-n junction diode can be derived for both the abrupt- and linear-graded junctions using Poisson's equation and continuity equations.

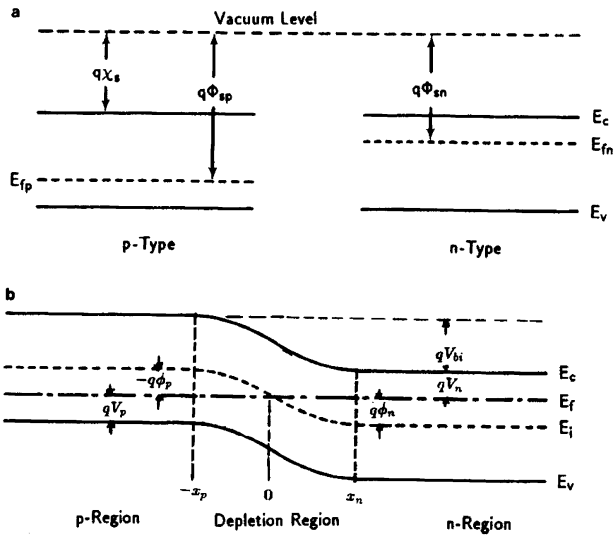


FIGURE 11.1. Energy band diagrams for an isolated n- and p-type semiconductor (a) before contact, and (b) in intimate contact.

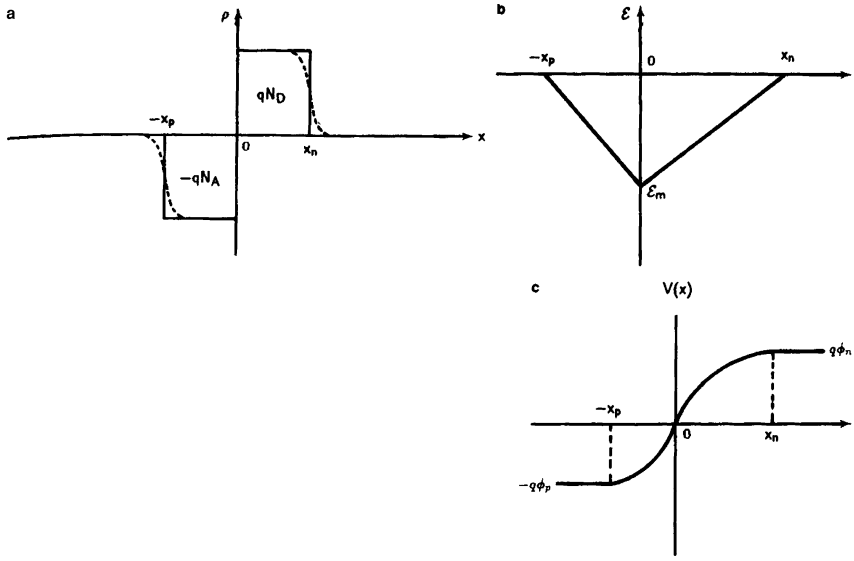


FIGURE 11.2. (a) Space-charge distribution, (b) electric field, and (c) potential distribution for an abrupt p-n junction diode.

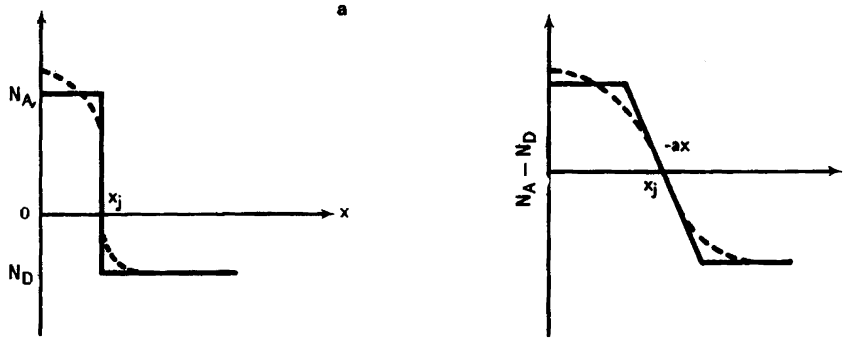


FIGURE 11.3. Impurity profile for (a) a shallow-diffused junction (i.e., an abrupt or step junction) and (b) a deep-diffused junction (i.e., a linearly graded junction).

In thermal equilibrium, the Fermi level is constant throughout the entire p-n junction, as shown in Figure 11.1b. The one-dimensional (1-D) Poisson's equation, which relates the charge density ρ to the potential $V(x)$, is given by

$$\frac{\mathrm{d}^2V(x)}{\mathrm{d}x^2} = -\frac{\rho}{\varepsilon_0\varepsilon_\mathrm{s}} = \left(\frac{q}{\varepsilon_0\varepsilon_\mathrm{s}}\right)(n - p - N_\mathrm{d} + N_\mathrm{a}). \tag{11.1}$$

As shown in Figure 11.1b, the electron and hole densities in the n and p regions of the junction can be expressed in terms of the intrinsic carrier density, n_i, and the electrostatic potential, ϕ, and are given by

$$n = n_\mathrm{i} \exp\left(\frac{\phi_\mathrm{n}}{V_\mathrm{T}}\right) \tag{11.2}$$

and

$$p = n_\mathrm{i} \exp\left(\frac{-\phi_\mathrm{p}}{V_\mathrm{T}}\right), \tag{11.3}$$

where n_i is the intrinsic carrier density, $V_\mathrm{T} = k_\mathrm{B}T/q$ is the thermal voltage, and ϕ_n, ϕ_p denote the electrostatic potential in the n and p regions of the diode, respectively. Using proper boundary conditions, expressions for the potential, electric field, and charge distribution in the different regions of the p-n junction can be derived using (11.1) to (11.3). Figure 11.1b shows the three distinct regions in a p-n junction, namely, the n- and p-quasineutral regions (QNR) away from the metallurgical junction, and the space-charge (or depletion) region (SCR), which is occupied by the ionized shallow acceptors in the p-depletion region and the ionized shallow donors in the n-depletion region. In addition to these three distinct regions, a transition region of a few Debye lengths may also be presented in the boundary region between the QNR and the SCR interfacial layers. This transition layer is usually much smaller than the depletion layer width, and hence may be neglected in the diode analysis.

In the n- and p-quasineutral regions, the total charge density is equal to 0, and (11.1) becomes

$$\frac{d^2V(x)}{dx^2} = 0 \tag{11.4}$$

and

$$n - p - N_d + N_a = 0. \tag{11.5}$$

In the n-quasineutral region, N_a is assumed equal to 0 (or $N_a \ll N_d$), and $p \ll n$. The electrostatic potential ϕ_n at the depletion layer edge of the n-quasineutral region can be derived by assuming $N_a = p = 0$ in (11.5) and then substituting the result into (11.2), which yields

$$\phi_n = V_T \ln\left(\frac{N_d}{n_i}\right). \tag{11.6}$$

Similarly, the potential distribution at the depletion edge of the p-quasineutral region can be written as

$$\phi_p = -V_T \ln\left(\frac{N_a}{n_i}\right). \tag{11.7}$$

Therefore, the built-in or diffusion potential of a p-n junction diode between the n- and p-quasineutral regions can be obtained from (11.6) and (11.7), and one has

$$V_{bi} = \phi_n - \phi_p = V_T \ln\left(\frac{N_d N_a}{n_i^2}\right), \tag{11.8}$$

where V_{bi} is known as the built-in or diffusion potential of a p-n junction diode in thermal equilibrium. For simplicity, the free-carrier density in the depletion region is assumed equal to 0 (i.e., $n = p = 0$). Thus, (11.1) becomes

$$\frac{d^2V(x)}{dx^2} = \left(\frac{q}{\varepsilon_s \varepsilon_0}\right)(N_a - N_d). \tag{11.9}$$

Equation (11.9) may be used in solving the potential and electric field distributions in the junction space-charge region of the abrupt- and a linear-graded p-n junction shown in Figures 11.3a and 11.3b. The abrupt junction approximation can be applied to the shallow-diffused step junction or an ion-implanted junction, while the linear-graded junction approximation is more suitable for a deep-diffused p-n junction diode.

Figure 11.2a shows the impurity distribution in the space-charge region of an abrupt p-n junction diode. It is noted that the boundary layer effect (i.e., the spreading of space charges a few Debye lengths into the quasineutral regions) shown by the dotted line is neglected in the present analysis. In the depletion region, free carriers are negligible, and the Poisson equation in the n- and p-space-charge

regions are given respectively by

$$\frac{\mathrm{d}^2V(x)}{\mathrm{d}x^2} = \begin{cases} -\dfrac{qN_\mathrm{d}}{\varepsilon_0\varepsilon_\mathrm{s}} & \text{for } 0 < x < x_n, \quad (11.10) \\ \dfrac{qN_\mathrm{a}}{\varepsilon_0\varepsilon_\mathrm{s}} & \text{for } -x_\mathrm{p} < x < 0, \quad (11.11) \end{cases}$$

where x_n and x_p denote the depletion layer widths in the n and p regions, respectively. The charge-neutrality condition in the depletion region of the junction requires that

$$N_\mathrm{a}x_\mathrm{n} = N_\mathrm{d}x_\mathrm{p}, \tag{11.12}$$

which shows that the depletion layer width on either side of the junction space-charge region is inversely proportional to the doping density. The total depletion layer width W_d of the junction is given by

$$W_\mathrm{d} = x_\mathrm{n} + x_\mathrm{p}. \tag{11.13}$$

From (11.12), it is seen that if N_a is much greater than N_d, then x_n will be much larger than x_p, and the depletion region will spread mostly into the n region. Thus, for $x_\mathrm{n} \gg x_\mathrm{p}$, $W_\mathrm{d} \approx x_\mathrm{n}$, and one has a one-sided abrupt p-n junction diode. In this case, the depletion layer width on the heavily doped p region becomes negligible compared to the depletion layer width on the lightly doped n region. As a result, one can solve the Poisson equation for the lightly doped side (i.e., n region) to obtain basic information on the junction characteristics. Integration of (11.10) once from x to x_n yields the electric field

$$\mathcal{E}(x) = -\frac{\mathrm{d}V(x)}{\mathrm{d}x} = \left(\frac{qN_\mathrm{d}}{\varepsilon_0\varepsilon_\mathrm{s}}\right)(x - x_\mathrm{n}), \tag{11.14}$$

which is obtained from the boundary condition that $\mathrm{d}V(x)/\mathrm{d}x = 0$ at $x = x_\mathrm{n}$. Since the maximum electric field occurs at $x = 0$, (11.14) can be expressed as

$$\mathcal{E}(x) = \mathcal{E}_\mathrm{m}\left(1 - \frac{x}{x_\mathrm{n}}\right) \quad \text{for } 0 < x < x_\mathrm{n}, \tag{11.15}$$

where $\mathcal{E}_\mathrm{m} = qN_\mathrm{d}x_\mathrm{n}/\varepsilon_0\varepsilon_\mathrm{s}$ is the maximum electric field strength at $x = 0$. It is noted that the electric field is negative throughout the entire depletion region, and varies linearly with distance from $x = 0$ to either side of the junction. As illustrated in Figure 11.2b, the electric field on the right-hand side of the junction (i.e., the n region) is negative since the force exerted by the electric field is offset by the electron diffusion to the left from the quasineutral n region.

Similarly, the electric field in the p region is also negative in order to retard the diffusion of holes to the right-hand side of the junction. Thus, the electric field for $x < 0$ can be written as

$$\mathcal{E}(x) = -\left(\frac{qN_\mathrm{a}}{\varepsilon_\mathrm{s}\varepsilon_0}\right)(x + x_\mathrm{p}) \quad \text{for } -x_\mathrm{p} < x < 0. \tag{11.16}$$

The potential in the n region can be obtained by integrating (11.14) once more, yielding

$$V(x) = V_{\mathrm{n}} - \left(\frac{qN_{\mathrm{d}}x_{\mathrm{n}}^2}{2\varepsilon_{\mathrm{s}}\varepsilon_0}\right)\left(1 - \frac{x}{x_{\mathrm{n}}}\right)^2 \quad \text{for } 0 < x < x_{\mathrm{n}}, \tag{11.17}$$

where $V_{\mathrm{n}} = V_{\mathrm{T}} \ln(N_{\mathrm{c}}/N_{\mathrm{d}})$ is the potential difference between the conduction band edge and the Fermi level at the depletion edge of the n-quasineutral region. Similarly, the potential in the p region is given by

$$V(x) = V_{\mathrm{p}} + \left(\frac{qN_{\mathrm{a}}x_{\mathrm{p}}^2}{2\varepsilon_0\varepsilon_{\mathrm{s}}}\right)\left(1 - \frac{x}{x_{\mathrm{p}}}\right)^2 \quad \text{for } -x_{\mathrm{p}} < x < 0, \tag{11.18}$$

where $V_{\mathrm{p}} = V_{\mathrm{T}} \ln(N_{\mathrm{v}}/N_{\mathrm{a}})$ is the potential at the edge of the p-depletion region.

The built-in potential V_{bi}, which is defined as the total potential change from the quasineutral p region to the quasineutral n region, is equal to $(\phi_{\mathrm{n}} - \phi_{\mathrm{p}})$, as given by (11.8). It is noted that most of the potential drop and the depletion region are on the lightly doped side of the junction. The depletion layer width can be obtained by solving (11.12), (11.17), and (11.18) at $x = 0$, and the result is

$$W_{\mathrm{d}} = x_{\mathrm{n}} + x_{\mathrm{p}} = \left[\left(\frac{2\varepsilon_{\mathrm{s}}\varepsilon_0 V_{\mathrm{bi}}}{q}\right)\left(\frac{N_{\mathrm{d}} + N_{\mathrm{a}}}{N_{\mathrm{d}}N_{\mathrm{a}}}\right)\right]^{1/2}. \tag{11.19}$$

Equation (11.19) shows that the depletion layer width depends on the doping density of the lightly doped n-base region (i.e., for $N_{\mathrm{a}} \gg N_{\mathrm{d}}$, $W_{\mathrm{d}} \approx (2\varepsilon_{\mathrm{s}}\varepsilon_0 V_{\mathrm{bi}}/qN_{\mathrm{d}})^{1/2}$, which varies inversely with the square root of the doping density.

For a linear-graded p-n junction, the space-charge distribution in the depletion region is given by

$$N_{\mathrm{a}} - N_{\mathrm{d}} = -ax, \tag{11.20}$$

where a is the slope of the doping impurity density profile (cm^{-4}). Thus, the Poisson equation for a linear-graded p-n junction diode can be expressed by

$$\frac{\mathrm{d}^2V(x)}{\mathrm{d}x^2} = -\left(\frac{q}{\varepsilon_0\varepsilon_{\mathrm{s}}}\right)ax. \tag{11.21}$$

Using the same procedures as for the step-junction diode described above, one can derive the depletion layer width W_{d} and the built-in potential V_{bi} for a linear-graded p-n junction diode, which yields

$$W_{\mathrm{d}} = \left[\frac{12\varepsilon_0\varepsilon_{\mathrm{s}}V_{\mathrm{bi}}}{qa}\right]^{1/3} \tag{11.22}$$

and

$$V_{\mathrm{bi}} = 2V_{\mathrm{T}} \ln\left(\frac{aW_{\mathrm{d}}}{2n_{\mathrm{i}}}\right). \tag{11.23}$$

Comparing (11.19) and (11.22), one finds that for a linear-graded junction, W_{d} depends on $(V_{\mathrm{bi}}/N_{\mathrm{d}})^{1/3}$, while for an abrupt junction, it depends on $(V_{\mathrm{bi}}/N_{\mathrm{d}})^{1/2}$.

11.3. p-n Junction Diode Under Bias Conditions

When an external bias voltage is applied to a p-n junction diode, the thermal equilibrium condition is disrupted and a current flow across the junction results. Since the resistance across the depletion region is many orders of magnitude larger than the resistance in the quasineutral regions, the voltage drops across both the n- and p-quasineutral regions are negligible compared to the voltage drop across the depletion region. Thus, it is reasonable to assume that the voltage applied to a p-n junction diode is roughly equal to the voltage drop across the depletion layer region. The current–voltage (I–V) characteristics of a p-n junction diode under reverse- and forward-bias conditions are discussed next.

The current flow in a p-n junction depends on the polarity of the applied bias voltage. Under forward-bias conditions, the current increases exponentially with applied voltage. Under reverse-bias conditions, the current flow is limited mainly by the thermal generation current and hence depends very little on the applied voltage. Figure 11.4 shows the energy band diagrams for a p-n junction diode under (a) zero-bias, (b) forward-bias, and (c) reverse-bias conditions. As shown in Figure 11.4b, when a forward-bias voltage V (i.e., positive polarity applied to the p-side and negative to the n-side) is applied to the p-n junction, the potential barrier across the junction will decrease to $(V_{bi} - V)$. In this case, the potential barrier for the majority carriers at the junction is reduced, and the depletion layer width is decreased. Thus, under forward-bias conditions a small increase in applied voltage will result in a large increase in current flow across the junction. On the other hand, if a reverse-bias voltage is applied to the junction, then the potential barrier across the junction will increase to $(V_{bi} + V)$, as shown in Figure 11.4c. Therefore, under a reverse-bias condition the potential barrier for the majority carriers and the depletion layer width will increase with increasing reverse-bias voltage. As a result, current flow through the junction becomes very small, and the junction impedance is extremely high.

The abrupt junction approximation is used to analyze the I–V characteristics of a step-junction diode under bias conditions. In the analysis it is assumed that (1) the entire applied voltage drop is only across the junction space-charge region, and is negligible in the n- and p-quasineutral regions; (2) the solution of Poisson's equation obtained under thermal equilibrium conditions can be modified to the applied bias case, and (3) the total potential across the junction space-charge region changes from V_{bi} for the equilibrium case to $(V_{bi} \pm V)$ when a bias voltage is applied to the p-n junction. Thus, the depletion layer width for a step-junction diode under bias conditions is given by

$$W_d = x_n + x_p = \left[\left(\frac{2\varepsilon_0\varepsilon_s(N_a + N_d)}{qN_aN_d}\right)(V_{bi} \pm V)\right]^{1/2}, \qquad (11.24)$$

where the plus sign is for the reverse-bias case, and the minus sign is for the forward-bias case. Equation (11.24) shows that for a step-junction diode under reverse-bias conditions, the depletion layer width W_d is proportional to the square root of the applied voltage.

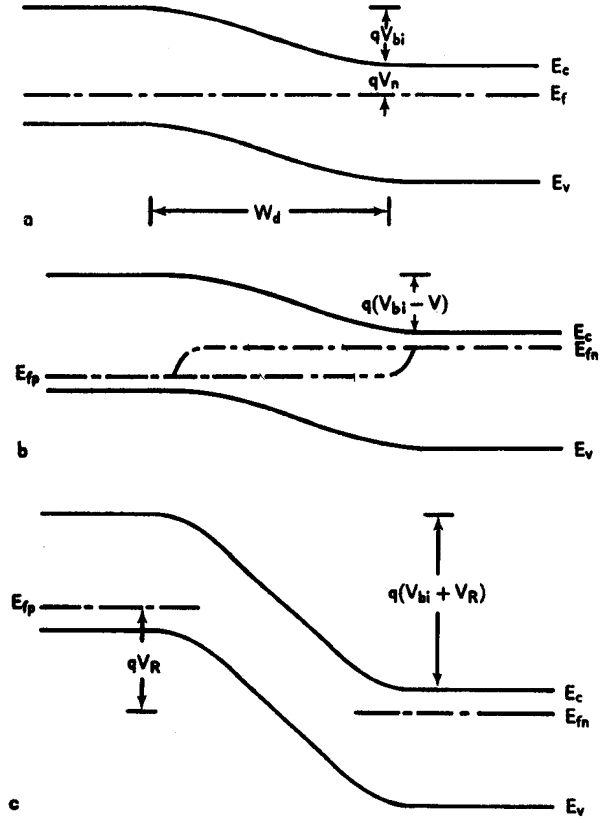


FIGURE 11.4. Energy band diagram of a p-n junction diode under (a) zero-bias, (b) forward-bias, and (c) revised-bias conditions.

Similarly, for a linear-graded junction, the depletion layer width under bias conditions can be expressed by

$$W_{\mathrm{d}} = \left[\frac{12\varepsilon_0\varepsilon_{\mathrm{s}}(V_{\mathrm{bi}} \pm V)}{qa}\right]^{1/3}. \tag{11.25}$$

The relationship between the maximum electric field and the applied bias voltage in the junction space-charge region can be derived as follows. For a step-junction diode, assuming $W_{\mathrm{d}} \approx x_{\mathrm{n}}$ (i.e., $N_{\mathrm{a}} \gg N_{\mathrm{d}}$ and $x_{\mathrm{n}} \gg x_{\mathrm{p}}$), the maximum electric field at the junction can be derived from (11.15) and (11.19), and one has

$$\mathcal{E}_{\mathrm{m}} = \frac{qN_{\mathrm{d}}x_{\mathrm{n}}}{\varepsilon_0\varepsilon_{\mathrm{s}}} \approx \frac{2(V_{\mathrm{bi}} - V)}{W_{\mathrm{d}}}. \tag{11.26}$$

Similarly, the maximum electric field versus applied bias voltage for a linearly graded junction is given by

$$\mathcal{E}_{\mathrm{m}} = \frac{3(V_{\mathrm{bi}} - V)}{2W_{\mathrm{d}}}. \tag{11.27}$$

Another important parameter that needs to be considered is the depletion capacitance in the space-charge region of a p-n junction. A p-n junction diode can be viewed as a parallel-plate capacitor filled with positive and negative fixed charges arising from the ionized donor and acceptor impurities in the depletion region, which determine the junction capacitance of the diode. For a step-junction diode with doping densities of N_{a} and N_{d} in the p and n regions, respectively, the transition capacitance per unit area may be derived from the total space charge Q_{s} per unit area on either side of the depletion region. Thus, one can write

$$C_{\mathrm{j}} = \frac{\mathrm{d}Q_s}{\mathrm{d}V} = \frac{\mathrm{d}(qN_{\mathrm{a}}x_{\mathrm{p}})}{\mathrm{d}V} = \frac{\mathrm{d}(qN_{\mathrm{d}}x_{\mathrm{n}})}{\mathrm{d}V}. \tag{11.28}$$

If one assumes that N_{a} and N_{d} are constant and independent of the position, and uses the relations $x_{\mathrm{p}} = (N_{\mathrm{d}}/N_{\mathrm{a}})x_{\mathrm{n}}$ and $W_{\mathrm{d}} = x_{\mathrm{n}} + x_{\mathrm{p}}$, then the small signal transition capacitance per unit area can be derived from (11.28), and one has

$$C_{\mathrm{j}} = \sqrt{\frac{q\varepsilon_0\varepsilon_{\mathrm{s}}}{2(1/N_{\mathrm{d}} + 1/N_{\mathrm{a}})(V_{\mathrm{bi}} - V)}}. \tag{11.29}$$

For a one-sided step-junction diode (i.e., $N_{\mathrm{a}} \gg N_{\mathrm{d}}$), (11.29) predicts that the transition capacitance due to fixed charges in the depletion region is directly proportional to the square root of the doping density, and varies inversely with the square root of the applied bias voltage for $|-V| \gg V_{\mathrm{bi}}$. The transition capacitance per unit area for a one-sided step-junction diode is equal to $\varepsilon_0\varepsilon_{\mathrm{s}}/W_{\mathrm{d}}$. Thus, from (11.29) and assuming $N_{\mathrm{a}} \gg N_{\mathrm{d}}$ one obtains

$$C_{\mathrm{j}} = \frac{\mathrm{d}Q_{\mathrm{s}}}{\mathrm{d}V} = \frac{\varepsilon_{\mathrm{s}}\varepsilon_0}{W_{\mathrm{d}}} = \left[\frac{q\varepsilon_0\varepsilon_{\mathrm{s}}N_{\mathrm{d}}}{2(V_{\mathrm{bi}} \pm V)}\right]^{1/2}. \tag{11.30}$$

Equation (11.30) shows that the inverse of the capacitance ($1/C_{\mathrm{j}}^2$) square varies linearly with the applied voltage V. Thus, a plot of $1/C_{\mathrm{j}}^2$ versus V yields a straight line. The slope of this straight line yields the doping density of the lightly doped semiconductor (i.e., n region), and the intercept of the $1/C_{\mathrm{j}}^2$ plot on the voltage axis gives the built-in potential V_{bi}. It is noted that if the doping density is not uniform across the lightly doped n region, the doping density profile can be determined using a differential C–V technique similar to the one described above, except that the doping densities are determined piecewise at a small incremental voltage across the n region, and the entire doping profile in the n region can be determined by this method. Figure 11.5 shows a typical $1/\mathrm{C}_{\mathrm{j}}^2$ versus V plot for a step junction diode. The doping density in the substrate and the built-in potential can be determined from this plot.

The transition capacitance for a linear-graded junction diode can be derived in a similar way as that of the step-junction diode discussed above. Thus, from (11.25),

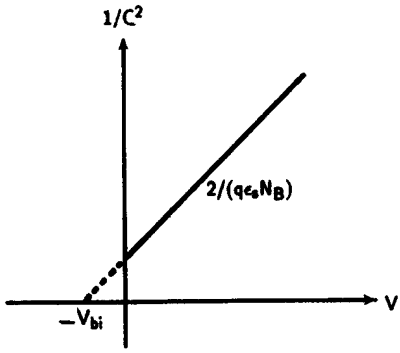


FIGURE 11.5. Inverse of capacitance squared versus applied reverse-bias voltage for a one-sided step-junction diode.

the transition capacitance for a linear-graded junction can be expressed by

$$C_j = \frac{dQ_s}{dV} = \frac{\varepsilon_s \varepsilon_0}{W_d} = \left[\frac{qa(\varepsilon_s \varepsilon_0)^2}{12(V_{bi} \pm V)}\right]^{1/3}, \tag{11.31}$$

which shows that the transition capacitance for a linear-graded junction diode is inversely proportional to the cube root of the applied reverse-bias voltage.

11.4. Minority Carrier Distribution and Current Flow

In order to derive the current density equations for an ideal p-n junction diode, it is necessary to find the minority carrier density distributions at the edges of the depletion layer near both the p- and n-quasineutral regions under applied bias conditions. Figure 11.6 shows a schematic diagram of a p-n junction diode to be used for deriving the current density equations in the n- and p-quasineutral regions as well as in the depletion region. The cross-sectional area of the diode perpendicular to the current flow is assumed equal to A. As illustrated in Figure. 11.6, the minority carrier densities at the edge of the quasineutral p (at $x = -x_p$) and n (at $x = x_n$) regions can be related to the majority carrier densities at the edge of the depletion region under bias conditions. These are given by

$$p_n(x_n) = p_{p0}(-x_p)\ \exp[-q(V_{bi} - V)/k_B T], \tag{11.32}$$

which is the hole density at the depletion edge of the n-quasineutral region, and

$$n_p(-x_p) = n_{n0}(x_n)\ \exp[-q(V_{bi} - V)/k_B T] \tag{11.33}$$

is the electron density at the depletion edge of the p region. It is noted that $p_{p0}(-x_p) = N_a(-x_p)$ and $n_{n0}(x_n) = N_d(x_n)$ denote the majority carrier densities at the edges of the p-quasineutral and n-quasineutral regions, respectively. If the applied voltage V is set equal to 0, then $n_p = n_{p0} = n_{n0} \exp(-qV_{bi}/k_B T)$ and $p_n = p_{n0} = p_{p0} \exp(qV_{bi}/k_B T)$, which are valid only for the low-level injection

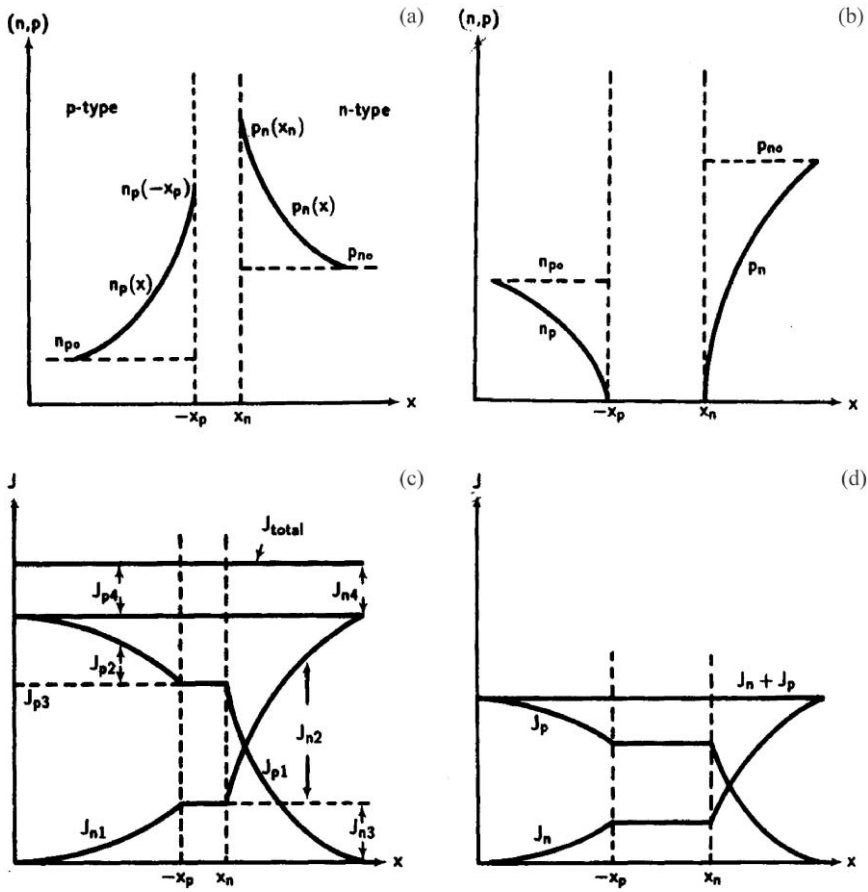


FIGURE 11.6. Minority carrier distribution under (a) forward-bias and (b) reverse-bias conditions current components under (c) forward-bias and (d) reverse-bias conditions: j_{p1} and j_{n1} are injected minority hole and electron currents; j_{n2} and j_{p2} are majority electron and hole currents recombining with j_{p1} and j_{n1}, respectively; j_{n3} and j_{p3} are electron and hole recombination currents in the space-charge region.

case. The excess carrier densities at the depletion layer edges under bias conditions can be obtained from (11.32) and (11.33) by subtracting their equilibrium densities, which yields

$$p'_n(x_n) = p_n(x_n) - p_{n0}(x_n) = p_{n0}(x_n)(e^{qV/k_BT} - 1), \tag{11.34}$$

$$n'_p(-x_p) = n_p(-x_p) - n_{p0}(-x_p) = n_{p0}(-x_p)(e^{qV/k_BT} - 1). \tag{11.35}$$

If (11.34) and (11.35) are used as the boundary conditions, then the expressions for the spatial distributions of the minority carrier densities can be derived by solving the continuity equations in the quasineutral regions of a p-n junction

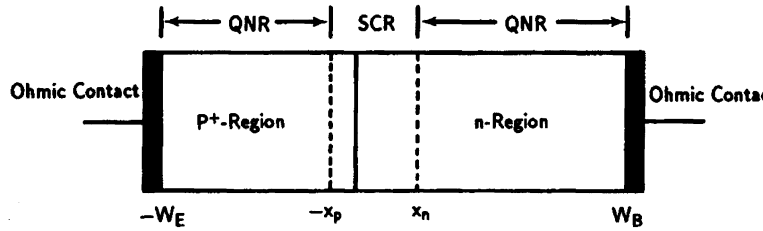


FIGURE 11.7. Schematic diagram of a p-n junction diode showing the dimensions and boundaries of the n- and p-quasi-neutral regions and the space-charge region.

diode. It is seen from (11.32) to (11.35) that the majority carrier density is insensitive to the applied bias voltage, while the minority carrier density depends exponentially on the applied bias voltage. It can be shown that the current flow in a p-n junction diode is in fact governed by the diffusion of the minority carriers across the p-n junction.

The derivation of electron and hole current densities in a p-n junction diode may be obtained from the continuity equations given in Chapter 6. The one-dimensional (i.e., x-direction) continuity equation for the excess hole density injected into the n-quasi-neutral region under steady-state conditions is given by

$$D_{\mathrm{p}}\frac{\mathrm{d}^2 p'_{\mathrm{n}}}{\mathrm{d}x^2} - \frac{p'_{\mathrm{n}}}{\tau_{\mathrm{p}}} = 0. \tag{11.36}$$

The solution of (11.36) can be expressed by

$$p'_{\mathrm{n}}(x) = A\mathrm{e}^{-(x-x_{\mathrm{n}})/L_{\mathrm{p}}} + B\mathrm{e}^{(x-x_{\mathrm{n}})/L_{\mathrm{p}}}, \tag{11.37}$$

where A and B are constants to be determined by the boundary conditions given by (11.34) and (11.35), and $L_{\mathrm{p}} = (D_{\mathrm{p}}\tau_{\mathrm{p}})^{1/2}$ is the hole diffusion length; D_{p} and τ_{p} denote the hole diffusion constant and hole lifetime, respectively.

The solutions of (11.36) can be obtained by considering two special cases, namely, the long-base diode with base width larger than the hole diffusion length (i.e., $W_{\mathrm{B}} \gg L_{\mathrm{p}}$) and the short-base diode with base width smaller than the hole-diffusion length (i.e., $L_{\mathrm{p}} \gg W_{\mathrm{B}}$). The current densities in the n- and p-quasineutral regions as well as in the depletion region, as shown in Figure 11.7, for both the long-base and short-base diodes will be derived next.

For a long-base diode, the base width in the n-quasineutral region is much larger than the hole diffusion length. As a result, the excess hole density $p'_{\mathrm{n}}(x)$ will decrease exponentially with increasing distance x, and the constant B in (11.37) can be set equal to 0. Constant A can be determined from (11.34) and (11.37) at $x = x_{\mathrm{n}}$, and one obtains

$$p'_{\mathrm{n}}(x) = p_{\mathrm{n0}}(\mathrm{e}^{qV/k_{\mathrm{B}}T} - 1)\mathrm{e}^{-(x-x_{\mathrm{n}})/L_{\mathrm{p}}}. \tag{11.38}$$

Equation (11.38) is the excess hole density in the n-quasineutral region of the p-n junction. Figures 11.6a and b show the distributions of the minority carriers

in the p- and n-quasineutral regions under forward- and reverse-bias conditions, respectively, and Figures 11.6c and d show the corresponding current densities under forward- and reverse-bias conditions. The excess carriers inside the depletion region are assumed equal to 0. The hole current density in the n-quasineutral region is contributed to only by the diffusion of excess holes in this region. Thus, from (11.38) one obtains

$$J_{\mathrm{p}}(x) = -q D_{\mathrm{p}} \frac{\mathrm{d}p'_{\mathrm{n}}}{\mathrm{d}x} = \left(\frac{q D_{\mathrm{p}} n_{\mathrm{i}}^2}{N_{\mathrm{d}} L_{\mathrm{p}}} \right) (\mathrm{e}^{qV/k_{\mathrm{B}}T} - 1)\, \mathrm{e}^{-(x-x_{\mathrm{n}})/L_{\mathrm{p}}}, \qquad (11.39)$$

where $p_{\mathrm{n0}} = n_{\mathrm{i}}^2/N_{\mathrm{d}}$ is used in the preexponential factor in (11.39). As shown in Figure 11.6c, the hole current density (J_{p1}) has a maximum value at the depletion layer edge, at $x = x_{\mathrm{n}}$, and decreases exponentially with x in the n region. This is a result of the recombination of the injected excess holes with the majority electrons in the n-quasineutral region before they reach the ohmic contact. Since the total current density (J_{total}) across the entire diode is invariant under steady-state conditions, the majority electron current (J_{n2}), which supplies electrons for recombination with holes, must increase with x away from the junction and reaches a maximum at the ohmic contact in the n-quasineutral region. Similarly, the minority electrons injected into the p-quasineutral region will contribute to the electron current flow (i.e., J_{n1}) in this region, and it can be derived in a similar way to that of the hole current density in the n-quasineutral region described above. Thus, the electron current density in the p-quasineutral region can be written as

$$J_{\mathrm{n}}(x) = q D_{\mathrm{n}} \frac{\mathrm{d}n'_{\mathrm{p}}(x)}{\mathrm{d}x} = \left(\frac{q D_{\mathrm{n}} n_{\mathrm{i}}^2}{N_{\mathrm{a}} L_{\mathrm{n}}} \right) (\mathrm{e}^{qV/k_{\mathrm{B}}T} - 1)\mathrm{e}^{(x+x_{\mathrm{p}})/L_{\mathrm{n}}}, \qquad (11.40)$$

which is obtained by assuming that the width of the p region is much larger than the electron diffusion length (i.e., $W_{\mathrm{E}} \gg L_{\mathrm{n}}$) in the p-quasineutral region. Note that x is negative in the p region and positive in the n region, and is equal to 0 at the metallurgical junction.

It is seen in Figure 11.6c that if the recombination current (i.e., J_{n4} or J_{p4}) in the depletion region is neglected, then the total current density in a p-n junction diode can be obtained by adding the injected minority hole current density evaluated at $x = x_{\mathrm{n}}$ and the injected minority electron current density evaluated at $x = -x_{\mathrm{p}}$. From (11.39) and (11.40) one obtains the total current density flow in a p-n junction as

$$J = J_{\mathrm{p1}}(x_{\mathrm{n}}) + J_{\mathrm{n1}}(-x_{\mathrm{p}}) = J_0(\mathrm{e}^{qV/k_{\mathrm{B}}T} - 1), \qquad (11.41)$$

where

$$J_0 = q n_{\mathrm{i}}^2 \left(\frac{D_{\mathrm{p}}}{N_{\mathrm{d}} L_{\mathrm{p}}} + \frac{D_{\mathrm{n}}}{N_{\mathrm{a}} L_{\mathrm{n}}} \right) \qquad (11.42)$$

is the saturation current density. Since J_0 is proportional to n_{i}^2, its value depends exponentially on the temperature and energy band gap of the semiconductor (i.e., $J_0 \propto n_{\mathrm{i}}^2 \propto \exp(-E_{\mathrm{g}}/k_{\mathrm{B}}T)$). For a silicon p-n junction diode, the value of J_0 will

double roughly for every 10°C increase in temperature. Equation (11.41) is known as the Shockley diode equation for an ideal p-n junction diode.[3]

Next consider the current flow in a short-base p-n junction diode, which has a base width W_B and an emitter width W_E much smaller than the minority carrier diffusion lengths (i.e., $W_B \ll L_p$) in the n-base region and (i.e., $W_E \ll L_n$) in the p-emitter region. In this case, the recombination loss in the p- and n-quasineutral regions is negligible, and hence the injected minority carriers are expected to recombine at the ohmic contact regions of the diode. It can be shown that the excess hole density in the n-base region of a short-base diode can be expressed by

$$p'_n(x) = p_{n0}(e^{qV/k_BT} - 1)\left[1 - \frac{(x - x_n)}{W'_B}\right], \tag{11.43}$$

where $W'_B = W_B - x_n$ is the width of the quasineutral n-base region. Equation (11.43) is obtained by replacing the exponential term in (11.38) by $[1 - (x - x_n)/W'_B]$, which was obtained from the boundary condition $p'_n(x) = 0$ at $x = W_B$. The boundary condition at $x = x_n$ is identical for both the short- and long-base diodes discussed above. Equation (11.43) predicts that the excess hole density in the n-base region decreases linearly with distance x. Thus, the hole current density can be derived from (11.43), and one has

$$J_p = -qD_p \frac{dp'_n}{dx}\bigg|_{x=x_n} = \left(\frac{qD_p n_i^2}{N_d W'_B}\right)(e^{qV/k_BT} - 1), \tag{11.44}$$

which shows that the hole current density in the n-base region is constant (i.e., the recombination loss in the base region is negligible). If the width of the p-emitter layer is smaller than the electron diffusion length (i.e., $W_E \ll L_n$), then the electron current density in the p^+-emitter region is given by

$$J_n = qD_n \frac{d'n_p}{dx}\bigg|_{x=-x_p} = \left(\frac{qD_n n_i^2}{N_a W'_E}\right)(e^{qV/k_BT} - 1)\,. \tag{11.45}$$

Therefore, the total current density for a short-base diode is equal to the sum of J_n and J_p, given by (11.44) and (11.45), which reads

$$J = J_n + J_p = qn_i^2\left(\frac{D_n}{N_a W'_E} + \frac{D_p}{N_d W'_B}\right)(e^{qV/k_BT} - 1). \tag{11.46}$$

Equation (11.46) shows that the current flow in a short-base diode is independent of the minority carrier diffusion lengths in the p and n regions of the diode, but varies inversely with the n- and p-layer thickness.

A comparison of the current density equations for a long-base diode and a short-base diode reveals that the preexponential factor for the former depends inversely on the minority carrier diffusion length, while the preexponential factor for the latter depends inversely on the thickness of the n and p regions of the diode. This is easy to understand, because for a long-base diode the width of the n-base region is much larger than the minority carrier diffusion length, and hence one

can expect that the hole current density in the n-base region will be influenced by the recombination loss of holes in the n-base region. However, this is not the case for the short-base diode, in which little or no recombination loss of holes in the n-base region is expected. It is, however, seen that both (11.41) and (11.46) predict the same exponential dependence of the current density on the applied bias voltage under forward-bias conditions and a very small saturation current density under reverse-bias conditions. It should be pointed out that under the reverse-bias condition, the saturation current density is contributed to by the thermal generation currents produced in both the n- and p-quasineutral regions of the junction. It is also noted that if one side of the junction is heavily doped, then the reverse saturation current will be determined by the thermal generation current produced on the lightly doped side of the junction. However, if the band gap narrowing and Auger recombination effects are taken into account in the heavily doped emitter region, then the saturation current density may be determined by the current flow in the heavily doped region of the junction.

The ideal diode analysis presented above is based on the assumption that the total current flow in a p-n junction diode is due solely to the diffusion current components produced in the n- and p-quasineutral regions. This approximation is valid as long as the recombination current in the junction space-charge region is negligible compared to the diffusion currents produced in the quasineutral regions. However, for a practical silicon p-n junction diode and p-n junction diodes fabricated from III-V compound semiconductors such as GaAs and InP, recombination in the junction space-charge region may become important and need to be considered. In this case, the ideal diode equation described above may be inadequate under small forward-bias conditions, and hence one needs to add the recombination current component (i.e., J_{n4} or J_{p4}) generated in the junction space-charge region to the total current density given by (11.41) for a long-base diode.

The generation-recombination current density in the junction space-charge region of a p-n diode can be derived using the Shockley–Read–Hall (SRH) model discussed in Chapter 6. For simplicity, it is assumed that the electron and hole capture cross-sections at the mid-gap recombination center are equal. Under this condition, the net recombination-generation rate for electrons and holes in the junction space-charge region is given by[2]

$$U_r = \frac{n_i^2(e^{qV/k_BT} - 1)}{[p + n + 2n_i \cosh(E_t - E_i)/k_BT]\tau_0}, \tag{11.47}$$

where E_t is the activation energy of the recombination center; E_i is the intrinsic Fermi level; $np = n_i^2 \exp(qV/k_BT)$ and $\tau_0 = 1/(N_t \upsilon_{th} \sigma)$ are used in (11.47). It is noted that the recombination rate given by (11.47) is positive under forward-bias conditions when the recombination process prevails, and becomes negative under reverse-bias conditions when the generation process is dominant in the junction space-charge region.

The total recombination-generation current density in the junction space-charge region can be obtained by integrating the recombination rate given in (11.47) over the entire depletion region from $x = 0$ to $x = W$, which is

$$J_{\mathrm{gr}} = q \int_0^W U_{\mathrm{r}} \, \mathrm{d}x. \tag{11.48}$$

Although the above integration cannot be readily carried out, it is possible to obtain an analytical expression for the recombination current density in the junction space-charge region if certain assumptions are made. For example, if one assumes that the recombination process is via a mid-gap trap center (i.e., $E_{\mathrm{t}} = E_{\mathrm{i}}$ and $n = p = n_{\mathrm{i}} \exp(qV/2k_{\mathrm{B}}T)$ for a maximum recombination rate, U_{max}), then the recombination current density under forward-bias conditions can be expressed as

$$J_{\mathrm{r}} = \frac{qW' n_{\mathrm{i}}^2 (\mathrm{e}^{qV/k_{\mathrm{B}}T} - 1)}{2n_{\mathrm{i}}\tau_0 (\mathrm{e}^{qV/2k_{\mathrm{B}}T} + 1)} \approx \left(\frac{qW' n_{\mathrm{i}}}{2\tau_0} \right) \mathrm{e}^{qV/2k_{\mathrm{B}}T}, \tag{11.49}$$

where $\tau_0 = (\tau_{\mathrm{n0}}\tau_{\mathrm{p0}})^{1/2}$ is the effective carrier lifetime associated with the recombination of excess carriers in the junction space-charge region of width W', and for $\exp(qV/2k_{\mathrm{B}}T) \gg 1$. It is interesting to note that if one calculates the ratio of the diffusion current and the recombination current components from (11.46) and (11.49), one finds that the recombination current component is important only in the small forward-bias regime, while the diffusion current becomes the dominant current component in the intermediate forward-bias regime.

Under reverse-bias conditions, the numerator in (11.47) reduces to $(-n_{\mathrm{i}}^2)$, and thus U_{r} becomes negative, which implies a net generation rate inside the junction space-charge region. The generation current density can be determined from the product of maximum generation rate and the depletion layer width W_{i}, namely,

$$J_{\mathrm{g}} = \int_0^{W_{\mathrm{i}}} qU \, \mathrm{d}x = q|U_{\mathrm{m}}|W_{\mathrm{i}} = \frac{qn_{\mathrm{i}}W_{\mathrm{i}}}{\tau_{\mathrm{e}}} \cong \left(\frac{n_{\mathrm{i}}}{\tau_{\mathrm{e}}} \right) \left[\left(\frac{q\varepsilon_0\varepsilon_{\mathrm{s}}}{2N_{\mathrm{d}}} \right) (V_{\mathrm{bi}} + V) \right]^{1/2}. \tag{11.50}$$

Equation (11.50) was obtained by assuming that the generation center coincides with the intrinsic Fermi level (i.e., $E_{\mathrm{t}} = E_{\mathrm{i}}$) and the depletion region is dominated by the lightly doped n region. The result shows that the generation current density varies linearly with the intrinsic carrier density, n_{i}, and the square root of the reverse-bias voltage.

It should be noted that the reverse saturation current density of a p-n junction diode is in general much smaller than that of the Schottky barrier diode discussed

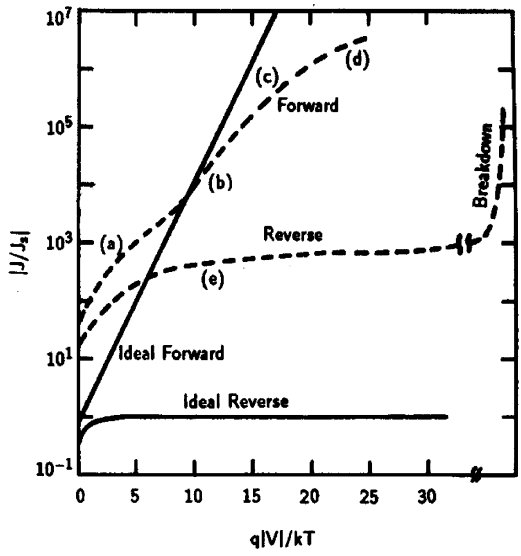


FIGURE 11.8. Current-voltage (I–V) characteristics of a practical silicon p-n diode; (a) generation-recombination (g-r) regime, (b) diffusion regime, (c) high-injection regime, (d) series resistance effect, (e) reverse leakage current due to g-r current and surface effects. After Moll,[3] by permission.

in Chapter 10. This is due to the fact that the saturation current of a p-n junction diode depends exponentially on the energy band gap of the semiconductor, while the saturation current of a Schottky diode depends exponentially on the barrier height. Since the barrier height is usually smaller than the energy band gap, the saturation current of a Schottky diode can be several orders of magnitude higher than that of a p-n junction diode under same temperature condition. Furthermore, one also expects that the saturation current of a p-n junction diode will have a stronger temperature dependence than that of a Schottky barrier diode due to the exponential dependence of the current density on both the temperature and band gap energy.

Figure 11.8 shows the I–V characteristics of a practical silicon p-n junction diode under forward and reverse-bias conditions.[3] The solid line corresponds to the ideal I–V curve predicted from the Shockley diode equation, while the dashed line corresponds to the I–V curve for a practical silicon p-n junction diode in which the recombination-current and series-resistance effects are also included under forward-bias conditions.

11.5. Diffusion Capacitance and Conductance

The transition capacitance derived in Section 11.3 is the dominant junction capacitance under reverse-bias conditions. However, under forward-bias conditions, when a small ac signal superimposed on a dc bias voltage is applied to a p-n junction diode, another capacitance component, known as the diffusion capacitance,

becomes the dominant component. This diffusion capacitance is associated with the minority carrier rearrangement in the quasineutral regions of the p-n junction under forward-bias conditions. The diffusion capacitance of a p-n junction diode under forward bias conditions can be derived using the small-signal time-varying voltage and current density equations:

$$V(t) = V_0 + v_1 e^{i\omega t}, \tag{11.51}$$

$$J(t) = J_0 + j_i e^{i\omega t}, \tag{11.52}$$

where V_0 and J_0 denote the dc bias voltage and the current density; v_1 and j_1 are the amplitude of the small-signal voltage and current density applied to the p-n junction, respectively. The small-signal condition is satisfied if $v_1 \ll k_B T/q$. When a small ac signal is applied to the junction, the minority hole density in the n-quasineutral region can be expressed as

$$p_n = p_{n0} \exp\left[\frac{q(V_0 + v_1 e^{i\omega t})}{k_B T}\right]. \tag{11.53}$$

Since $v_1 \ll V_0$, an approximate solution can be obtained by expanding the exponential term of (11.53), which yields

$$p_n \approx p_{n0} \exp\left(\frac{qV_0}{k_B T}\right)\left(1 + \frac{qv_1}{k_B T} e^{i\omega t}\right). \tag{11.54}$$

The first term in (11.54) is the dc component, while the second term corresponds to the small-signal component at the depletion layer edge of the n-quasineutral region. A similar expression for the electron density in the p^+-quasineutral region can also be derived. Substituting the ac component of p_n given by (11.54) into the continuity equation yields

$$D_p \frac{\partial^2 \tilde{p}_n}{\partial x^2} - \frac{\tilde{p}_n}{\tau_p} = i\omega \tilde{p}_n, \tag{11.55}$$

where

$$\tilde{p}_n = p_{n1} e^{i\omega t} = \left(\frac{p_{n0} q v_1}{k_B T}\right) \exp\left(\frac{qV_0}{k_B T}\right) e^{i\omega t}. \tag{11.56}$$

Now substituting (11.56) into (11.55), one obtains

$$\frac{\partial^2 \tilde{p}_n}{\partial x^2} - \frac{\tilde{p}_n}{D_p \tau_p^*} = 0, \tag{11.57}$$

where

$$\tau_p^* = \frac{\tau_p}{(1 + i\omega\tau_p)} \tag{11.58}$$

is the effective hole lifetime, which is frequency-dependent. The solution of (11.57) is given by

$$\tilde{p}_n = p_{n1} e^{-(x - x_n)/L_p^*}, \tag{11.59}$$

where $L_p^* = \sqrt{D_p \tau_p^*}$ is the effective hole diffusion length in the n-quasineutral region. Similar to the solution given by (11.41) for the dc current density, the solution for the ac hole current density is obtained by substituting (11.59) into (11.39) and evaluating the hole current density at $x = x_n$, which yields

$$j_p(x_n) = -qD_p \frac{dp_n}{dx}\bigg|_{x=x_n} = \left(\frac{qv_1}{k_B T}\right)\left(\frac{qD_p n_i^2}{N_d L_p^*}\right)\exp\left(\frac{qV_0}{k_B T}\right). \quad (11.60)$$

Similarly, the ac electron current density at $x = -x_p$ in the p^+-quasineutral region can be expressed as

$$j_n(-x_p) = qD_n \frac{dn_p}{dx}\bigg|_{x=-x_p} = \left(\frac{qv_1}{k_B T}\right)\left(\frac{qD_n n_i^2}{N_a L_n^*}\right)\exp\left(\frac{qV_0}{k_B T}\right). \quad (11.61)$$

The total ac current density is equal to the sum of $j_p(x_n)$ and $j_n(-x_p)$ given by (11.60) and (11.61), respectively, and can be written as

$$j_1 = j_p(x_n) + j_n(-x_p) = \left(\frac{qv_1}{k_B T}\right)\left(\frac{qD_p n_i^2}{N_d L_p^*} + \frac{qD_n n_i^2}{N_a L_n^*}\right)\exp\left(\frac{qV_0}{k_B T}\right). \quad (11.62)$$

The small-signal admittance (Y) of the p-n diode can be obtained from (11.62), and one obtains

$$Y = \frac{j_1}{v_1} = G_d + i\omega C_d = \left(\frac{q}{k_B T}\right)\left(\frac{qD_p n_i^2}{N_d L_p^*} + \frac{qD_n n_i^2}{N_a L_n^*}\right)\exp\left(\frac{qV_0}{k_B T}\right), \quad (11.63)$$

where $L_p^* = L_p/\sqrt{1 + i\omega\tau_p}$ and $L_n^* = L_n/\sqrt{1 + i\omega\tau_n}$ denote the effective hole- and electron-diffusion lengths, respectively. It is noted that both L_p^* and L_n^* depend on the frequency of the ac signals. At very low frequencies, $\omega\tau_{p,n} \ll 1$, the diffusion capacitance and conductance of a p-n diode can be obtained from (11.63), which yield

$$C_{d0} \approx \left(\frac{q^2 n_i^2}{2k_B T}\right)\left(\frac{L_p}{N_d} + \frac{L_n}{N_a}\right)\exp\left(\frac{qV_0}{k_B T}\right), \quad (11.64)$$

$$G_{d0} \approx \left(\frac{q^2 n_i^2}{k_B T}\right)\left(\frac{D_p}{N_d L_p} + \frac{D_n}{N_a L_n}\right)\exp\left(\frac{qV_0}{k_B T}\right). \quad (11.65)$$

Equation (11.63) shows that the diffusion capacitance varies inversely with the square root of the frequency and the minority carrier lifetimes, while the conductance increases with the square root of the frequency and the minority carrier lifetime. The small-signal analysis presented above for a p-n junction diode reveals that under forward-bias conditions the diffusion capacitance will become the dominant junction capacitance. It increases exponentially with the dc forward-bias voltage. Thus, the equivalent circuit of a p-n junction diode under small-signal operation should include both the transition and diffusion capacitances in parallel with ac-conductance and the series resistances that account for

the voltage drop across the ohmic contacts and the quasineutral regions of the diode.

11.6. Minority Carrier Storage and Transient Behavior

As discussed in the previous section, under forward-bias conditions, electrons are injected from the n-quasineutral region into the p-quasineutral region, while holes are injected from the p-quasineutral region into the n-quasineutral region. This will lead to a current flow and minority carrier storage in both the n- and p-quasineutral regions of the p-n junction. In this section, the minority carrier storage and transient behavior in a p-n junction diode are described.

Although in principle one could predict the transient behavior of minority carriers by solving the continuity equations, it is usually difficult to obtain an analytical solution by this approach. Fortunately, one can solve the problem more readily by using the charge-control method, as will be discussed next.

The total injected minority carrier charge per unit area stored in the n-quasineutral region can be found by integrating the excess hole density distribution across the n-quasineutral region. For a long-base diode, this is given by

$$\begin{aligned} Q'_p &= q \int_{x_n}^{W_B} p'_n(x)\, dx \\ &= q \int_{x_n}^{W_B} p_{n0}(e^{qV/k_BT} - 1)\, e^{-(x-x_n)/L_p} dx \\ &= q L_p p_{n0}(e^{qV/k_BT} - 1). \end{aligned} \tag{11.66}$$

Equation (11.66) shows that the minority carrier charge storage is proportional to both the minority carrier diffusion length and the minority carrier density at the depletion layer edge. The stored minority carrier charge (holes) given by (11.66) can be related to the hole injection current density given by (11.39) in the n-quasineutral region evaluated at $x = x_n$, namely

$$Q'_p = \left(\frac{L_p^2}{D_p}\right) J_p(x_n) = \tau_p J_p(x_n). \tag{11.67}$$

Equation (11.67) shows that the hole charge stored in the n-quasineutral base region is equal to the product of the hole lifetime and the hole current density. Thus, a long hole lifetime will result in more hole storage in the n-base region. This is expected since the injected holes can stay longer and diffuse deeper into the n-base region for long hole lifetime.

Similarly, the minority carrier storage in a short-base diode can be obtained by substituting (11.43) into (11.66), and using (11.44) for hole current density. This

yields

$$Q'_{\rm p} = \frac{q(W_{\rm B} - x_{\rm n})p_{\rm n0}}{2}(e^{qV/k_{\rm B}T} - 1) = \left[\frac{(W_{\rm B} - x_{\rm n})^2}{2D_{\rm p}}\right] J_{\rm p} = \tau_{\rm tr} J_{\rm p}, \tag{11.68}$$

which shows that for a short-base diode the minority carrier storage is not dependent on the minority carrier lifetime, but instead varies linearly with the average transit time $\tau_{\rm tr}$ across the n-base region. The term inside the square brackets of (11.68) denotes the average transit time for a hole to travel across the n-quasineutral region.

Another important diode parameter under forward-bias conditions that is associated with the minority carrier storage in the quasineutral regions is the diffusion capacitance. The diffusion capacitance per unit area for hole storage in the n-quasineutral region can be derived using the definition $C_{\rm d} = {\rm d}Q'_{\rm p}/{\rm d}V$, where $Q'_{\rm p}$ is given by (11.67) and (11.68) for long- and short-base diodes, respectively. Thus, the diffusion capacitance due to hole charge storage in the n-base region is given by

$$C_{\rm d} = \left(\frac{q^2 L_{\rm p} p_{\rm n0}}{k_{\rm B}T}\right) \exp\left(\frac{qV}{k_{\rm B}T}\right) \tag{11.69}$$

for the long-base diode, and

$$C_{\rm d} = \left(\frac{q^2(W_{\rm B} - x_{\rm n})p_{\rm n0}}{2k_{\rm B}T}\right) \exp\left(\frac{qV}{k_{\rm B}T}\right) \tag{11.70}$$

for the short-base diode. It is seen that the diffusion capacitance is important only under forward-bias conditions, and is negligible under reverse-bias conditions when the transition capacitance becomes the dominant component.

The transient behavior of the minority carrier storage in a p-n junction diode is very important when the diode is used in switching applications. This is because the switching time of a p-n diode depends on the amount of stored charge that must be injected and removed from the quasineutral regions of the diode. For example, one may shorten the switching time by reducing the stored charge in the quasineutral regions of the diode. This can be achieved by either reducing the minority carrier lifetime or by limiting the forward current flow in the diode. For switching applications the forward- to reverse-bias transition must be nearly abrupt, and the transit time must be short. In a switching diode the turnoff time is limited by the speed at which the stored holes can be removed from the n-quasineutral base region. When a reverse-bias voltage is suddenly applied across a forward-biased junction, the current can be switched in the reverse direction quickly. This is due to the fact that the gradient near the edge of the depletion region can make only a small change in the number of stored holes in the n-quasineutral region. Figure 11.9a shows a qualitative sketch of the transient decay of the excess stored holes in a long-base p-n diode. Figure 11.9b shows the basic switching circuit, and

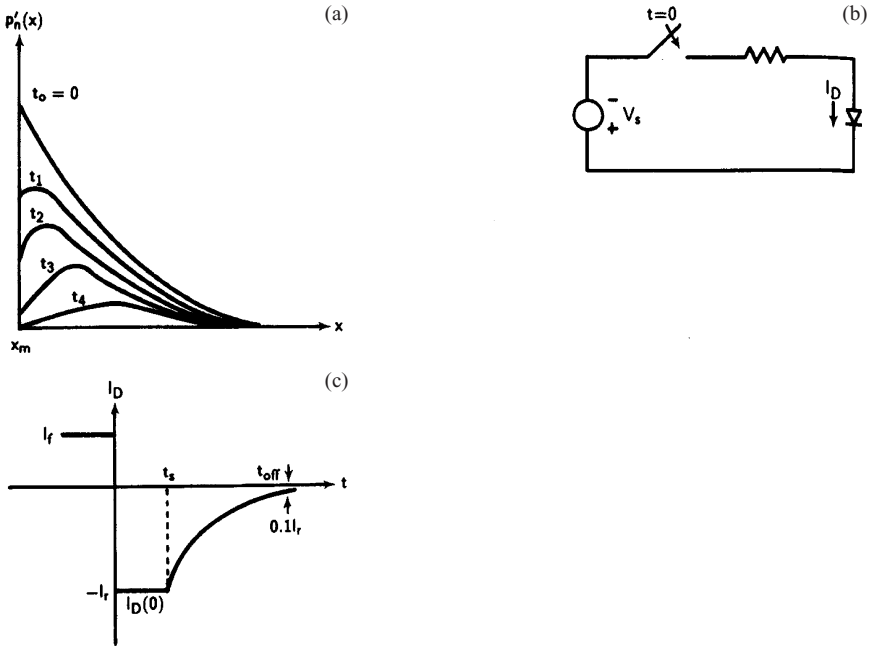


FIGURE 11.9. Transient behavior of a p-n junction diode: (a) transient decay of the minority hole density, (b) basic circuit diagram, and (c) transient response of the current from forward- to reverse-bias conditions.

Figure 11.9c displays the transient response of the current from the forward- to the reverse-bias conditions. It is seen that the turnoff time constant t_{off} shown in Figure 11.9c is the time required for the current to drop to 10% of the initial reverse current, I_r. This turnoff time can be estimated by considering a p^+-n junction diode under forward-bias conditions. In this case, the charge of the stored excess holes in the n-quasineutral region is given by

$$Q'_p = qA \int_{x_n}^{W_B} p'_n(x)\, dx, \tag{11.71}$$

where W_B is the n-base width, $p'_n(x)$ is the excess hole density in the n-base region, and A is the diode cross-sectional area. By integrating the continuity equation for the excess hole density given by (6.56) once from $x = x_n$ to $x = W_B$ and using (11.71), one obtains

$$I_p(x_n) - I_p(W_B) = \frac{dQ_s}{dt} + \frac{Q_s}{\tau_p}. \tag{11.72}$$

Equation (11.72) is known as the charge-control equation for a long-base diode. It is noted that $I_p(W_B)$ for a long-base diode can be set equal to 0. Thus, the steady-state forward-bias current can be obtained by setting $dQ_s/dt = 0$ in (11.72),

which yields

$$I_f = I_p(x_n) = \frac{Q_{sf}}{\tau_p}, \tag{11.73}$$

or

$$Q_{sf} = I_f \tau_p. \tag{11.74}$$

If the reverse-bias current is designated as I_r during the turnoff period, then (11.72) becomes

$$-I_r = \frac{dQ_s}{dt} + \frac{Q_s}{\tau_p}. \tag{11.75}$$

Using (11.73) as the initial condition, the solution of (11.75) is a time-dependent storage charge equation, which reads

$$Q_s(t) = \tau_p[-I_r + (I_f + I_r)e^{-t/\tau_p}]. \tag{11.76}$$

The turnoff time t_{off}, which is defined as the time required to move the minority holes out of the n-quasineutral region in order to reduce Q_s to zero, can be obtained by solving (11.76), which yields

$$t_{off} = \tau_p \ln\left(1 + \frac{I_f}{I_r}\right), \tag{11.77}$$

which shows that the turnoff time or switching time is directly proportional to the minority carrier lifetime and the ratio of the forward current to the reverse current in the diode. Thus, the switching speed of a p-n junction diode can be increased by shortening the minority carrier lifetimes in a n-p junction diode. Gold impurity is often used as an effective mid-gap recombination center in silicon switching diodes and transistors for reducing the minority carrier lifetimes and increasing the switching speed in these devices. Another approach, such as adding a Schottky barrier diode to the collector-base junction of a bipolar junction transistor (BJT) to form a Schottky-clamped BJT, has been widely used to reduce the minority carrier storage time in a switching transistor.

11.7. Zener and Avalanche Breakdowns

In this section, the junction breakdown phenomena in a p-n junction diode are described. As described in Section 11.2, the depletion layer width and the maximum electric field in the space-charge region of a p-n junction will increase with increasing reverse-bias voltage. Increasing the maximum field strength in the depletion region will eventually lead to junction breakdown phenomena commonly observed in a p-n junction diode under large reverse bias. There are two types of junction breakdown commonly observed in a p-n diode: Zener breakdown and avalanche breakdown. Zener breakdown occurs when valence electrons gain sufficient energy from the electric field and then tunnel through the

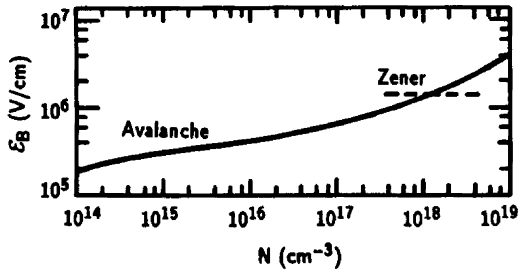


FIGURE 11.10. Critical electric fields for avalanche and Zener breakdowns in silicon as a function of dopant density. After Grove,[4] with permission by John Wiley & Sons, Inc.

forbidden gap into the conduction band. In this case electron–hole pairs are created by the large reverse-bias voltage, which results in a current flow. Avalanche breakdown is different from Zener breakdown in that the electric field is usually much higher. In avalanche breakdown, electrons (or holes) gain sufficient energy from the electric field and then engage in collisions. Between collisions of these high-energy electrons (or holes) they break the covalent bonds in the lattice and thus create more electron–hole pairs during the collisions. In this process, every electron (or hole) interacting with the lattice will create additional electrons (or holes), and all these electrons can participate in further avalanche collisions under high field conditions. This avalanche process will eventually lead to a sudden multiplication of carriers in the junction space-charge region where the maximum electric field becomes large enough to cause avalanche multiplication. It is noted that avalanche multiplication (or impact ionization) is probably the most important mechanism in junction breakdown, since the avalanche breakdown voltage imposes an upper limit on the reverse *I–V* characteristics of a p-n junction diode as well as other bipolar junction devices. Both Zener and avalanche breakdowns are nondestructive processes. Values of the breakdown voltage for each of these two processes depend on the junction structure and the doping concentration of the p-n junction. Figure 11.10 shows the critical electric fields for the avalanche and Zener breakdowns as a function of doping concentration in a silicon crystal.[4,5] Both of these breakdown phenomena are very important in practical device applications. The physical mechanisms and mathematical derivation of the avalanche and Zener breakdowns are given next.

Avalanche multiplication is an important mechanism in the junction breakdown phenomena because the avalanche breakdown voltage determines the maximum reverse-bias voltage that can be applied to a p-n junction without destroying the device. The avalanche multiplication mechanism has been widely used in achieving the internal current gain of an avalanche photodiode (APD) or to generate microwave power in an IMPATT diode.

The basic ionization integral, which determines the breakdown condition, can be derived as follows. As shown in Figure 11.11, consider the case in which impact ionization is initiated by electrons. The electron current $I_n(0)$ enters on the left-hand side (p region) of the depletion layer region of width equal to W, at

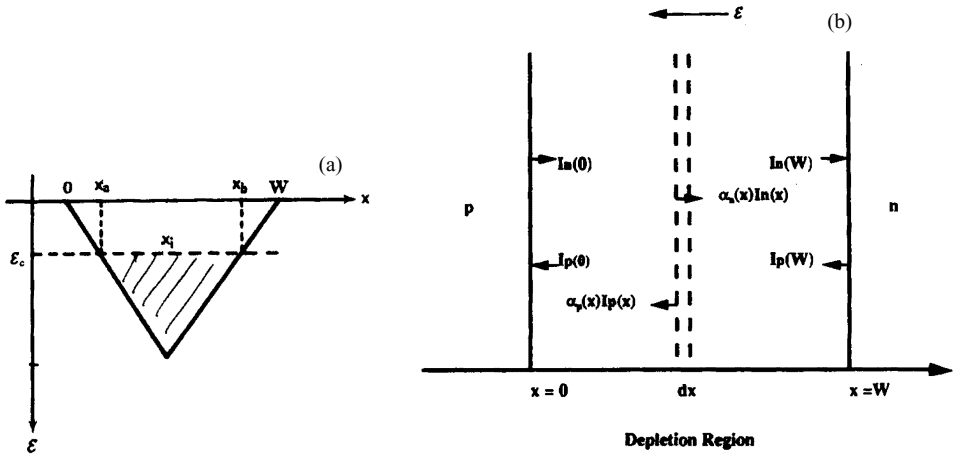


FIGURE 11.11. Schematic representation of (a) the electric field distribution and (b) the avalanche process in the space-charge region showing that ionization occurs in the hight field portion of the space-charge region (i.e., x_i).

$x = 0$. If the electric field in the depletion region is large enough (i.e., $\mathcal{E} \geq \mathcal{E}_c$), then electron–hole pairs will be created by impact ionization, and the electron current I_n will increase with distance through the depletion region, reaching a maximum value of $I_n(W) = M_n I_n(0)$ at $x = W$. Similarly, the hole current $I_p(x)$ will increase from $x = W$ to $x = 0$ as it moves through the depletion region from right to left in the junction space-charge region. Figure 11.11b shows the current flows due to the avalanche multiplication process of electrons and holes in the depletion region under large reverse-bias conditions. The total current $I = I_p(x) + I_n(x)$ is constant under steady-state conditions. The incremental electron current at x is equal to the number of electron–hole pairs generated per second in the interval dx, which is given by

$$d\left(\frac{I_n}{q}\right) = \left(\frac{I_n}{q}\right)\alpha_n dx + \left(\frac{I_p}{q}\right)\alpha_p \, dx, \tag{11.78}$$

or

$$\frac{dI_n}{dx} - (\alpha_n - \alpha_p)I_n = \alpha_p(I_n + I_p) = \alpha_p I, \tag{11.79}$$

where α_n and α_p denote the electron and hole ionization coefficients (cm^{-1}), respectively. If one introduces the boundary conditions $I_n(0) = I_{n0}$ at $x = 0$ and $I = I_n(W) = M_n I_{n0}$ at $x = W$, then the solution of (11.79) is given by

$$I_n(x) = \frac{I\{1/M_n + \int_0^x \alpha_p \, \exp[-\int_0^x (\alpha_n - \alpha_p)du]dx\}}{\exp[-\int_0^x (\alpha_n - \alpha_p) \, du]}, \tag{11.80}$$

where M_n is the multiplication factor of electrons, defined by

$$M_n = \frac{I_n(W)}{I_n(0)}. \tag{11.81}$$

Solving (11.80) and (11.81), one obtains the electron multiplication factor as

$$M_n = \frac{1}{\exp\left[-\int_0^w (\alpha_n - \alpha_p)\, dx\right] - \int_0^w \alpha_p \exp\left[-\int_0^x (\alpha_n - \alpha_p)\, du\right] dx}, \tag{11.82}$$

or

$$M_n = \frac{1}{1 - \int_0^w \alpha_n \exp\left[-\int_0^x (\alpha_n - \alpha_p) du\right] dx}. \tag{11.83}$$

Note that (11.83) is obtained by using the relation

$$\exp\left[-\int_0^w (\alpha_n - \alpha_p)\, dx\right] = 1 - \int_0^w (\alpha_n - \alpha_p)\, \exp\left[-\int_0^x (\alpha_n - \alpha_p)\, du\right] dx. \tag{11.84}$$

The avalanche breakdown voltage is referred to as the critical bias voltage in which the impact ionization occurs in the junction space-charge region and the multiplication factor M_n becomes infinity. When the avalanche multiplication process is initiated by the electron, the breakdown condition can be obtained from (11.83) with $M_n \to \infty$, which yields

$$\int_0^w \alpha_n\, \exp\left[-\int_0^x (\alpha_n - \alpha_p)\, du\right] dx = 1. \tag{11.85}$$

Similarly, if the avalanche multiplication is initiated by holes instead of electrons, then the ionization integral given by (11.85) becomes

$$\int_0^w \alpha_p\, \exp\left[-\int_0^x (\alpha_p - \alpha_n)\, du\right] dx = 1. \tag{11.86}$$

Equations (11.85) and (11.86) should yield the same breakdown condition within the depletion region of the diode regardless of whether the avalanche process is initiated by electrons or holes. For a semiconductor such as GaP that has equal ionization coefficients (i.e., $\alpha_n = \alpha_p = \alpha$), (11.85) and (11.86) can be simplified to

$$\int_0^W \alpha\, dx = 1. \tag{11.87}$$

If the ionization coefficients for both electrons and holes are independent of the position in the depletion region, then (11.82) becomes

$$M_n = \frac{(1 - \alpha_p/\alpha_n)\, \exp[(\alpha_n - \alpha_p)W]}{(1 - \exp[(\alpha_n - \alpha_p)W])}. \tag{11.88}$$

In general, the ionization coefficient α is a strong function of the electric field, since the energy necessary for an ionizing collision is imparted to the carriers by

the electric field. The field-dependent ionization coefficient can be expressed by an empirical formula given by

$$\alpha = A \ \exp\left(-\frac{B}{\mathcal{E}}\right), \tag{11.89}$$

where A and B are material constants; $\mathcal{E}$ is the electric field, which can be calculated for each material from the solution of Poisson's equation. For silicon, $A = 9 \times 10^5$ cm^{-1} and $B = 18 \times 10^6$ V/cm. It is noted that not only does the ionization coefficient vary with the electric field and the position in the depletion region, but the width of the depletion region also changes with the applied bias voltage. Thus, it is usually difficult to evaluate the avalanche multiplication factor M from (11.85) or (11.86). Instead, an empirical formula for M given by

$$M = \frac{1}{[1 - (V_R/V_B)^n]} \quad (2 < n < 6) \tag{11.90}$$

is often used. Here, V_R denotes the applied reverse-bias voltage, and V_B is the breakdown voltage given by

$$V_B = \frac{\mathcal{E}_m W}{2} = \frac{\varepsilon_0 \varepsilon_s \mathcal{E}_m^2}{2qN_B} \tag{11.91}$$

for a one-sided abrupt junction diode, and

$$V_B = \frac{2\mathcal{E}_m W}{3} = \left(\frac{4\mathcal{E}_m^{3/2}}{3}\right)\left(\frac{2\varepsilon_0 \varepsilon_s}{qa}\right)^{1/2} \tag{11.92}$$

for a linear-graded junction diode. It is noted that N_B is the background doping density in the lightly doped base region of the junction; a is the impurity gradient coefficient, and $\mathcal{E}_m$ is the maximum electric field in the junction space-charge region. An approximate universal expression for calculating the breakdown voltage as a function of energy band gap and doping density in an abrupt p-n junction diode is given by

$$V_B \approx 60\left(\frac{E_g}{1.1}\right)^{3/2}\left(\frac{N_B}{10^{16}}\right)^{-3/4}, \tag{11.93}$$

where E_g is the band gap energy in eV. For a linear-graded junction diode, the breakdown voltage is given by

$$V_B \approx 60\left(\frac{E_g}{1.1}\right)^{1.2}\left(\frac{a}{3 \times 10^{20}}\right)^{-0.4}. \tag{11.94}$$

Using (11.93), the breakdown voltage V_B for a silicon p^+-n step-junction diode with $N_d = 10^{16}$ cm^{-3} was found equal to 60 V at $T = 300$ K, and for a GaAs p^+-n diode with similar doping density the breakdown voltage V_B was found to be 75 V. Figure 11.12 shows the avalanche breakdown voltage versus impurity density for a one-sided abrupt junction and a linear-graded junction diode formed on Ge, Si, GaAs, and GaP, respectively.[5] The dashed line indicates the maximum

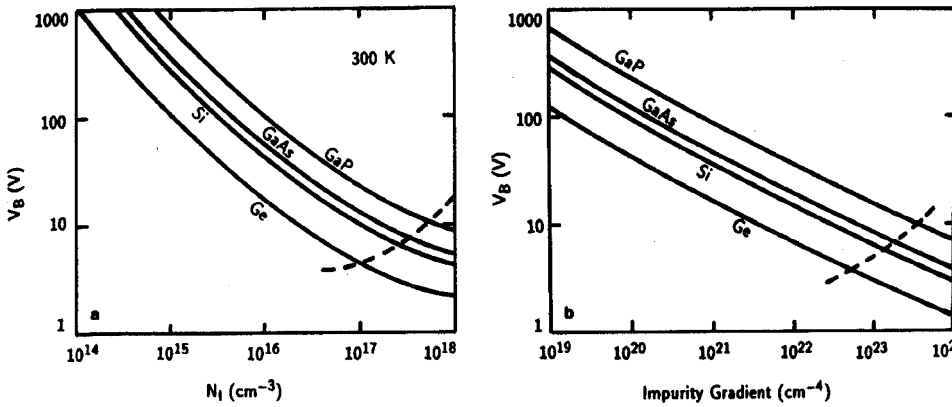


FIGURE 11.12. Avalanche breakdown voltage versus impurity density for (a) a one-sided abrupt junction and (b) a linearly graded diode in Ge, Si, GaAs, and GaP. The dashed line indicates the maximum doping density beyond which the tunneling mechanism will dominate the voltage breakdown characteristics. After Sze and Gibbons,[5] by permission.

doping density beyond which the tunneling mechanism will dominate the voltage breakdown characteristics. The Zener breakdown phenomenon in a p-n junction diode is discussed next.

As shown in Figure 11.10, when the doping density increases, the width of the space-charge region will decrease and the critical field at which avalanche breakdown occurs will also increase. At very high doping density, the electric field required for the avalanche breakdown to occur exceeds the field strength necessary for the Zener breakdown to take place, and hence the latter becomes more likely to occur. To explain the Zener breakdown mechanism, Figures 11.13a and b show the energy band diagram under reverse-bias conditions and the triangle potential barrier for a heavily doped p^+-n^+ junction diode, respectively. The probability for

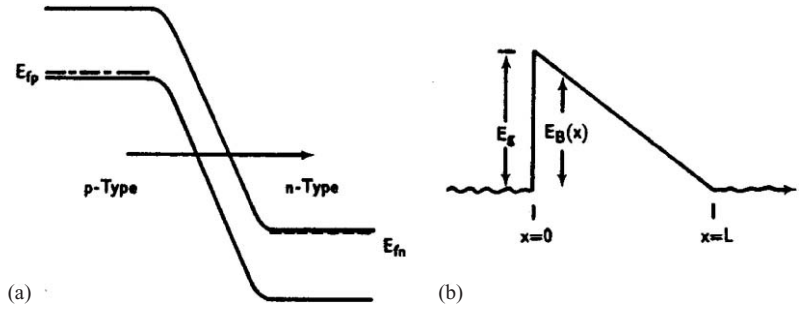


FIGURE 11.13. (a) Energy band diagram of a Zener diode under reverse-bias conditions. (b) The probability of tunneling across the junction is represented by tunneling through a triangle potential barrier.

electrons to tunnel from the valence band to the conduction band under high field conditions can be calculated using the tunneling of electrons through a triangular potential barrier. The energy barrier height, $E_B(x)$, decreases linearly from E_g at $x = 0$ to 0 at $x = L$. The probability of tunneling, T_x, can be derived from the WKB (Wentzel–Kramers–Brillouin)[6] approximation, which reads

$$T_x \approx \exp\left[-2\int_0^L \sqrt{\frac{2m^*}{\hbar^2}\left(E_g - q\mathcal{E}x\right)}\,dx\right]$$
$$= \exp(-B/\mathcal{E}) = \exp(-qBL/E_g), \tag{11.95}$$

where

$$B = \frac{4(2m^*)^{1/2}E_g^{3/2}}{3q\hbar}. \tag{11.96}$$

In (11.95), L is the tunneling distance, and $\mathcal{E} = E_g/qL$ is the average electric field in the junction space charge region. Therefore, the Zener tunneling probability decreases exponentially with decreasing electric field or increasing tunneling distance. If n is the number of valence electrons tunneling through the barrier, and v_{th} is the thermal velocity of electrons, then the tunneling current can be written as

$$I_t = Aqnv_{th}T_x, \tag{11.97}$$

where A is the cross-sectional area of the diode, and T_x is the tunneling probability given by (11.95). Equations (11.95) through (11.97) enables one to estimate the tunneling probability, the tunneling distance, and the electric field for a given tunneling current. It is seen that p-n diodes exhibiting Zener breakdown generally have a lower breakdown voltage than that of avalanche diodes. For example, in a silicon p-n junction diode with doping densities on both sides of the junction greater than 10^{18} cm^{-3}, Zener breakdown will occur at a voltage less than -6 V, while avalanche breakdown will occur at a much higher reverse-bias voltage.

11.8. Tunnel Diodes

In 1958, L. Esaki discovered a new device, known as the tunnel diode, when he observed a negative differential resistance and microwave oscillation in a heavily doped germanium p^{++}-n^{++} junction diode under forward-bias conditions. The current flow in a forward-bias tunnel diode can be attributed to the quantum-mechanical tunneling of charged carriers through the thin potential barrier across the junction.

A tunnel diode is formed when the densities of the shallow-donor and shallow-acceptor impurities in both the p^+ and n^+ regions of the junction are doped to the middle-10^{19} cm^{-3} range. Figure 11.14a shows the energy band diagram of a tunnel diode under equilibrium conditions ($V = 0$). Figure 11.14b shows the energy band diagram under a small forward-bias voltage with a triangle potential barrier height of $q\chi_B \approx E_g$. Figure 11.14c displays the forward current–voltage (I–V)

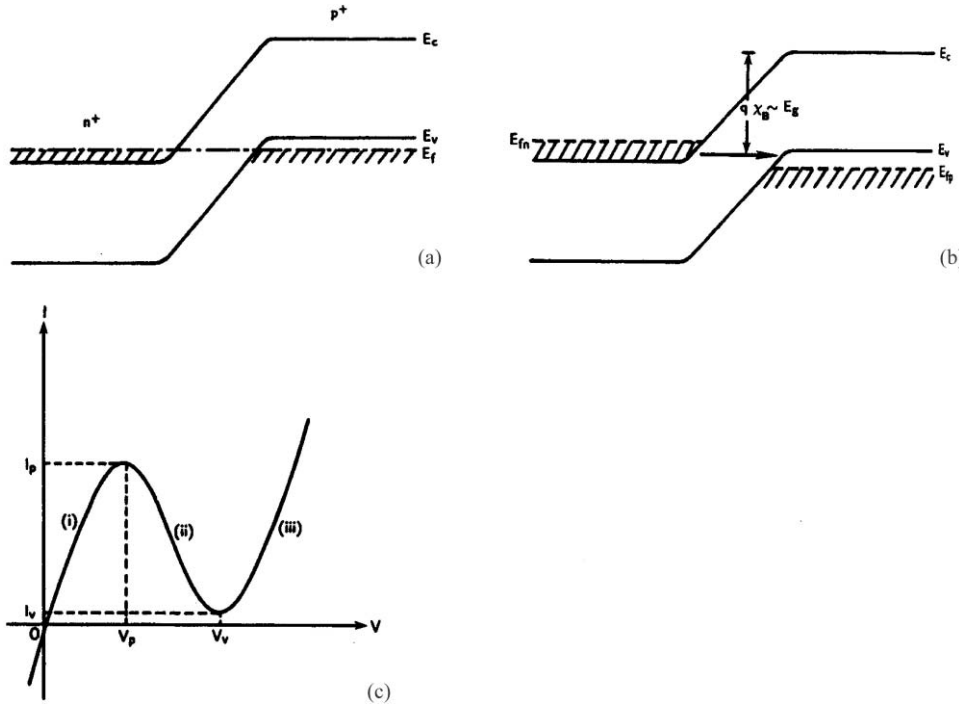


FIGURE 11.14. Energy band diagram (a) for a tunnel diode in equilibrium, (b) under forward-bias conditions, and (c) *I–V* characteristics under forward-bias conditions.

characteristics of a tunnel diode. Due to the high doping densities on both sides of the junction, the Fermi levels on either side of the junction are located a few $k_B T$ inside the conduction and valence bands, as shown in Figure 11.14a. For a tunnel diode, the depletion layer width under zero bias condition is on the order of 50 to 100 Å, which is much smaller than that of a standard p-n junction diode.

The electron tunneling process from the valence band to the conduction band, which is dominated in a tunnel diode under forward-bias condition, can be explained using the quantum-mechanical tunneling mechanism. As shown in Figure 11.14a, at $V = 0$ and $T = 0$ K, the states above the Fermi level in the conduction band of the n^+ region are empty, and the states below the Fermi level in the valence band of the p^+ region are completely filled. Therefore, under this condition no tunneling of electrons from the conduction band to the valence band will take place, and the tunneling current is equal to 0. This situation will usually prevail even at room temperature. When a forward-bias voltage is applied to the tunnel diode as shown in Figure 11.14b, the quasi-Fermi level in the n^+ region will move above the quasi-Fermi level in the p^+ region. As a result, it is possible for some of the electrons in the conduction band of the n^+ region to tunnel through the thin potential barrier across the junction into the empty states in the valence band of the p^+ region. The tunneling probability in this case will depend on the thickness

of the potential barrier across the junction, which will increase with decreasing barrier thickness.

A typical current–voltage (I–V) characteristic curve for a tunnel diode under forward-bias conditions is illustrated in Figure 11.14c, where I_p and V_p denote the peak current and peak voltage, while I_v and V_v are the valley current and valley voltage, respectively. The I–V characteristics under forward-bias conditions may be divided into three regions: (i) the low-bias (i.e., $V < V_p$) regime, where the current increases monotonically with voltage to a peak value I_p at voltage V_p; (ii) the intermediate-bias (i.e., $V_p < V < V_v$) regime, where the current decreases with increasing voltage to a minimum current I_v at voltage V_v; and (iii) the high-bias regime (i.e., $V > V_v$), where the current increases exponentially with applied voltage. In general, the current components contributing to the forward I–V characteristics of a tunnel diode shown in Figure 11.14c are dominated by the band-to-band tunneling current, the excess current, and the diffusion current. For $V < V_v$ the diffusion current is the dominant current component. In the negative resistance regime (i.e., regime (ii)) the current is dominated by the band-to-band tunneling through the thin triangle potential barrier of the junction.

Tunneling mechanisms and physical insight in a tunnel diode can be understood with the aid of a simple model using the triangle potential barrier shown in Figure 11.14b under forward-bias conditions. If the barrier height of the triangle potential barrier is assumed equal to the band gap energy (i.e., $\approx E_g$), and n is the density of electrons in the conduction band available for tunneling, then using the WKB method the tunneling probability of electrons across a triangle potential barrier is given by

$$T_t \approx \exp\left[-2\int_0^W |k(x)|\mathrm{d}x\right] \approx \exp\left(-\frac{4\sqrt{2qm_e^*}E_g^{3/2}}{3\hbar\varepsilon}\right), \tag{11.98}$$

where $|k(x)| = \sqrt{2m^*/\hbar^2(E_g/2 - q\mathcal{E}x)}$ is the absolute value of the electron momentum; $\mathcal{E} = E_g/W$ is the average electric field across the depletion region of the tunnel diode; W is the depletion layer width. It is seen in Figure 11.14b that if the states in the valence band of the p^+ region are mostly empty, then the tunneling current due to the band-to-band tunneling from n^+ to p^+ region is given by

$$I_t = Aqv_{th}nT_t, \tag{11.99}$$

where A is the cross-sectional area of the tunnel diode, v_{th} is the thermal velocity of the tunneling electrons, and T_t is the tunneling probability given by (11.98).

The tunneling current given by (11.99) is relatively insensitive to temperature. For example, the peak current of a typical germanium tunnel diode varies by only $\pm 10\%$ over a temperature range from -50 to 100°C. Since the tunneling time across a tunnel diode is very short, the switching speed of a tunnel diode is usually very fast. A wide variety of device and circuit applications using tunnel diodes, including microwave oscillators, multivibrators, low-noise microwave amplifiers, and high-speed logic circuits, have been reported in the literature. In addition to microwave and digital circuit applications, a tunnel diode can also be used as a test

vehicle in tunneling spectroscopy for studying fundamental physical parameters such as electron energy states in a solid and excitation modes in a p-n junction device. Finally, it should be noted that a tunnel diode is a two-terminal device, and hence it is not easy to incorporate such a device structure in many integrated circuit applications.

11.9. p-n Heterojunction Diodes

A p-n heterojunction diode can be formed using two semiconductors of different band gaps and with opposite doping impurities. Examples of p-n heterojunction diodes are Ge/GaAs, Si/SiGe, AlGaAs/GaAs, InGaAs/InAlAs, InGaP/GaAs, InGaAs/InP, and GaN/InGaN heterostructures. The heterojunction diodes offer a wide variety of important applications for laser diodes, light-emitting diodes (LEDs), photodetectors, solar cells, junction field-effect transistors (JFETs), modulation-doped field-effect transistors (MODFETs or HEMTs), heterojunction bipolar transistors (HBTs), quantum cascade lasers, quantum well infrared photodetectors (QWIPs), quantum dot lasers, and quantum dot infrared photodetectors. With recent advances in MOCVD and MBE epitaxial growth techniques for III-V compound semiconductors and SiGe/Si systems, it is now possible to grow extremely high-quality III-V heterojunction structures with layer thickness of 100 Å or less for quantum dots, superlattices, and multiquantum-well (MQW) device applications.

Figure 11.15a shows the energy band diagram for an isolated n-Ge and p-GaAs semiconductor in thermal equilibrium, and Figure 11.15b shows the energy band diagram of an ideal n-Ge/p-GaAs heterojunction diode. The energy band diagrams for ideal n-Ge/n-GaAs, p-Ge/n-GaAs, and p-Ge/p-GaAs heterostructures with no interface states are illustrated in Figures 11.16a, b, and c, respectively. Although the energy band gaps and dielectric constants for Ge and GaAs are quite different, the lattice constants for both materials are nearly identical (5.658 Å for Ge and 5.654 Å for GaAs). As a result, high-quality lattice-matched Ge/GaAs heterojunction structures can be formed in this material system. As shown in Figures 11.15b and 11.16, the energy band diagram for a heterojunction diode is much more complicated than that of a p-n homojunction due to the presence of energy band discontinuities in the conduction band (ΔE_c) and the valence band (ΔE_v) at the metallurgical junction of the two materials. In Figure 11.15b, subscripts 1 and 2 refer to Ge and GaAs, respectively; the energy discontinuity step arises from the difference of band gap and work function in these two semiconductors. The conduction band offset at the heterointerface of the two materials is equal to ΔE_c, and the valence band offset is ΔE_v. Based on the Anderson model,[6] the conduction and valence band offsets (ΔE_c and ΔE_v) can be obtained from the energy band diagram shown in Figure 11.15a, and are given, respectively, by

$$\Delta E_c = q(\chi_1 - \chi_2), \tag{11.100}$$

$$\Delta E_v = (E_{g2} - E_{g1}) - \Delta E_c = \Delta E_g - \Delta E_c, \tag{11.101}$$

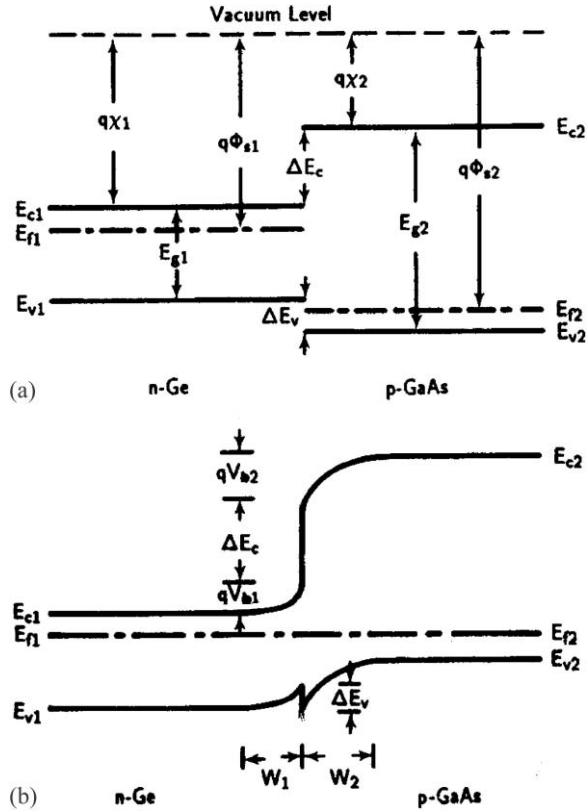


FIGURE 11.15. Energy band diagrams for (a) an isolated n-Ge and p-GaAs semiconductor in equilibrium, and (b) n-Ge and p-GaAs brought into intimate contact to form an n-p heterojunction diode.

which shows that the conduction band offset is equal to the difference in the electron affinity of these two materials, and the valence band offset is equal to the band gap difference minus the conduction band offset. From (11.101) it is noted that the sum of the conduction band and valence band offsets is equal to the band gap energy difference of the two semiconductors. When these two semiconductors are brought into intimate contact, the Fermi level (or chemical potential) must line up in equilibrium. As a result, electrons from the n-Ge will flow to the p-GaAs, and holes from the p-GaAs side will flow to the n-Ge side until the equilibrium condition is reached (i.e., the Fermi energy is lined up across the heterojunction). As in the case of a p-n homojunction, the redistribution of charges creates a depletion region across both sides of the junction. Figure 11.15b shows the energy band diagram for an ideal n-Ge/p-GaAs heterojunction diode in equilibrium, and the band offset in the conduction and valence bands at the Ge/GaAs interface is clearly shown in this figure. The band bending across the depletion region indicates that a built-in

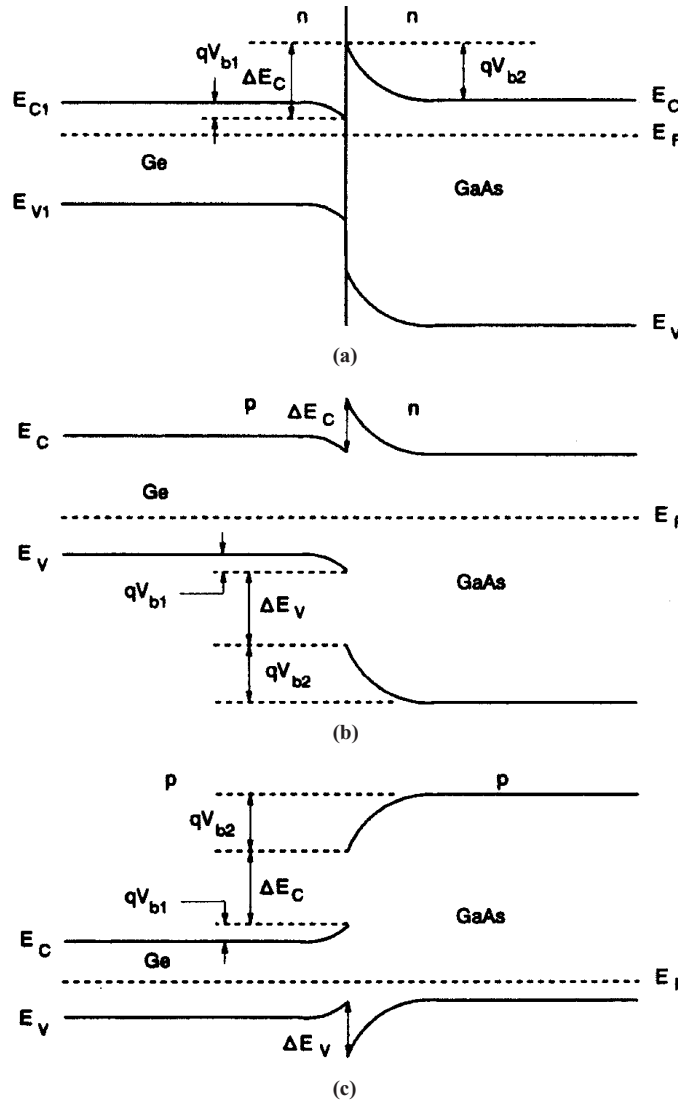


FIGURE 11.16. Energy band diagrams for (a) n-Ge/n-GaAs, (b) p-Ge/n- GaAs, and (c) p-Ge/p- GaAs heterojunction diodes in thermal equilibrium.

potential exists on both sides of the junction. The total built-in potential, V_{bi}, is equal to the sum of the built-in potentials on each side of the junction, i.e.,

$$V_{bi} = V_{b1} + V_{b2}, \tag{11.102}$$

where V_{b1} are V_{b2} are the band bending potentials in p-Ge and n-GaAs, respectively. It is noted that the discontinuity in the electrostatic field at the interface is due to the

difference in dielectric constants of these two semiconductors. Using the depletion approximation, V_{b1} and V_{b2} in the n-Ge and p-GaAs regions can be expressed, respectively, by

$$V_{b2} = \frac{\varepsilon_1 N_{D1}}{\varepsilon_1 N_{D1} + \varepsilon_2 N_{A2}} V_{bi}, \tag{11.103}$$

$$V_{b1} = \frac{\varepsilon_2 N_{A2}}{\varepsilon_1 N_{D1} + \varepsilon_2 N_{A2}} V_{bi}. \tag{11.104}$$

If one assumes the existence of a Schottky barrier at the heterointerface, the solution of Poisson's equation yields the depletion layer widths on either side of the step-heterojunction diode under bias conditions, which are given, respectively, by

$$W_1 = \left[\frac{2N_{A2}\varepsilon_1\varepsilon_2(V_{bi} - V)}{qN_{D1}(\varepsilon_1 N_{D1} + \varepsilon_2 N_{A2})}\right]^{1/2}, \tag{11.105}$$

$$W_2 = \left[\frac{2N_{D1}\varepsilon_1\varepsilon_2(V_{bi} - V)}{qN_{A2}(\varepsilon_1 N_{D1} + \varepsilon_2 N_{A2})}\right]^{1/2}. \tag{11.106}$$

Thus, the total depletion layer width of the heterojunction can be obtained from (11.105) and (11.106) as

$$W_d = W_1 + W_2,$$

$$W_d = \left[\frac{2\varepsilon_1\varepsilon_2(V_{bi} - V)(N_{A2}^2 + N_{D1}^2)}{q(\varepsilon_1 N_{D1} + \varepsilon_2 N_{A2})N_{D1}N_{A2}}\right]^{1/2}. \tag{11.107}$$

Equations (11.105) and (11.106) are derived from Poisson's equation in the depletion region of the heterojunction diode. The two boundary conditions are given, respectively, by

$$W_1 N_{D1} = W_2 N_{A2}, \tag{11.108}$$

$$\varepsilon_1 \mathcal{E}_1 = \varepsilon_2 \mathcal{E}_2, \tag{11.109}$$

where ε_i and $\mathcal{E}_i$ $(i = 1, 2)$ denote the dielectric constants and electric fields in regions 1 and 2, respectively. From (11.103) and (11.104), the ratio of the relative voltage drop across regions 1 and 2 of the two semiconductors is given by

$$\frac{(V_{b1} - V_1)}{(V_{b2} - V_2)} = \frac{N_{A2}\varepsilon_2}{N_{D1}\varepsilon_1}. \tag{11.110}$$

The transition capacitance per unit area for the p-n heterojunction can be derived from (11.107), and one obtains

$$C_j = \left[\frac{qN_{D1}N_{A2}\varepsilon_1\varepsilon_2}{2(\varepsilon_1 N_{D1} + \varepsilon_2 N_{A2})(V_{bi} - V)}\right]^{1/2} = \sqrt{\frac{\varepsilon_1\varepsilon_2}{W_d}}. \tag{11.111}$$

The current–voltage (I–V) characteristics for an n-Ge/p-GaAs heterojunction diode can be derived using the energy band diagram shown in Figure 11.15b and the thermionic emission theory for a Schottky barrier diode described in Chapter 10. Since the relative magnitudes of the current components in a heterojunction are determined by the potential barriers involved, for the n-Ge/p-GaAs heterojunction

diode shown in Figure 11.15b the hole current from p-GaAs to n-Ge is expected to dominate the current flow because of the lower potential barrier ($= V_{b2}$) for hole injection and the higher potential barrier ($= V_{b1} + \Delta E_c + V_{b2}$) for electron injection. Therefore, to derive the current–voltage relationship for the n-p heterojunction diode shown in Figure 11.15b, only the hole current need be considered. At zero bias, the barrier to hole flow from p-GaAs to n-Ge is equal to qV_{b2}, and in the opposite direction it is $(\Delta E_v - qV_{b1})$. Under thermal equilibrium conditions, the two oppositely directed fluxes of holes must be equal, since the net current flow is zero. Thus, one can write

$$A_1 \exp[-(\Delta E_v - qV_{b1})/k_B T] = A_2 \exp(-qV_{b2}/k_B T), \tag{11.112}$$

where constants A_1 and A_2 depend on the doping levels and carrier effective masses in the diode.

If one applies a forward-bias voltage V_a across the junction, then the portions of the voltage drops on the two sides of the junction are determined by the relative doping densities and dielectric constants of the materials, and are given, respectively, by

$$V_2 = K_2 V_a, \quad \text{where } K_2 = \frac{N_{D1}\varepsilon_1}{N_{D1}\varepsilon_1 + N_{A2}\varepsilon_2} \tag{11.113}$$

and

$$V_1 = K_1 V_a, \quad \text{where } K_1 = 1 - K_2. \tag{11.114}$$

The energy barriers are equal to $q(V_{b2} - V_2)$ on the p-GaAs side and $[\Delta E_v - q(V_{bi} - V_1)]$ on the n-Ge side of the junction. Using (11.112), the net hole flux from right to the left under forward-bias conditions can be expressed by

$$\phi_p = A_1 \exp(-qV_{b2}/k_B T)[\exp(qV_2/k_B T) - \exp(qV_1/k_B T)]. \tag{11.115}$$

If the conduction mechanism is governed by thermionic emission, the current density due to hole injection from p-GaAs to n-Ge of the n-p heterojunction diode shown in Figure 11.15b has a similar form to that given by (11.115). Thus, the hole current density can be written as

$$\begin{aligned} J_p &= A \exp(-qV_{b2}/k_B T)\,[\exp(qV_2/k_B T) - \exp(qV_1/k_B T)] \\ &\approx J_0 \left(1 - \frac{V_a}{V_{bi}}\right)[\exp(qV_a/k_B T) - 1], \end{aligned} \tag{11.116}$$

where

$$J_0 = \left(\frac{qA^* T V_{bi}}{k_B}\right) e^{-qV_{bi}/k_B T}. \tag{11.117}$$

It is noted that (11.116) was obtained using the approximation $e^{q(V_{b1}-V_1)/k_B T} \approx (q/k_B T)(V_{bi} - V_a)$ and the total applied voltage $V_a = V_1 + V_2$. From (11.116), it is seen that the current–voltage (I–V) relationship for an n-p heterojunction diode is somewhat different from that of a metal–semiconductor Schottky diode. The main difference is that the reverse saturation current density is not a constant,

but increases linearly with the applied reverse-bias voltage for $V_a \gg V_{bi}$. Under forward-bias conditions, as in the case of a Schottky barrier diode, the current–voltage relationship can be approximated by an exponential dependence of the form e^{-qV_a/nk_BT}, where n is the diode ideality factor.

Practical applications of heterojunction structures for a wide variety of devices such as solar cells, LEDs, laser diodes, photodetectors, HEMTs, and HBTs will be discussed further in Chapters 12, 13, and 16. Finally, it should be mentioned that multilayer heterojunction structures such as superlattices, multiple quantum wells and quantum dots, and the modulation-doped heterostructures grown using MBE and MOCVD techniques have been widely reported for photonic and high-speed device applications. Figure 11.17 shows (a) the energy band diagrams for a modulation-doped $Al_{0.3}GA_{0.7}As/GaAs$ heterostructure and a single-period $Al_{0.3}Ga_{0.7}As/GaAs$ heterostructure, (b) a comparison of the electron mobilities as a function of temperature in the undoped GaAs quantum well and in several bulk GaAs samples of diferent doping concentrations, and (c) a comparison of electron mobilities versus temperature in the triangle potential well of undoped GaAs with a single-period modulation-doped AlGaAs/GaAs heterostructure and in bulk GaAs doped to 10^{17} cm^{-3}. Using a modulation-doping technique, the electron mobility in the undoped GaAs quantum well can be greatly enhanced because the impurity scattering due to ionized donor impurities in the AlGaAs layer can be eliminated in the undoped GaAs quantum well. Typical layer thickness for the modulation-doped heterostructure devices is around 100 Å. The modulation-doping technique has been widely used in heterojunction field-effect transistors such as HEMTs for high-speed device applications. Other applications using superlattices and multiple quantum well heterostructures include resonant tunneling devices and quantum well infrared photodetectors (QWIPs), LEDs and laser diodes (LDs). These devices will be discussed further in Chapter 12, 13, and 16.

11.10. Junction Field-Effect Transistors

The junction field-effect transistor (JFET) is a three-terminal device, which consists of the source, the gate, and the drain electrode. In a JFET the lateral current flow between the source and drain electrodes is controlled by the applied vertical electric field via a controlled gate, which is formed by a p^+-n junction. Figure 11.18 shows the basic device structure of an n-channel JFET formed on a lightly doped n-type epilayer grown on a p-type substrate. The heavily doped n^+-source–drain regions and the p^+-gate region can be formed using either the thermal diffusion or ion implantation technique. In IC fabrication, the ion-implantation technique is more widely used since better control of geometries, doping densities, and profiles for both the source and drain regions can be obtained using this technique.

Since current flow in a JFET is due to the majority carriers in the channel that is formed between the p^+-gate and the p-substrate, the JFET is also known as a unipolar transistor. The unique feature of a JFET is that the conductivity in the

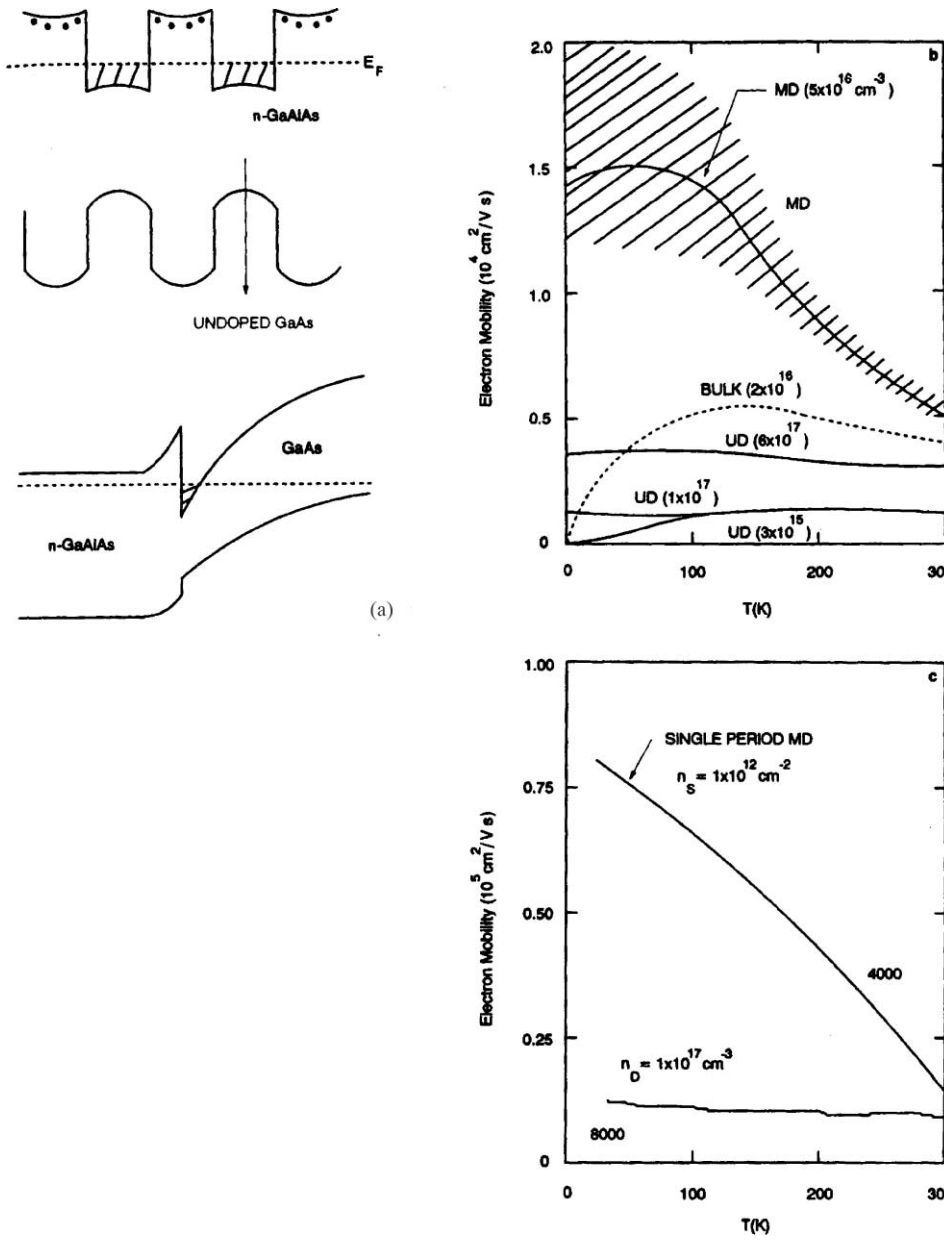


FIGURE 11.17. (a) Energy band for an AlGaAs/GaAs modulation-doped heterostructure, (b) electron mobility versus temperature for the modulation-doped structure and bulk GaAs with different doping concentrations, and (c) electron mobility versus temperature for the single-period modulation-doped AlGaAs/GaAs heterostructure. After Dingle et al.,[7] by permission.

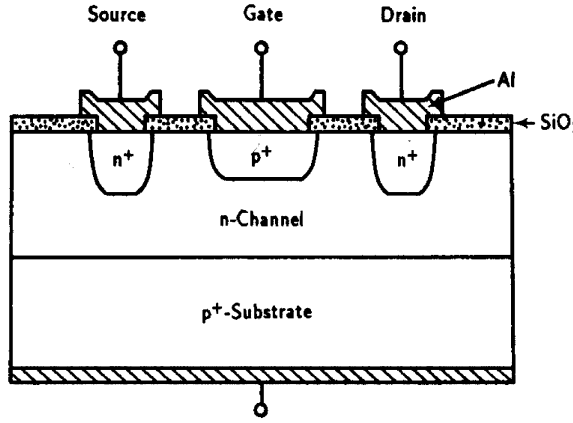


FIGURE 11.18. The device structure of an n-channel junction field-effect transistor (JFET).

channel can be controlled by the reverse-bias voltage at the gate electrode. The gate bias voltage is used to change the depletion layer width in the p^+-gate/n-channel space-charge region. If the current flow in the channel is due to electrons, then one has an n-channel JFET. On the other hand, if the current flow in the channel is due to holes, then one has a p-channel JFET (in this case the source and drain electrodes are p^+ doping, and the substrate is n-type).

The dc characteristics of a JFET can be analyzed using a one-dimensional (1-D) JFET structure as shown in Figures 11.19a and b under different bias conditions. In this figure, L is the channel length between the source and drain, Z is the depth of the channel, $2a$ is the channel width, and the drain current is along the x-direction of the channel length. If the channel length (L) is much larger than the channel width ($2a$), then the change in channel width along the channel is small compared to the channel width. Therefore, the electric field in the depletion region of the gate junction is assumed perpendicular to the channel (i.e., along the y-direction), while the electric field inside the neutral n-channel may be assumed in the x-direction only. The gradual-channel approximation was first introduced by Shockley to analyze the current–voltage (I_D–V_D) characteristics of a JFET. By assuming a one-side abrupt junction at the gate region with its doping density N_A much larger than N_D in the channel, the depletion layer will extend mainly into the channel region. Under normal operating conditions, a reverse-bias voltage is applied across the gate electrode so that free carriers are depleted from the channel and the space-charge region extends into the channel. Consequently, the cross-sectional area of the channel is reduced, and channel conduction is also reduced. Thus, the current flow in the channel is controlled by the gate voltage. The depletion layer width in this case can be expressed as

$$W_d(x) = \sqrt{\frac{2\varepsilon_0\varepsilon_s[V(x) + V_{bi} - V_g]}{qN_D}}. \tag{11.118}$$

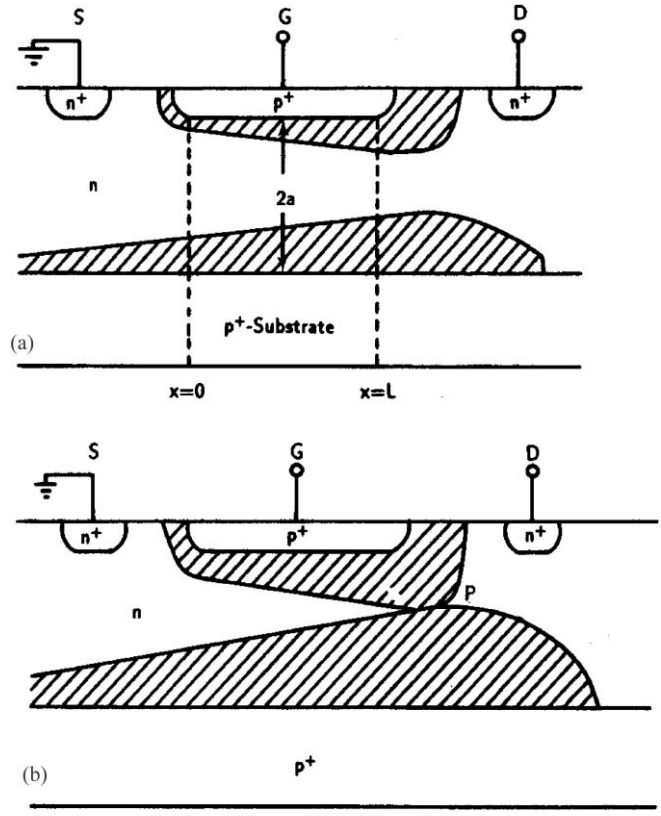


FIGURE 11.19. Schematic diagram of an n-channel JFET showing (a) the source (S), gate (G), and drain (D) regions, the dimensions of the channel, and the depletion region (shaded area) in the channel under small gate bias voltage and (b) at pinch-off condition.

The resistance in the channel region can be expressed by

$$R = \frac{\rho L}{A} = \frac{L}{q\mu_{\mathrm{n}} N_{\mathrm{D}} A} = \frac{L}{2q\mu_{\mathrm{n}} N_{\mathrm{D}} Z(a - W_{\mathrm{d}})}. \tag{11.119}$$

The drain current I_{D} in the n-channel is due to the drift component only, and is given by

$$I_{\mathrm{D}} = Aqn\mu_{\mathrm{n}}\mathcal{E}_x = 2q\mu_{\mathrm{n}} N_{\mathrm{D}}[a - W_{\mathrm{d}}(x)]\, Z\, \frac{\mathrm{d}V}{\mathrm{d}x}, \tag{11.120}$$

where $A = 2(a - W_{\mathrm{d}})Z$ is the cross-sectional area of the channel. Substituting (11.118) for $W_{\mathrm{d}}(x)$ into (11.120) and integrating the equation from $x = 0$ to $x = L$ with corresponding voltages from 0 to V_{D} yields

$$I_{\mathrm{D}} \int_0^L \frac{\mathrm{d}x}{2q\mu_{\mathrm{n}} N_{\mathrm{D}} Z} = \int_0^{V_{\mathrm{D}}} \left\{ a - \left[\left(\frac{2\varepsilon_0 \varepsilon_{\mathrm{s}}}{qN_{\mathrm{D}}} \right) (V + V_{\mathrm{bi}} - V_{\mathrm{g}}) \right]^{1/2} \right\} \mathrm{d}V. \tag{11.121}$$

Now integrating and rearranging the terms in (11.121), one obtains

$$I_D = G_0 \left\{ V_F - \frac{2}{3} \left(\frac{2\varepsilon_0 \varepsilon_s}{qa^2 N_D} \right)^{1/2} [(V_d + V_{bi} - V_g)^{3/2} - (V_{bi} - V_g)^{3/2}] \right\}. \tag{11.122}$$

Equation (11.122) is a general expression of the current–voltage (I_D–V_D) relation for a JFET; it can also be expressed in terms of the pinch-off voltage and pinch-off current as

$$I_D = I_p \left\{ \left(\frac{V_D}{V_p} \right) - \frac{2}{3} \left[\frac{(V_D + V_g + V_{bi})}{V_p} \right]^{3/2} + \frac{2}{3} \left[\frac{(V_g + V_{bi})}{V_p} \right]^{3/2} \right\}, \tag{11.123}$$

where

$$I_p = \frac{q^2 \mu_n N_D^2 Z a^3}{\varepsilon_0 \varepsilon_s L}, \tag{11.124}$$

$$V_p = \frac{q N_D a^2}{2\varepsilon_0 \varepsilon_s}, \tag{11.125}$$

where V_p is the pinch-off voltage (i.e., $V_p = V_D + V_g + V_{bi}$, and $W_d = a$ at $x = L$) and I_p is the pinch-off current.

The *I–V* characteristics for a JFET can be analyzed in two regions with pinch-off as the boundary condition. At low drain voltages (i.e., $V_D \ll V_g + V_{bi}$), (11.123) becomes

$$I_D = G_0 \left\{ 1 - \left[\frac{\varepsilon_0 \varepsilon_s}{2q N_d (a - W_d)(V_{bi} - V_g)} \right]^{1/2} \right\} V_D, \tag{11.126}$$

where $G_0 = 2q\mu_n N_d Z(a - W_d)/L$ is the channel conductance. Equation (11.126) shows that at a given gate voltage, a linear relationship between I_D and V_D prevails in this region. The inverse square-root dependence of the drain current on the gate voltage is a direct result of assuming an abrupt junction for the gate-channel junction. It is seen in (11.126) that the drain current reaches a maximum value when the gate voltage is zero and decreases with increasing gate voltage. In addition, this equation also predicts zero drain current when the gate voltage is large enough to deplete the entire channel region. If the drain voltage is further increased, the depletion layer width will also increase. Eventually, the two depletion regions touch each other at the drain electrode as shown in Figure 11.19b. This pinch-off condition occurs when the depletion width W_d is equal to a at the drain electrode. For a p^+-n junction, solving (11.118) yields the corresponding value of the drain voltage, which is given by

$$V_{DS} = \frac{q N_D a^2}{2\varepsilon_0 \varepsilon_s} - V_{bi} \quad \text{for} \quad V_g = 0, \tag{11.127}$$

where V_{DS} is the saturation drain voltage. The pinch-off condition is reached at this drain voltage, and both the source and drain regions are completely separated by a

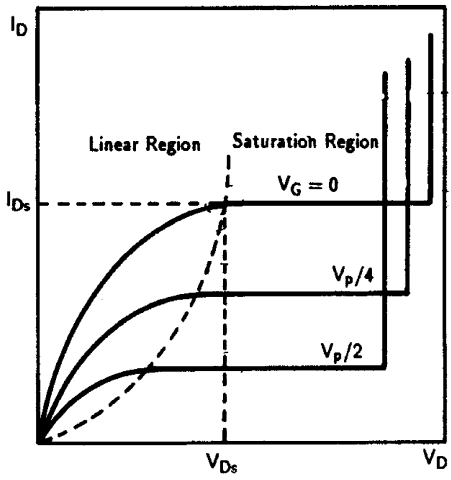


FIGURE 11.20. Output I–V characteristics of a JFET for different gate bias voltages, showing the linear and saturation regions of the device operation.

reverse-bias depletion region. The location of point P in Figure 11.19b is called the pinch-off point, and the corresponding drain current is called the saturation drain current I_{DS}, which can flow through the depletion region. Beyond the pinch-off point, as V_D is increased further, the depletion region near the drain will expand and point P will move toward the source region. However, the voltage at point P remains the same as V_{DS}. As a result, the potential drop in the channel from the source to point P remains the same, and the current flow in the channel also stays constant. Thus, for drain voltages larger than V_{DS}, the current flow in the channel is independent of V_D and is equal to I_{DS}. Under this condition, the JFET is operating in the saturation regime, and the expression for the saturation current can be deduced from (11.122) in the form

$$I_{DS} = G_0 \left\{ \frac{qN_D a^2}{6\varepsilon_0 \varepsilon_s} - (V_{bi} - V_g) \left[1 - \frac{2}{3} \left(\frac{2\varepsilon_0 \varepsilon_s (V_{bi} - V_g)}{qN_D a^2} \right)^{1/2} \right] \right\}, \tag{11.128}$$

or

$$I_{DS} = I_p \left\{ \frac{1}{3} - \frac{(V_g + V_{bi})}{V_p} + \frac{2}{3} \left[\frac{V_g + V_{bi}}{V_p} \right]^{3/2} \right\}. \tag{11.129}$$

The corresponding drain saturation voltage is given by

$$V_{DS} = V_p - V_g - V_{bi}. \tag{11.130}$$

Based on the above analysis, the I_D versus V_D curve can be divided into three different regimes: (1) the linear regime at low drain voltages, (2) a regime with less than a linear increase of drain current with drain voltage, and (3) a saturation regime where the drain current remains constant as the drain voltage is further increased. This is illustrated in Figure 11.20.

The JFETs are often operated in the saturation regime in which the output drain current does not depend on the output drain voltage but depends only on the input

gate voltage. Under this condition, the JFET may be used as an ideal current source controlled by an input gate voltage. The transconductance of a JFET can be obtained by differentiating (11.122) with respect to gate voltage, which yields

$$\begin{aligned} g_{\mathrm{m}} &\equiv \left.\frac{\partial I_{\mathrm{D}}}{\partial V_{\mathrm{g}}}\right|_{V_{\mathrm{D}}=\text{constant}} \\ &= G_0\left(\frac{2\varepsilon_0\varepsilon_{\mathrm{s}}}{qN_{\mathrm{D}}a^2}\right)^{1/2}[(V_{\mathrm{bi}}-V_{\mathrm{g}}+V_{\mathrm{D}})^{1/2}-(V_{\mathrm{bi}}-V_{\mathrm{g}})^{1/2}] \quad (11.131) \\ &= \left(\frac{I_{\mathrm{p}}}{V_{\mathrm{p}}}\right)\left\{1-\left[\frac{(V_{\mathrm{g}}+V_{\mathrm{bi}})}{V_{\mathrm{p}}}\right]^{1/2}\right\}. \end{aligned}$$

It is noted from (11.131) that in the saturation regime, the transconductance has a maximum value given by

$$g_{\mathrm{ms}} = G_0\left\{1-\left[\frac{2\varepsilon_0\varepsilon_{\mathrm{s}}}{qN_{\mathrm{D}}a^2}(V_{\mathrm{bi}}-V_{\mathrm{g}})\right]^{1/2}\right\}. \quad (11.132)$$

Theoretical analysis presented in this section for a JFET is based on several simplified assumptions. For example, it is assumed that the depletion layer width is controlled solely by the gate-channel junction and not by the channel–substrate junction. In reality, there will be a variation in potential across the channel–substrate junction along the channel, with maximum potential and depletion width occurring near the drain region. As a result, this simplified assumption may lead to a disagreement between the theoretical predictions and experimental data on I_{D}–V_{D} characteristics of a practical JFET. In general, the simple model presented here is valid only for a long-channel JFET device. For a short-channel JFET with $L/a < 2$, the saturation mechanism becomes more complex and the above theories require refinement in order to obtain good agreement between the theory and experiment.

Based on the above theoretical analysis, it is clear that the dc characteristics of a JFET are usually quite sensitive to doping density and thickness of the channel region. Therefore, a precise control of thickness and doping density in the channel region of a JFET is very important. An n-channel silicon JFET can be made with excellent control using the epitaxial growth technique. As for the source, drain, and gate regions, because the density and location of the doping impurity in these regions can be controlled better using ion implantation rather than thermal diffusion, the ion-implantation approach is preferable to the thermal-diffusion method. In fact, both the n-channel and p-gate regions for the silicon JFET are usually formed using the ion-implantation technique.

Problems

11.1. Derive the electric field, potential distribution, and depletion layer width for a linearly graded p-n junction diode, and show that the depletion layer

width and the built-in potential under zero-bias conditions are given by (11.22) and (11.23), respectively.

11.2. If the impurity gradient a for a Si and a GaAs linearly graded p-n junction diode is equal to 10^{22} cm^{-4} at 300 K, calculate the depletion layer width, built-in potential, and breakdown voltage for these two diodes. Repeat for $a = 10^{20}$ cm^{-4}. Calculate and plot junction capacitance versus applied voltage for both diodes.

11.3. The small-signal ac characteristics of a p-n junction diode are important for circuit applications. The diode admittance can be obtained by solving the small-signal carrier distribution from the steady-state continuity equation. The small-signal condition is satisfied if the applied ac signal, v_1, is small compared to the thermal voltage, V_T ($= k_B T/q$). Draw a small-signal equivalent circuit for the diode by including the circuit elements r_s, G_d, C_d, and C_j, where r_s is the series resistance caused by the ohmic drop across the neutral semiconductor regions and the contacts, $G_d = I/V_T$ is the small-signal conductance, $C_d \approx \tau_p I/2V_T$ is the diffusion capacitance, and C_j is the transition capacitance that arises from the junction space-charge layer. Note that $I = I_p(0)$ is the dc current density for a p$^+$-n junction diode.

11.4. From the results obtained in Problem 11.3, calculate the small-signal conductance and capacitance for a long-base silicon p$^+$-n diode if $N_A = 5 \times 10^{18}$ cm^{-3}, $N_D = 2 \times 10^{16}$ cm^{-3}, $\tau_n = 2 \times 10^{-8}$ s and $\tau_p = 5 \times 10^{-8}$ s, $A = 2 \times 10^{-4}$ cm^2, and $T = 300$ K.

(a) For forward-bias voltages $V_f = 0.1$, 0.3, 0.5, and 0.7 V.

(b) For reverse-bias voltages $V_R = -0.5$, -5, -10, and -20 V.

(c) What is the series resistance of the n-neutral region if the thickness is equal to 2 mm?

11.5. Consider the minority carrier charge storage effect in a long-base p$^+$-n diode.

(a) Show that the turnoff time t_{off} of holes in the n region is given by

$$t_{off} = \tau_p \ln\left(1 + \frac{I_f}{I_r}\right) . \tag{1}$$

The above equation is obtained from the charge-control equation for a long-base p$^+$-n diode. It can be shown that an exact analysis by solving the time-dependent diffusion equation would yield

$$\text{erf}\left(\frac{t_{off}}{\tau_p}\right)^{1/2} = \frac{I_f}{(I_f + I_r)} , \tag{2}$$

where erf (x) is the error function.

(b) Plot t_{off}/τ_p versus I_r/I_f using expressions (1) and (2) given above.

11.6. Under forward-bias conditions, the space-charge recombination current can be calculated from the Shockley–Read–Hall (SRH) model via a mid-gap recombination center. The recombination rate derived from the SRH model

is given by

$$U_r = \frac{n_i^2(e^{qV_a/k_BT} - 1)}{[p + n + 2n_i \cosh (E_t - E_i)/k_BT]\tau_0}, \tag{1}$$

where $\tau_0 = 1/N_t\sigma v_{th}$ is the effective carrier lifetime.

(a) Find the conditions of maximum recombination rate from (1).

(b) If the recombination current in the junction space-charge region can be derived from

$$J_r = q\int_{-x_p}^{x_n} U_r \, dx \ , \tag{2}$$

where U_r is given by (1), and x_n and x_p are the depletion layer widths in the n and p regions, respectively, show that the recombination current density can be expressed by

$$J_r = \left(\frac{qW'n_i}{2\tau_0}\right) e^{qV_a/2k_BT} \ , \tag{3}$$

where W' is the portion of the depletion region in which the recombination current is dominant, which is valid for $qV_a > 3k_BT$.

11.7. Consider a long-base silicon p^+ -n diode. If the diode parameters are given by $N_A = 10^{19}\,cm^{-3}$, $N_D = 5 \times 10^{16}\,cm^{-3}$, $D_p = 4\,cm^2/s$, $D_n = 20\,cm^2/s$, $\tau_p = 10^{-8}$ s, $\tau_n = 10^{-6}$ s, $n_i = 1.4 \times 10^{10}\,cm^{-3}$, and $A = 10^{-4}\,cm^2$:

(a) Calculate the hole injection current into the n region for forward-bias voltages of 0.1, 0.3, 0.5, and 0.7 V at 300 K.

(b) Repeat (a) for the electron injection current into the p^+-n region.

(c) What is the total injection current for $V = 0$, 0.3, and 0.5 V if the silicon p-n diode is a short-base diode with p-emitter width of $W_E = 5 \times 10^{-5}$ cm and n-base width of $W_B = 10^{-3}$ cm?

11.8. If the silicon p-n diode given in Problem 11.7 has a mid-gap recombination center in the junction space-charge region with large defect density, calculate the recombination current for the cases $N_t = 10^{13}$, 10^{14}, and 10^{15} cm^{-3} and $V = 0.2$, 0.4, and 0.6 V assuming that capture cross-section $\sigma = 10^{-15}\,cm^2$, $v_{th} = 10^7$ cm/s, and recombination occurs throughout the entire depletion region. (*Hint*: use the expression for the recombination current given by Problem 11.6.)

11.9. The diode parameters for a Ge-, Si-, and GaAs- p^+-n short-base diode are the same as those given by Problem 11.7, except that the intrinsic carrier densities are 2.5×10^{13}, 9.65×10^9, and $10^6\,cm^{-3}$ for Ge, Si, and GaAs at 300 K, respectively. Calculate the total injection current for these diodes. Explain why GaAs is more suitable for high-temperature applications than Si and Ge. (Given: $E_g = 0.67.1.12$, and 1.43 eV for Ge, Si, and GaAs at 300 K.)

11.10. (a) Plot the energy band diagram for an $Al_xGa_{1-x}As$/GaAs p^+-n heterojunction for $x = 0.3$ ($E_g = 1.8$ eV).

(b) If $N_A = 5 \times 10^{18}$ cm^{-3}, $N_D = 2 \times 10^{17}$ cm^{-3}, $D_n = D_p = 5$ cm^2/s, and $\tau_n = \tau_p = 10^{-9}$ s, calculate the injection currents in both regions of the diode, assuming $A = 2 \times 10^{-5}$ cm^2.

(c) Plot the injection current versus temperature (100–400 K) for this diode.

11.11. The onset of Zener breakdown in an abrupt silicon p-n diode takes place when the maximum electric field approaches 10^6 V/cm. If the doping density in the p region is 5×10^{19} cm^{-3}, what would be the doping density in the n region in order to achieve a Zener breakdown voltage of 2 V? Repeat the calculation for $N_A = 10^{20}$ cm^{-3} and $V_{ZB} = 3$ V.

11.12. Plot the energy band diagrams for a p^+-$Al_{0.3}Ga_{0.7}As$/n-GaAs and an n-Ge/p-GaAs heterojunction diode. The band gap energy for $Al_xGa_{1-x}As$ can be calculated using the equation $E_g = 1.424 + 1.247x$eV (for $0 < x < 0.45$). Calculate the values of ΔE_c and ΔE_v for both cases.

References

1. W. Shockley, *Bell Syst. Tech. J.*, **28**, 435 (1949); *Electrons and Holes in Semiconductors*, D. Van Nostrand, Princeton, N. J. (1950).
2. C. T. Sah and R. N. Noyce, *Proc. IRE* **45**, 1228 (1957).
3. J. L. Moll, *Proc. IRE* **46**, 1076 (1958).
4. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Chapter 6, Wiley, New York (1967), p. 192.
5. S. M. Sze and G. Gibbons, *Appl. Phys. Lett*. **8**, 111 (1966).
6. R. L. Anderson, *Solid State Electron*. **5**, 341 (1962).

Bibliography

R. Dingle, H. L. Stormer, A. C. Gossard, and W. Wiegmann, *Appl. Phys. Lett*. **33**, 665 (1978).
W. R. Frensky and H. Kroemer, *Phys. Rev. B* **16**, 2642 (1977).
R. N. Hall, *Phys. Rev*. **87**, 387 (1952).
A. G. Miles and D. L. Feucht, *Heterojunctions and Metal–Semiconductor Junctions*, Academic Press, New York (1972).
J. L. Moll, *Physics of Semiconductors*, McGraw-Hill, New York (1964).
R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, Wiley, New York (1977).
C. T. Sah, R. N. Noyce, and W. Shockley, *Proc. IRE* **45**, 1228 (1957).
W. Shockey, *Electrons and Holes in Semiconductors*, Van Nostrand, Princeton (1950).
W. Shockley and W. T. Read, *Phys. Rev*. **87**, 835 (1952).
S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1981).
E. S. Yang, *Microelectronic Devices*, McGraw-Hill, New York (1988).

12
Solar Cells and Photodetectors

12.1. Introduction

Photonic devices play an important role in a wide variety of applications in the areas of photovoltaic (PV) power generation, optical communications, data transmission and signal processing, detection, sensors and optical imaging, and displays and light sources. Recent advances in III-IV compound semiconductor growth and processing technologies have enabled these applications to become a reality. As a result, various photonic devices such as laser diodes (LDs), light-emitting diodes (LEDs), solar cells, and photodetectors using III-V semiconductors have been developed for use in power generation, optical communications, displays and solid-state light sources, data transmission, and signal processing. Depending on the device structures and operating modes, photonic devices can in general be divided into three categories: (i) PV devices (i.e., solar cells), which convert sunlight directly into electricity by generating electron–hole pairs in a solar cell via internal PV effect, (ii) photodetectors, which detect photons or optical signals and convert them into electrical signals via internal photoelectric effects, and (iii) LEDs and LDs, which convert electrical energy into incoherent (for LEDs) and coherent (for LDs) optical radiation by electrical injection into the junction region of a p-n junction diode. In this chapter, the basic device physics and structures, the operation principles, and the general characterstics of solar cells and photodetectors fabricated from elemental and compound semiconductors will be depicted.

The solar cell, which utilizes the internal PV effect in a semiconductor, will be discussed first. Solar cells may be formed using a p-n junction, a Schottky barrier, or a metal–insulator–semiconductor (MIS) structure fabricated on various semiconductor materials. The basic device physics, cell structures and characteristics, design criteria, and performance limitations for different types of solar cells are described in Section 12.2. It is interesting to note that prior to 1973, a majority of solar cell research was focused mainly on the development of silicon p-n junction solar cells for space applications. However, in recent years most of the efforts have been shifted toward the development of various low-cost and high-efficiency solar cells for both terrestrial and space power generation as well as for consumer electronics applications. It is well known that for

terrestrial applications, cost and conversion efficiency are the two key factors that determine the viability and compatibility of the PV system with other types of power generation systems using fossil fuel, nuclear, hydrogen fuel cell, windmill, and geothermal technologies. Recent advances in several thin-film PV technologies using amorphous silicon (a-Si) thin films, Cu(ln,Ga)Se_2(CIGS), and CdTe absorber materials show excellent potential of meeting both low-cost and high-efficiency criteria for large-scale terrestrial power generation. On the other hand, multijunction tandem solar cells using III-V compound semiconductors with different band gaps and concentrator solar cells using a multijunction approach have achieved much higher conversion efficiency than a single-junction solar cell. For example, a mechanically stacked lnGaP/GaAs/lnGaAs 3-junction solar cell has achieved a conversion efficiency of 33.3% at 1-sun AM1.5G conditions, and a concentrator operation of InGaP/InGaAs/Ge 3-junction solar cell has demonstrated a world-record 36% efficiency at 100-sun AM1.5G conditions. Large-band-gap III-V compound semiconductor materials such as GaAs and InGaP are particularly attractive for concentrator solar cell applicaitons since they can be operated at a much higher temperature than that of a silicon solar cell. Conversion efficiency over 20% AM1.5G has been achieved in a single-junction solar cell fabricated from CIGS,Si,GaAs and InP material systems.

Photodetectors, which employ the internal photoelectric effects to detect photons in a semiconductor device, are presented in Section 12.3. A p-n junction or a Schottky barrier photodiode can be very fast and sensitive when operating under reverse-bias conditions. If sensitivity is the main concern, then an avalanche photodiode (APD) may be used to obtain the necessary internal current gain and quantum efficiency. The APD has achieved the highest-gain bandwidth product among all photodetectors. On the other hand, a silicon p-i-n photodiode can offer both sensitivity and speed in the visible to near-infrared (IR) spectral range. In fact, various photodetectors covering a broad range of wavelengths from ultraviolet (UV) to visible, near-IR, mid-wavelength infrared (MWIR), long-wavelength (LWIR), and far-infrared spectral ranges have been developed for a wide variety of applications. For example, GaN and SiC Schottky barrier and p-i-n photodiodes have been developed for solar-blind and UV light detection. Extremely high sensitivity photomultipliers are commercially available for photon counting in the visible to near-IR (0.3–0.9 μm) spectral range, while p-i-n photodiodes and APDs fabricated from Si, GaAs, InGaAs, InGaAsP, and Ge cover the wavelength ranges from UV, visible, to near-IR (0.4–1.8 μm). Quantum-well infrared photodetectors (QWIPs) and $Hg_xCd_{1-x}Te$ (MCT) photoconductors and p-n diodes have been developed for 3–5 μm (MWIR) and 8–12 μm (LWIR) detection, while extrinsic photoconductors (impurity-band photoconductors such as As- or Sb-doped Si, and Cu-doped Ge photoconductors) can extend the detection wavelengths into the far-IR spectral regime (e.g., $\lambda > 30$ μm). Large-format (e.g., 640 × 480), highly uniform GaAs/AlGaAs QWIP focal plane arrays (FPAs) with excellent noise equivalent temperature difference (NEDT) of a few tens of mK have been developed for IR imaging camera applications in the 8–12 μm atmospheric spectral window. Multicolor QWIP FPAs using InGaAs/AlGaAs and

GaAs/AlGaAs material systems have also been demonstrated for MWIR and LWIR imgaing array applications.

12.2. Photovoltaic Devices (Solar Cells)

12.2.1. Introduction

Although practical solar cells have become available only since the mid-1950s, scientific investigation of the PV effect started as early as in 1839, when the French scientist Henri Becquerel discovered that an electric current was produced by shining a light onto certain chemical solutions. The PV effect was first observed in the metal selenium in 1877. This material was used for many years in light meters that required only very small amounts of electric power. A more detailed understanding of the basic principles, provided by Einstein in 1905 and Schottky in 1930, was required before efficient solar cells could be made. The first silicon solar cell with a conversion efficiency of 6% AM0 was demonstrated by Chapin, Pearson, and Fuller in 1954, which was used primarily in specialized applications for orbiting space satellites. Today, single- and multicrystalline silicon solar cells with 1-sun conversion efficiencies ranging from 14.7% to around 25% AM1.5G have been demonstrated using different fabrication processing steps and device structures. There are several competing PV technologies available for the production of commercial PV modules for terrestrial power generation and consumer electronics applications. They are (i) silicon solar cell modules made from single-crystal and polycrystalline silicon, (ii) low-cost thin-film solar cell modules fabricated from a-Si: H, Cu(In,Ga)Se_2(CIGS), and CdTe materials, and (iii) high-efficiency multijunction tandem cells and concentrator solar cells using III-V compound semiconductors such as InGaP/GaAs and InGaP/GaAs/Ge material systems for cell fabrication. These solar cells and solar cell modules can be used in a wide variety of applications for consumer electronics, office and residential systems, remote irrigation systems and relay stations, and remote village and off-grid industrial systems.

A solar cell using a p-n junction or a Schottky barrier structure can convert sunlight directly into electricity. In order to calculate the conversion efficiency of a solar cell, one needs to know the exact incident solar irradiance power under different insolution conditions. Figure 12.1 shows the solar irradiance spectra for two air-mass (AM) conditions.[1] The top curve is the solar irradiance spectrum measured above the earth's atmosphere, and is defined as the air-mass zero (AM0) insolution. The irradiant power of the sun under AM0 conditions is 136.61 mW/cm^2 (or 1,366.1 W/m^2). The bottom curve is the solar irradiance spectrum measured under AM1 conditions. The AM1 solar spectrum represents the sunlight on the earth's surface when the sun is at its zenith. The total incident power of sunlight under AM1 condition is 92.5 mW/cm^2. The AM1.5G (global) condition is the solar irradiance of both diffuse and direct components that are incident on a sun-facing 37° tilted surface, and has an average incident power of 100 mW/cm^2. This is the

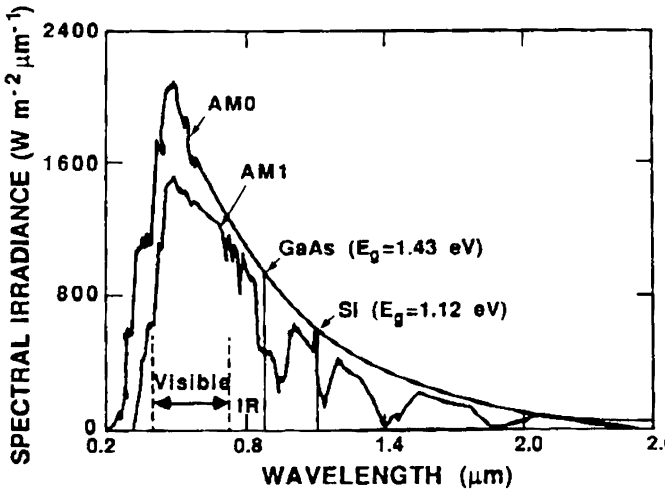


FIGURE 12.1. Solar irradiance versus wavelength under air-mass zero (AM0) and air-mass one (AM1) conditions. Also shown are the energy band gaps and the corresponding cutoff wavelengths for both GaAs and Si. After Thekaekara,[1] by permission.

most suitable incident solar irradiance power for calculating the conversion efficiency of a solar cell in the terrestrial environment because this is representative of conditions in the 48 contiguous states of the United States. Since the conversion efficiency will vary under different AM conditions, it is important to specify the exact AM (i.e., AMx; $x = 0$, 1, 1.5, or 2) condition in calculating the conversion efficiency of a solar cell.

In this section, different types of solar cells using p-n junction, Schottky barrier, MIS, p-n heterojunction, concentrator, and multijunction tandem solar cell structures are described. The generation and collection of photogenerated electron–hole pairs in a p-n junction solar cell are discussed first, followed by the derivation of equations for spectral response (quantum efficiency, η_q), short-circuit current (I_{sc}), open-circuit voltage (V_{oc}), and conversion efficiency (η_c) of a p-n junction solar cell. Formation of the front ohmic contact grids and antireflection (AR) coatings for a p-n junction solar cell will be discussed. Key PV technologies based on crystalline and polycrystalline silicon solar cells, amorphous silicon, $CuInGaSe_2$ and CdTe polycrystalline thin-film solar cells, and multijunction and concentrator solar cells fabricated from III-V compound semiconductors for terrestrial and space power generation will be presented.

12.2.2. Device Physics and General Characteristics of a p-n Junction Solar Cell

Since most solar cells use a p-n junction structure, it is important to consider the basic device physics and electrical characteristics of a p-n junction solar cell.

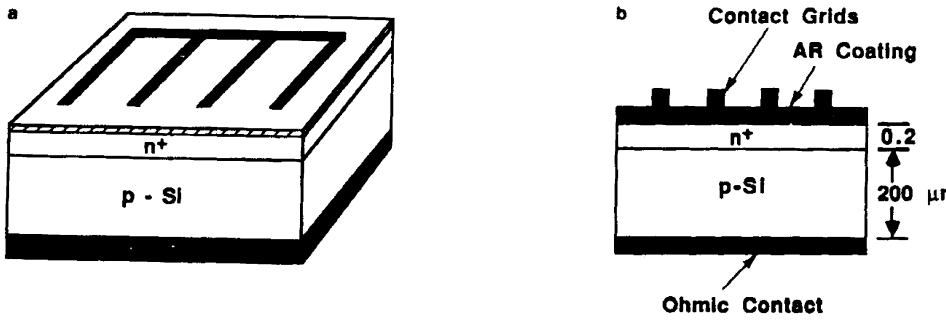


FIGURE 12.2. (a) A typical cell structure and (b) the cross-sectional view of a silicon n^+-p junction solar cell. Also shown are the front ohmic contact grids and the antireflection (AR) coating layer (e.g., Ta_2O_5).

Topics to be covered include current–voltage (*I–V*) characteristics, the derivatiion of photocurrent and quantum efficiency expressions, and analysis of the performance parameters of a p-n junction solar cell. Figure 12.2 shows (a) the cell structure and (b) the cross-sectional view of a typical n^+-p junction solar cell. The basic characteristics of an n-p junction solar cell are obtained by analyzing the *I–V* behavior under dark and illumination conditions. The key performance parameters such as the open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), fill factor (FF), and conversion efficiency (η_c) can be determined from the photo-*I–V* characteristics of a solar cell. To explain the basic operational principles of a p-n junction solar cell, Figure 12.3 shows the energy band diagrams, carrier generation, dark current components, *I–V* characteristics, and the equivalent circuit of an n^+-p junction solar cell under dark and illumination conditions. Figure 12.3a shows the electron–hole pairs generated by the absorbed photons in different regions of the solar cell, and Figure 12.3b shows the dark current components generated in different regions of the junction. The *I–V* curves under dark and illumination conditions are illustrated in Figure 12.3c. It is noted that the shaded area in the fourth quadrant of the photo- *I–V* curve represents the power generated in the solar cell. The equivalent circuit of a p-n junction solar cell is shown in Figure 12.3d, where R_s denotes the series resistance and R_p is the shunt resistance. The dark and photo-*I–V* characteristics of a p-n junction solar cell are discussed next.

(i) Dark I–V characteristics. As shown in Figure 12.3b, the dark current of a p-n junction solar cell under forward-bias conditions consists of three components: (1) the injection current due to injection of majority carriers across the p-n junction, (2) the recombination current due to the recombination of electrons and holes via deep-level traps in the junction space-charge region, and (3) the tunneling current due to multistep tunneling via deep-level defect states in the junction space-charge region. In a silicon p-n junction solar cell, the injection current is usually the dominant component. However, the recombination current can become a dominant component for solar cells fabricated from low-quality materials

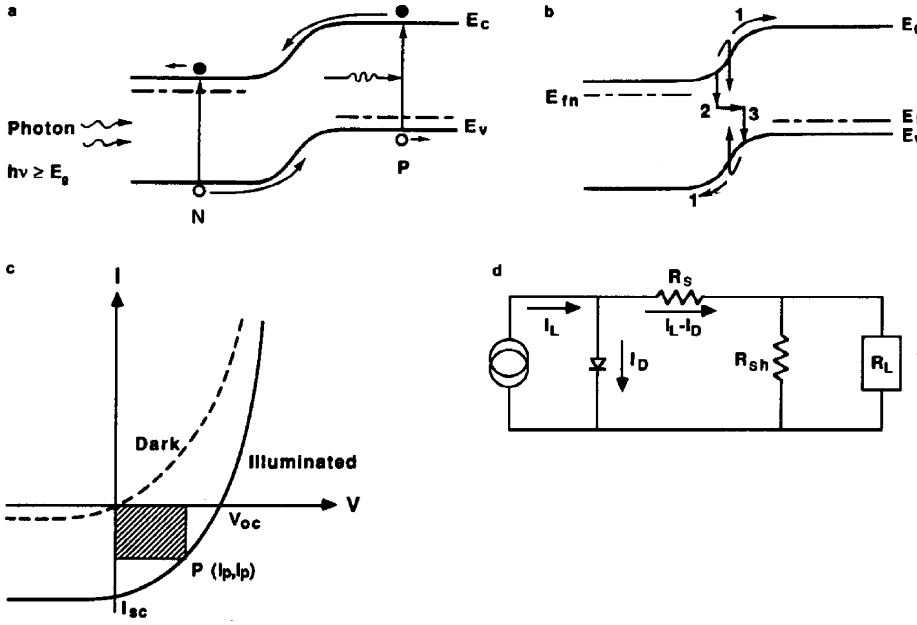


FIGURE 12.3. (a) The energy band diagram of an n^+-p junction solar cell under illumination; (b) the energy band diagram showing (1) the injection current, (2) the recombination current via a deep-level trap, and (3) the trap-assisted tunneling current in the junction space-charge region; (c) *I–V* characteristics under dark and illumination conditions; and (d) the equivalent circuit diagram.

such as amorphous and polycrystalline thin-film materials. The tuneling current component may become important in some solar cells such as Cu_2S/CdS heterojunction cells or MIS cells.

The injection current, which is due to the injection of holes from the p region into the n region and electrons from the n region into the p region of the junction, can be described by the ideal Shockley diode equation, which is given by

$$I_d = I_{01}[\exp(qV/k_BT) - 1], \tag{12.1}$$

where

$$I_{01} = qn_i^2 A_j \left(\frac{D_p}{L_p N_D} + \frac{D_n}{L_n N_A} \right) \tag{12.2}$$

is the reverse saturation current due to the injection of electrons and holes across the p-n junction; A_j is the junction area; n_i is the intrinsic carrier density; D_n and D_p denote the electron and hole diffusion coefficients; and L_n and L_p are the electron and hole diffusion lengths, respectively. Equation (12.1) is obtained by assuming

uniform doping in the n and p regions of the solar cell so that the quasineutral conditions prevail in both regions.

The recombination current in a p-n junction solar cell is due to the recombination of electrons and holes via deep-level defect centers inside the junction space-charge region. Based on the Shockley–Read–Hall (SRH) model, this recombination current component can be expressed by

$$I_{\mathrm{r}} = I_{02}[\exp(qV/mk_{\mathrm{B}}T) - 1], \tag{12.3}$$

where

$$I_{02} = \frac{qn_{\mathrm{i}}WA_{\mathrm{j}}}{2\sqrt{\tau_{\mathrm{n0}}\tau_{\mathrm{p0}}}}. \tag{12.4}$$

In (12.4), W is the depletion layer width; τ_{no} and τ_{po} are the minority electron and hole lifetimes in the p and n regions, respectively. It is noted that the diode ideality factor m in the exponent of (12.3) may vary between 1 and 2, depending on the location of the defect level in the forbidden gap. For example, m equals 2 if the recombination of electron–hole pairs is via a mid-gap recombination center (i.e., $E_{\mathrm{t}} = E_{\mathrm{i}}$), and is smaller than 2 if the recombination center is not located at the mid-gap or if the multilevel recombination centers exist in the junction space-charge region.

In general, the total dark current of a p-n junction solar cell can be represented by the sum of injection and recombination current components given by (12.1) and (12.3), namely,

$$I_{\mathrm{D}} = I_{\mathrm{d}} + I_{\mathrm{r}}. \tag{12.5}$$

The main difference between I_{d} and I_{r} given by (12.5) is that the injection current I_{d} varies with $n_{\mathrm{i}}^2\,\mathrm{e}^{qV/k_{\mathrm{B}}T}$, while the recombination current I_{r} varies with $n_{\mathrm{i}}\,\mathrm{e}^{qV/mk_{\mathrm{B}}T}$, which shows that the injection current depends more strongly on temperature than the recombination current. For a typical silicon p-n junction solar cell, values of the injection current density may vary from 10^{-8} to 10^{-12} A/cm^2, while values of the recombination current depend on the density of recombination centers in the junction space-charge region. In general, the recombination current is important only at low to moderate forward-bias regimes, and becomes less important at higher-bias regimes. For a high-quality p-n junction solar cell, the recombination current component can be neglected in (12.5). The tunneling current component, which may be important for an MIS solar cell or a CdS solar cell, can be neglected in (12.5).

(ii) Photo-I–V characteristics. The photocurrent generated in a p-n junction solar cell under 1-sun conditions is discussed next. When photons with energy $h\nu \geq E_{\mathrm{g}}$ impinge on a p-n junction solar cell, the rate of generation of electron–hole pairs as a function of distance x from the surface of the solar cell is given by

$$g_E(x) = \alpha\phi_0(1 - R)\,\mathrm{e}^{-\alpha x}, \tag{12.6}$$

where α is the optical absorption coefficient, ϕ_0 is the incident photon flux density (number of photons absorbed per unit area per second), and R is the reflection coefficient at the semiconductor surface. The photocurrent generated in a solar cell by the incident sunlight can be derived using the continuity equations for the excess carriers described in Chapter 6. For an n^+-p junction solar cell, the spatial distribution of the excess hole density, $\Delta p(x)$, generated in the n region of the solar cell can be obtained by solving the steady-state continuity equation given by

$$D_p \frac{d^2 \Delta p(x)}{dx^2} - \frac{\Delta p(x)}{\tau_p} = -\alpha \phi_0 (1 - R)\, e^{-\alpha x}. \tag{12.7}$$

Equation (12.7) has a general solution given by

$$\Delta p(x) = A \cosh\left(\frac{x}{L_p}\right) + B \sinh\left(\frac{x}{L_p}\right) + C e^{-\alpha x}, \tag{12.8}$$

where A and B are constants, which can be determined using the boundary conditions at the front surface of the cell where the recombination occurs (at $x = 0$) and at the depletion edge of the n quasineutral region (at $x = x_j$), and C can be determined using a particular solution for $\Delta p(x) = C e^{-\alpha x}$ in (12.7), which yields

$$C = \alpha \phi_0 (1 - R)\tau_p / (\alpha^2 L_p^2 - 1). \tag{12.9}$$

The first boundary condition is obtained from the fact that the diffusion current density is equal to the surface recombination current density at $x = 0$, which can be expressed as

$$D_p \frac{d\Delta p(0)}{dx} = s_p \Delta p(0) \quad \text{at } x = 0, \tag{12.10}$$

where $\Delta p(0) = p_n(0) - p_{no}$ is the excess hole density at $x = 0$, and s_p is the surface recombination velocity at $x = 0$. The second boundary condition is obtained at the edge of the space-charge region where the excess hole density is assumed equal to zero (i.e., holes are swept out by the high electric field in the depletion region). Thus, one can write

$$\Delta p(x_j) = 0 \quad \text{at } x = x_j, \tag{12.11}$$

where x_j is the junction depth. Constants A and B in (12.8) can be determined by solving the boundary conditions given by (12.10) and (12.11), and hence the photo-generated excess hole density in the n region is given by

$$\Delta p(x) = \frac{\alpha \phi_0 (1 - R)\tau_p}{(\alpha^2 L_p^2 - 1)} \times \left\{ -e^{-\alpha x} + \frac{(s_p + \alpha D_p) \sinh\left(\frac{X_j - x}{L_p}\right) + e^{-\alpha x_j}\left[s_p \sinh\left(\frac{x}{L_p}\right) + \left(\frac{D_p}{L_p}\right) \cosh\left(\frac{x}{L_p}\right)\right]}{s_p \sinh\left(\frac{x_j}{L_p}\right) + \left(\frac{D_p}{L_p}\right) \cosh\left(\frac{x_j}{L_p}\right)} \right\}. \tag{12.12}$$

Thus, the hole current density at $x = x_j$ generated by the absorbed photons of wavelength λ can be expressed by

$$\begin{aligned} J_p(\lambda) &= -qD_p \frac{d\Delta p}{dx}\Big|_{x=x_j} \\ &= \frac{q\phi_0(1-R)\alpha L_p}{(\alpha^2 L_p^2 - 1)} \\ &\times \left\{ -\alpha L_p e^{-\alpha x_j} + \frac{(s_p + \alpha D_p) - e^{-\alpha x_j}\left[s_p \cosh\left(\frac{x_j}{L_p}\right) + \left(\frac{D_p}{L_p}\right) \sinh\left(\frac{x_j}{L_p}\right)\right]}{s_p \sinh\left(\frac{x_j}{L_p}\right) + \left(\frac{D_p}{L_p}\right) \cosh\left(\frac{x_j}{L_p}\right)} \right\}, \end{aligned} \tag{12.13}$$

where $J_p(\lambda)$ is the photo-generated hole current density with wavelength λ in the n region of the n-p junction cell.

The photocurrent density due to electrons generated in the p-base region can be derived in a similar way to the hole current density generated in the n region derived above. The continuity equation for electrons in the p-base region is obtained by replacing $\Delta p(x)$ by $\Delta n(x)$, D_p by D_n, and τ_p by τ_n in (12.7). However, the boundary conditions for this case are given by

$$\begin{aligned} \Delta n(x) &= 0, \quad \text{at } x = x_j + W, \\ -D_n \frac{d\Delta n}{dx} &= s_n \Delta n, \quad \text{at } x = d, \end{aligned} \tag{12.14}$$

where W is the depletion layer width, d is the thickness of the solar cell, and $\Delta n(x) = n_p(x) - n_{p0}$ is the excess electron density. The photocurrent density per unit bandwidth due to electrons collected at the depletion edge of the p-base region is thus given by

$$\begin{aligned} J_n(\lambda) &= qD_n \frac{d\Delta n}{dx}\Big|_{x=x_j+W} \\ &= \frac{q\phi_0(1-R)\alpha L_n \exp[-\alpha(x_j + W)]}{(\alpha^3 L_n^2 - 1)} \\ &\times \left\{ \alpha L_n - \frac{s_n\left[\cosh\left(\frac{d}{L_n}\right) - e^{\alpha d}\right] + \left(\frac{D_n}{L_n}\right) \sinh\left(\frac{d}{L_n}\right) + \alpha D_n e^{-\alpha d}}{s_p \sinh\left(\frac{d}{L_n}\right) + \left(\frac{D_n}{L_n}\right) \cosh\left(\frac{d}{L_n}\right)} \right\}. \end{aligned} \tag{12.15}$$

In addition to the diffusion components of the photocurrent collected in the n and p quasineutral regions given by (12.13) and (12.15), the drift component of the photocurrent generated in the depletion region must also be considered. The electron–hole pairs generated in the depletion region are swept out by the built-in electric field in this region. The drift component of the photocurrent density

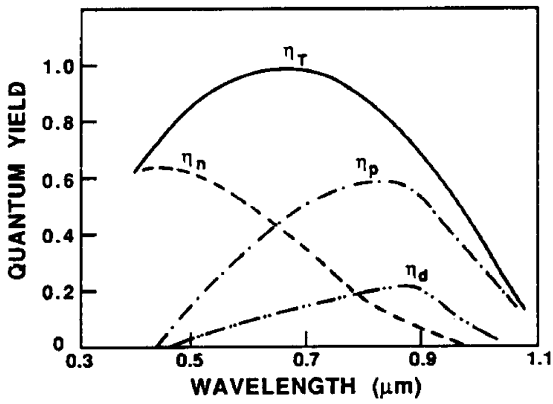


FIGURE 12.4. Normalized quantum yield versus wavelength in the n-emitter, p-base, and space-charge regions of a silicon p-n junction solar cell. The solid line denotes the total quantum yield, while the dashed and dotted lines are the quantum yields in different regions of the solar cell, assuming zero reflection loss at the top surface of the cell.

generated in the depletion region can be expressed by

$$J_{\mathrm{d}}(\lambda) = q \int_{x_{\mathrm{j}}}^{x_{\mathrm{j}}+w} g_E(\lambda)\,\mathrm{d}x = q\phi_0(1-R)\,\mathrm{e}^{-\alpha x_{\mathrm{j}}}(1-\mathrm{e}^{-\alpha W}). \tag{12.16}$$

Thus, the total photocurrent density generated in a p-n junction solar cell by the incident sunlight for a given wavelength λ is equal to the sum of (12.13), (12.15), and (12.16), which is

$$J_{\mathrm{L}}(\lambda) = J_{\mathrm{p}}(\lambda) + J_{\mathrm{n}}(\lambda) + J_{\mathrm{d}}(\lambda). \tag{12.17}$$

The quantum efficiency, which is defined as the number of electron–hole pairs generated per absorbed photon, for a p-n junction solar cell can be expressed by

$$\eta = \frac{J_{\mathrm{L}}(\lambda)}{q\phi_0(1-R)} \times 100\%, \tag{12.18}$$

where $J_{\mathrm{L}}(\lambda)$ is the photocurrent current density given by (12.17).

Figure 12.4 shows the quantum yield versus wavelength in the n^+-emitter (η_{n}), p-base (η_{p}), and the junction space-charge (η_{d}) regions as well as the total quantum yield (η_{T}) of a silicon n^+-p junction solar cell. In this plot the reflection loss (R) at the front surface of the solar cell is assumed equal to 0.

The total photocurrent density generated in a p-n junction solar cell under 1-sun conditions can be obtained by integrating (12.17) over the entire solar spectrum (under different AM conditions), which can be written as

$$J_{\mathrm{ph}} = \int_{\lambda_1}^{\lambda_2} J_{\mathrm{L}}(\lambda)\,\mathrm{d}\lambda, \tag{12.19}$$

where λ_1 and λ_2 denote the cutoff wavelengths at the short- and long-wavelength limits of the solar spectrum, respectively. For a typical p-n junction solar cell, λ_1 can be set at 0.3 μm, and λ_2 is determined by the cutoff wavelength or the energy band gap of the semiconductor (i.e., $\lambda_2 = \lambda_{\mathrm{c}} = 1.24/E_{\mathrm{g}}(\mu\mathrm{m})$). For a

silicon solar cell with $E_g = 1.12$ eV at 300 K, the cutoff wavelength is $\lambda_c = 1.1\ \mu m$.

(iii) Solar cell parameters. The equivalent circuit for a p-n junction solar cell is shown in Figure 12.3d, which is composed of the photocurrent component represented by a constant current source $I_{ph} = I_{sc}$, a dark current component I_D, a shunt resistance R_{sh}, and a series resistance R_s. If one neglects the effects of shunt resistance (assuming $R_{sh} \to \infty$), series resistance ($R_s \approx 0$), and the recombination current ($I_r = 0$) in the depletion region, then the photo-*I–V* characteristics of a p-n junction solar cell under illumination condition can be expressed by

$$I = -I_{ph} + I_{01}[\exp(qV/k_BT) - 1], \tag{12.20}$$

where I_{ph} is given by (12.19), and I_{01} is the injection current given by (12.2). The short-circuit current can be obtained by setting $V = 0$ in (12.20), which yields

$$I_{sc} = -I_{ph}, \tag{12.21}$$

which shows that the short-circuit current I_{sc} is equal to the photogenerated current $-I_{ph}$. The open-circuit voltage V_{oc} can be obtained by setting $I = 0$ in (12.20), and one obtains

$$V_{oc} = V_T \ln\left[\left(\frac{I_{sc}}{I_{01}}\right) + 1\right], \tag{12.22}$$

where $V_T = k_BT/q$ is the thermal voltage. It is seen from (12.22) that V_{oc} depends on the ratio of the short-circuit current and the dark current, and V_{oc} can be increased by keeping the ratio of I_{sc}/I_{01} as large as possible. This can be achieved by reducing the dark current, either by increasing the substrate doping density or by increasing the minority carrier lifetimes in the solar cell. Increasing the short-circuit current can also enhance V_{oc}, but it is not as drastic as reducing the dark current in the solar cell. In practice, the V_{oc} can be improved by incorporating a p-p$^+$ back surface field (BSF) structure in the n-p junction solar cell. The BSF structure not only can deflect the minority carriers back into the junction but can also reduce the back contact resistance of the cell. As a result, V_{oc}, J_{sc}, FF, and conversion efficiency can be improved with the BSF structure. Values of V_{oc} for a silicon p-n junction solar cell may vary between 0.5 and 0.7 V depending on the cell structure, doping densities, and other device parameters used in the cell's design and fabrication.

If one includes the series resistance R_s and neglects the shunt resistance effect (i.e., $R_{sh} \to \infty$) in the *I–V* equation, then the output current of the solar cell can be expressed as

$$I = I_D\{\exp[(V - IR_s)/V_T] - 1\} - I_{ph}, \tag{12.23}$$

and the output power is given by

$$P = |IV| = I\left[V_T \ln\left(\frac{I + I_{ph}}{I_D} + 1\right) + IR_s\right]. \tag{12.24}$$

The maximum output power can be calculated using the expression

$$P_{\mathrm{m}} = V_{\mathrm{m}} I_{\mathrm{m}}, \tag{12.25}$$

where

$$I_{\mathrm{m}} = (I_{\mathrm{sc}} + I_{01}) \left[\frac{(V_{\mathrm{m}}/V_{\mathrm{T}})}{(1 + V_{\mathrm{m}}/V_{\mathrm{T}})} \right]. \tag{12.26}$$

Here I_{m} is the current corresponding to the maximum power output, which is obtained by differentiating (12.24) with respect to current I and setting $\partial P/\partial I = 0$. It is noted that V_{m} is obtained by solving the equation given below using the iteration procedure:

$$\exp(V_{\mathrm{m}}/V_{\mathrm{T}})[1 + V_{\mathrm{m}}/V_{\mathrm{T}}] = \exp(V_{\mathrm{oc}}/V_{\mathrm{T}}). \tag{12.27}$$

Another important solar cell parameter known as the fill factor (FF), which measures the squareness of the photo-I–V curve shown in Figure 12.3c, is defined by

$$\mathrm{FF} = \frac{V_{\mathrm{m}} I_{\mathrm{m}}}{V_{\mathrm{oc}} I_{\mathrm{sc}}} = \left(\frac{V_{\mathrm{m}}}{V_{\mathrm{oc}}} \right) \left[1 - \frac{\mathrm{e}^{(V_{\mathrm{m}}/V_{\mathrm{T}})} - 1}{\mathrm{e}^{(V_{\mathrm{oc}}/V_{\mathrm{T}})} - 1} \right]. \tag{12.28}$$

Depending on the values of the diode ideality factor and the shunt and series resistances, the fill factor for a silicon p-n junction solar cell may vary between 0.75 and 0.85, while for a GaAs solar cell it may vary between 0.79 and 0.87.

Finally, the conversion efficiency of a p-n junction solar cell can be calculated by

$$\eta_{\mathrm{c}} = \left(\frac{P_{\mathrm{out}}}{P_{\mathrm{in}}} \right) \times 100\%, \tag{12.29}$$

where P_{in} is the input power from the sunlight, and P_{out} is the output power from the solar cell. The input power from the sunlight under 1-sun AM0, AM1, AM1.5G, and AM2 conditions are given by 135.3, 92.5, 100, and 69.1 mW/cm^2, respectively.

12.2.3. Design Considerations

It is clear from the above analysis that the performance of a solar cell is determined by both the device and physical parameters such as the minority carrier lifetimes and diffusion lengths, the doping densities, the series and shunt resistances, the AR coating, and the junction structures. Therefore, in order to obtain an optimal cell design, it is important to consider all the key device and material parameters that affect the conversion efficiency of a solar cell. These are discussed next.

(i) Spectral response. An important consideration in material selection for solar cell fabrication is to select a semiconductor with energy band gap that is matched with the peak irradiance power density of the solar irradiant spectrum. This will provide a maximum absorption of the incident sun power by the solar cell, and hence will enable the cell to produce an optimum spectral response. Figure 12.5 shows the maximum theoretical conversion efficiency of an ideal p-n junction solar

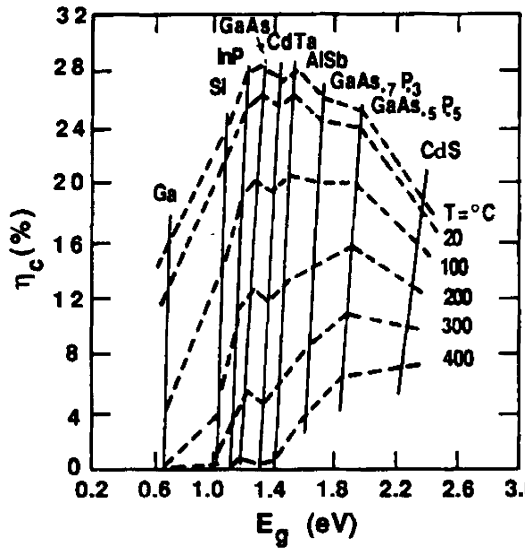


FIGURE 12.5. Maximum theoretical AM0 conversion efficiency versus energy band gap for different semiconductor materials. Dashed lines denote conversion efficiencies under different operating temperatures (0–400°C). After Wysocki and Rappaport,[2] by permission.

cell versus energy band gap for various semiconductor materials. It is noted that a single-junction GaAs solar cell has a maximum theoretical conversion efficiency of around 28% under AM0 conditions, while a silicon p-n junction solar cell has a maximum theoretical conversion efficiency of around 21% under AM0 conditions. The reasons a GaAs solar cell has a higher conversion efficiency than a silicon cell are that (a) the band gap energy for GaAs is better matched with the peak solar insolation spectrum than that of silicon, (b) GaAs has a larger band gap ($E_g = 1.43$ eV) and higher V_{oc} than silicon, and (c) GaAs is a direct band gap material, which has a larger absorption coefficient at peak solar irradiance, while silicon is an indirect band gap material.

(ii) Series resistance and contact grids. The series resistance, which is due to the contact and bulk resistances of the cell, can influence the shape of photo-*I–V* curve, fill factor, and conversion efficiency of a solar cell. For example, a large series resistance will increase the internal power dissipation and reduce the fill factor and output power of the solar cell. To reduce the effect of series resistance, both the contact and bulk resistances must be minimized. The contact resistance can be greatly reduced if optimal front contact grids are used. Unfortunately, reducing the sheet resistance of a solar cell is not an easy task. One way to reduce the sheet resistance is to employ a heavily doped surface layer and to design an optimized front contact grid in the cell. However, this will in turn reduce the minority carrier lifetime and diffusion length at the surface layer, and hence it could also decrease the short-circuit current. Therefore, a compromise between the doping density and the junction depth is necessary in order to achieve optimal design. There are a number of ways of making the front contact grids. The most common contact grids are formed with the rectangular metal grids (fingers); each grid line is equally spaced and set on top of the front surface of the solar cell. This is shown in

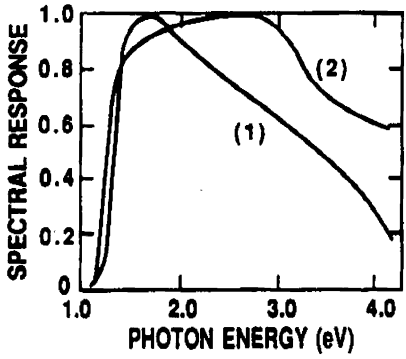


FIGURE 12.6. Relative spectral responses of (1) a normal silicon p-n junction solar cell, and (2) a violet (i.e., nonreflecting) silicon p-n cell that utilizes a texturized front surface to reduce reflection loss to near zero. After Lindmayer and Allison,[3] by permission.

Figure 12.2a for a silicon n^+-p junction solar cell. It is noted that using the contact grid structure allows the exposure of a major portion of the solar cell surface to sunlight and at the same time keeps the series resistance to a minimum value. It should be noted that the area covered by the front metal contact grids is usually less than 10% of the total solar cell area.

(iii) Antireflection coatings. Another important factor that must be considered in the solar cell design is the reflection loss at the front surface of the solar cell. For example, as much as 30–35% of the incident sunlight in the visible spectral range is reflected back to the air from the bare surface of a semiconductor without AR coatings. Therefore, it is important to reduce the surface reflection loss by using proper AR coatings on the solar cell. To illustrate the effect of reflection loss on the quantum efficiency of a solar cell, Figure 12.6 shows the spectral response of a regular silicon p-n junction solar cell (curve 1) and a violet silicon solar cell (curve 2). The reflection loss in the violet cell is reduced to near zero from UV to the visible wavelength regime when a texturized front surface is used. The texturized grooves on the front surface of a silicon solar cell can be formed using a preferential etching technique. As shown in Figure 12.6, the short-wavelength response of a violet cell is greatly improved over a regular cell as a result of using the texturized surface.

The most widely used technique for achieving near-zero reflection loss is by applying the AR coatings on the front surface of a solar cell. This is usually achieved by depositing a thin dielectric film of Ta_2O_5, SiO_2, or Si_3N_4 on the front side of a silicon solar cell with thickness equal to a quarter-wavelength of a selected incident monochromatic light corresponding to the peak response wavelength of the solar cell. The thickness of the dielectric film for a single AR coating can be calculated using the expression

$$d = \frac{\lambda_0}{4n_1}, \tag{12.30}$$

where λ_0 is the wavelength of incident sunlight at a selective wavelength, and n_1 is the refractive index of the dielectric film used for AR coatings. For example, using SiO_2 film ($n_1 = 1.5$) for AR coatings on a silicon solar cell, the film thickness

calculated using (12.30) is found to be $d = 80$ nm at $\lambda_0 = 0.48\,\mu$m, and $d = 100$ nm at $\lambda_0 = 0.60\,\mu$m. The minimum reflection loss for a quarter-wavelength AR coating may be calculated using the expression

$$R_{\min} = \left(\frac{n_1^2 - n_0 n_2}{n_1^2 + n_0 n_2}\right)^2, \tag{12.31}$$

where n_0, n_1, and n_2 are the refractive indices of air, AR coating film, and the solar cell material, respectively. From (12.31), it is found that a silicon solar cell coated with a 110-nm thick SiO_2 film has a reflection loss of only 7%, which is a drastic improvement over that of a silicon solar cell without AR coatings ($R = 35\%$). Among the various AR coating materials used today, Ta_2O_5 (with $n_1 = 2.25$) is probably the most widely used dielectric film used for AR coatings on silicon solar cells, which can be easily done using the sputtering technique. For example, by applying a 70-nm-thick Ta_2O_5 AR coating film on a silicon solar cell, the reflection loss can be reduced to about 5%. Therefore, it is clear that by carefully selecting a suitable AR coating film, it is possible to reduce the reflection loss of a solar cell to almost 0. It is noted that aside from the methods cited above for improving solar cell performance, there are other means that could also be employed to further improve the conversion efficiency of a solar cell. For example, the short-wavelength spectral response may be improved by using a shallow-junction structure with a thin p-emitter in a p-n junction solar cell. The V_{oc} of a p-n junction cell could be increased using a BSF structure (i.e., n/n^+ or p/p^+) and by increasing the doping density (to reduce the dark current) in the base region of the cell. Theoretical calculations reveal that $V_{oc} = 0.7$ V for a silicon n^+-p junction solar cell can be obtained using a 0.1 $\Omega \cdot$ cm ($N_A \approx 2 \times 10^{17}$ cm^{-3}) silicon material for the p-base layer. Recently, Martin Green's group at the University of New South Wales, in Sydney, Australia, has reported a record 24.7% AM1.5G efficiency for a PERL (passivated emitter, rear locally diffused) silicon solar cell. This is the highest ever reported conversion efficiency for a silicon solar cell. The advanced surface passivation method and an improved cell pattern design have contributed to the improvement in the cell efficiency. The PERL cell has achieved $V_{oc} = 706$ mV, $J_{sc} = 42.2$ mA/cm^2, and F.F. = 82.8%. Using their PERL cells, a silicon PV module has demonstrated 22.7% AMI.5G efficiency, which is the highest reported efficiency for a PV module made on any material.

For a GaAs p-n junction solar cell, the short-wavelength spectral response and contact resistance can be greatly improved by growing a wide-band-gap p^+-$Al_{0.9}Ga_{0.1}$ As window layer 0.3–0.5 μm thick on top of the GaAs p-n junction cell structure. Figure 12.7 shows the cross-sectional view of a high-efficiency AlGaAs/GaAs p-n junction solar cell. The reason for using a thin highly doped p^+-AlGaAs wide-band-gap window layer on top of the p^+-GaAs emitter layer is to reduce the surface recombination velocity and series resistance of the GaAs solar cell. Since the $Al_{0.9}Ga_{0.1}As$ window layer is a wide-band-gap ($E_g = 2.1$ eV) material, it is transparent to most of the visible sunlight. The AlGaAs window layer can reduce the surface recombination velocity of a GaAs solar cell to less

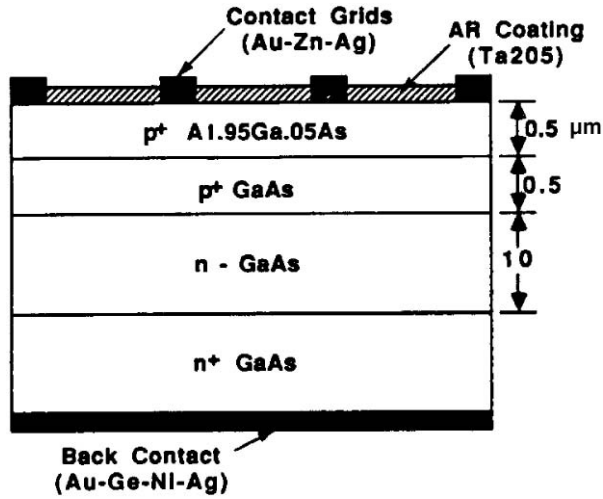


FIGURE 12.7. The cross-sectional view of a GaAs p-n junction solar cell with a p^+-AlGaAs window layer and a Ta_2O_5 layer for antireflection coating.

than 10^3 cm/s from 10^6 cm/s. Figure 12.8 shows the spectral response curves for a GaAs p-n junction solar cell with (curve 2) and without (curve 1) an AlGaAs window layer. It is clearly shown that adding an AlGaAs window layer to the GaAs p-n junction solar cell can indeed produce a significant improvement in the short-wavelength response. Maximum conversion efficiency of 24.2% AM1.5G for a GaAs single-junction solar cell has been achieved recently. The solar cell structure consists of an active GaAs single-junction solar cell grown on an inactive back-side AlGaAs/GaAs distributed Bragg reflector (DBR) grown on top of the GaAs substrate. This GaAs cell is capped with an AlGaAs window layer. The reflectivity of the backside DBR is approximately 70% for incident light with energy close to the band edge of GaAs. The cell is coated with a TiO_2/MgF_2 AR coating. The GaAs-based solar cells are preferred for powering of space satellites. The use of

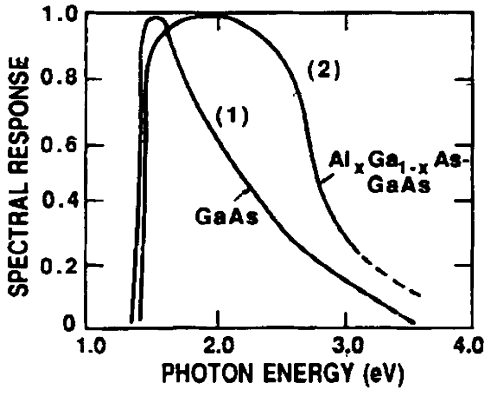


FIGURE 12.8. Spectral response curves (i.e., normalized quantum yield) for a GaAs p-n junction solar cell: (1) without window layer and (2) with an AlGaAs window layer.

Ge substrates could reduce weight significantly, and the conversion efficiency can be further increased using a GaInP/GaAs/Ge tandem cell structure. These types of solar cells are superior to silicon cells because of their lighter weight and higher resistance to the cosmic radiation in space.

12.2.4. Schottky Barrier and MIS Solar Cells

Although most commercial solar cells use a p-n junction structure, other structures such as Schottky barrier, MIS, heterojunction, and multijunction tandem cell structures have also been employed for cell fabrication. The Schottky barrier solar cell is easy to fabricate and has the simplest structure among all the solar cells. It has a better spectral response in the shorter-wavelength regime, and hence can produce higher short-circuit current. However, the conversion efficiency of a Schottky barrier solar cell is usually lower than a p-n junction solar cell due to its lower open-circuit voltage, which is a direct result of higher dark current due to the inherent barrier height limitation.

A Schottky barrier solar cell can be fabricated using either a thin semitransparent metal film or a grating-type structure deposited on the semiconductor substrate to form Schottky contacts. Figure 12.9a shows the cross-sectional view of a Schottky barrier solar cell with a 10-nm semitransparent metal film for Schottky contact, and Figure 12.9b shows the energy band diagram under illumination conditions.

In a Schottky barrier solar cell, the photocurrents are generated in the depletion and base regions of the cell. The collection of electron–hole pairs in the depletion region is similar to that of a p-n junction cell discussed in the previous section. The excess carriers generated in the depletion region are swept out by the built-in electric field in this region, leading to a photocurrent density per unit bandwidth given by

$$J_{\mathrm{d}}(\lambda) = q \int_0^W T(\lambda)\alpha\phi_0(\lambda)\mathrm{e}^{-\alpha x}\mathrm{d}x = qT(\lambda)\phi_0(\lambda)(1 - \mathrm{e}^{-\alpha W}), \tag{12.32}$$

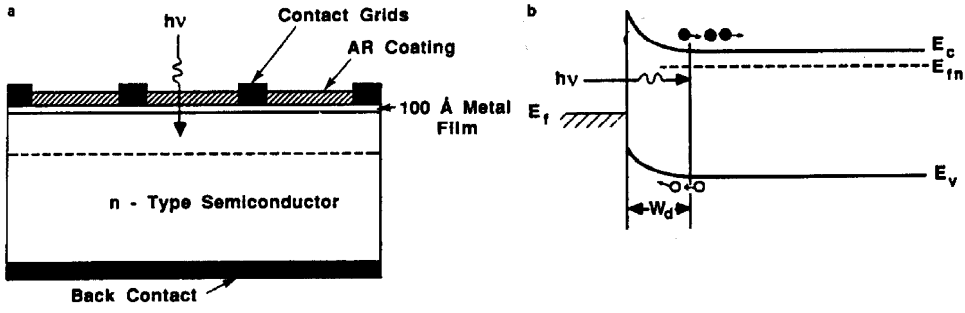


FIGURE 12.9. (a) Cross-sectional view and (b) energy band diagram of a metal–n-type semiconductor Schottky barrier cell under illumination conditions.

where $\phi_0(\lambda)$ is the incident photon flux density at wavelength λ, $T(\lambda)$ is the transmission coefficient of the metal film, and W is the depletion layer width given by

$$W = \sqrt{\frac{2\varepsilon_0\varepsilon_s(V_d - V)}{qN_D}}. \tag{12.33}$$

The photocurrent given by (12.32) is similar to that given by (12.16) for a p-n junction cell, except that in the latter case the transmission coefficient of light through the metal film (i.e., $T(\lambda)$) for the Schottky contact is replaced by the transmission coefficient $(1 - R)$ of light through a p-n junction solar cell.

The collection of photocurrent in the quasineutral base region of a Schottky barrier solar cell is similar to that in the base region of a p-n junction cell. Thus, the photocurrent density due to holes collected in the n-base region can be expressed by

$$J_p(\lambda) = \frac{q\phi_0\alpha L_p}{(1 + \alpha L_p)} T(\lambda)\, e^{-\alpha w}. \tag{12.34}$$

Equation (12.34) is obtained by assuming that the cell thickness is much larger than the hole diffusion length in the n-base region. The photocurrent generated in a Schottky barrier solar cell due to the incident monochromatic light of wavelength λ is equal to the sum of (12.32) and (12.34). Thus, the total photocurrent generated by the sunlight can be obtained by integrating the single-wavelength photocurrent from the UV (λ_1) to the cutoff wavelength (λ_2) of the semiconductor material, namely,

$$J_{ph} = \int_{\lambda_1}^{\lambda_2} [J_d(\lambda) + J_p(\lambda)]\, d\lambda, \tag{12.35}$$

where $\lambda_1 = 0.3\ \mu m$ and $\lambda_2 = \lambda_g$, the cutoff wavelength of the semiconductor.

Under forward-bias conditions, the dark current in a Schottky barrier solar cell is due primarily to the thermionic emission of majority carriers in the bulk semiconductor, which is given by

$$J_D = J_s[\exp(V/nV_T) - 1], \tag{12.36}$$

where $J_s = A^*T^2 \exp(-\phi_{Bn}/V_T)$ is the saturation current density, and A^* is the effective Richardson constant, which is equal to 110 A/(cm^2·K^2) for n-type silicon and 8.16 A/(cm^2·K^2) for n-type GaAs. Here A^* is equal to 79.2 A/(cm^2·K^2) for p-Si, and 74.4 A/cm^2-K^2 for p-GaAs. Since the barrier height is generally lower than the band gap energy of the semiconductor, one expects that the saturation current for a Schottky barrier solar cell will be much higher than that of a p-n junction solar cell. As a result, V_{oc} for a Schottky barrier solar cell is expected to be lower than that of a p-n junction solar cell. To overcome this problem, barrier height enhancement techniques described in Section 10.8 may be applied to the Schottky barrier solar cell in order to obtain a higher V_{oc} and conversion efficiency.

The open-circuit voltage, fill factor, maximum power output, and conversion efficiency for a Schottky barrier solar cell can be calculated in a similar way to

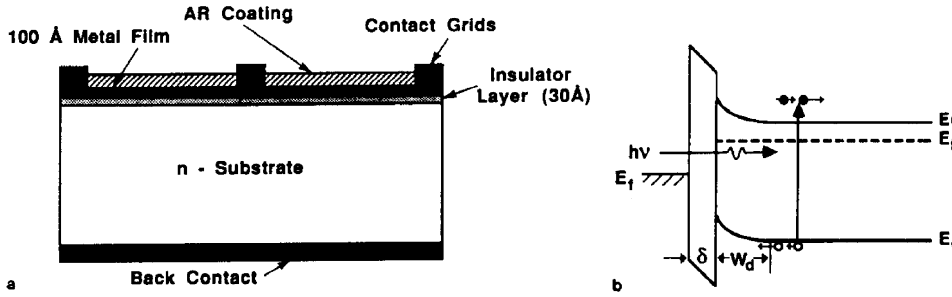


FIGURE 12.10. (a) Cross-sectional view and (b) energy band diagram of an MIS solar cell under illumination conditions.

those of the p-n junction solar cell discussed earlier. The short-circuit current and dark current can be calculated using (12.32)–(12.36). Typical values of V_{oc} from 0.4 to 0.55 V, fill factor (FF) of 0.6–0.76, and conversion efficiency η_c of 12–15% AM1.5G are achievable for a silicon Schottky barrier solar cell.

Another method to improve the value of V_{oc} in a Schottky barrier solar cell is to use an MIS structure. In this structure, a thin insulating layer with thickness of 1–3 nm is formed between the metal Schottky contact and the semiconductor, which results in an MIS solar cell structure. The MIS structure can increase the effective barrier height ($\Delta\phi_B = \delta\chi^{1/2}$), and hence can reduce the dark current of the MIS cell. As a result, the V_{oc} of an MIS solar cell is usually higher than that of a conventional Schottky barrier cell. Figure 12.10a shows the cross-sectional view of an MIS Schottky barrier solar cell and Figure 12.10b displays the energy band diagram under illumination conditions. In an MIS solar cell, current conduction under dark conditions is due to majority carriers tunneling through the thin insulating layer. This tunneling current can be described by

$$J_t = A^*T^2 \exp(-\phi_{Bn}/V_T) \exp(-\delta\chi^{1/2}) \exp(V/nV_T), \tag{12.37}$$

where δ is the thin insulating layer thickness in Å, χ is the mean incremental barrier height, and ϕ_{Bn} is the barrier height without the thin insulating layer (i.e., $\delta = 0$). Equation (12.37) reduces to (12.36) if $\delta = 0$. It is seen from (12.37) that the thin insulating layer in an MIS structure will limit only the majority carrier flow and not the minority carrier flow (or the photocurrent) as long as the thickness of the insulating layer remains very thin (e.g., $\delta \leq 30$ Å). Thus, the V_{oc} of an MIS solar cell will be higher than that of a conventional Schottky barrier solar cell. The V_{oc} of an MIS solar cell can be derived from (12.22) and (12.37), which yields

$$V_{oc} = nV_T \left[\ln\left(\frac{J_{sc}}{A^*T^2}\right) + \frac{\phi_{Bn}}{V_T} + \delta\chi^{1/2} \right]. \tag{12.38}$$

Conversion efficiencies of 15% for an Au–Si MIS cell and 17% for an Au–GaAs MIS cell under AM1 conditions have been reported. The main drawback to the MIS solar cells is the difficulty of controlling the thin insulating layer thickness in the

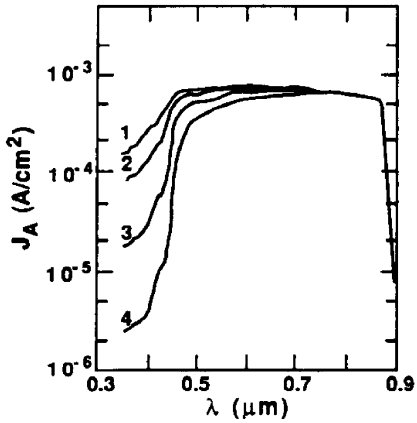


FIGURE 12.11. Calculated photocurrent density versus wavelength for a Au–p^+-n GaAs Schottky barrier solar cell under AM0 conditions. The dopant density of the n-GaAs substrate is $N_d = 10^{16}$ cm^{-3}, and the thicknesses and dopant densities for curves 1 through 4 are given by: curve 1, $N_a = 8.2 \times 10^{16}$ cm^{-3} and $W_p = 100$ Å; curve 2, $N_a = 2.2 \times 10^{16}$ cm^{-3} and $W_p = 200$ Å curve 3, $N_a = 4.4 \times 10^{16}$ cm^{-3} and $W_p = 500$ Å; curve 4, $N_a = 8.2 \times 10^{16}$ cm^{-3} and $W_p = 1000$Å. After Li,[4] by permission.

cell. When the insulating film thickness exceeds 5 nm, photocurrent suppression results and the conversion efficiency drops. This, in turn, will lower the value of V_{oc} and the conversion efficiency of the MIS solar cell. An alternative approach for solving the problems associated with low barrier height and high dark current in a Schottky barrier solar cell is to introduce a thin semiconductor layer of opposite doping type to the substrate to form a metal–p^+-n or metal–n^+-p Schottky barrier structure, as described in Section 10.8. Using this approach, enhancement of the effective barrier height for Au–p^+-n and Au–n^+-p GaAs Schottky diodes can be readily achieved, as shown earlier in Figures 10.21 and 10.22, respectively. Figure 12.11 shows the calculated photocurrent density versus wavelength for an ideal Au–p^+-n GaAs Schottky barrier solar cell under AM0 conditions for four different p-layer doping densities and thicknesses. Theoretical conversion efficiency as high as 21% for a metal–n^+-p GaAs Schottky barrier solar cell structure is predicted under AM0 conditions.

Finally, since a Schottky barrier solar cell offers several advantages such as low cost, simple structure, ease of fabrication, and low-temperature processing, it is clear that using a Schottky barrier structure could be an attractive and viable approach for fabricating low-cost photovoltaic modules for terrestrial power generation.

12.2.5. Heterojunction Solar Cells

A p-n heterojunction solar cell is formed using two semiconductor materials of different band gap energies and opposite dopant impurities. For example, a p-n heterojunction solar cell can be fabricated using a p^+-InGaP/n-GaAs, p^+-AlGaAs/n-GaAs, p^+-GaAs/n-Ge, n-CdS/p-CdTe, or n-CdS/p-CIGS material system. Figure 12.12a shows the energy band diagram for an n^+-GaAs/p-Ge heterojunction solar cell in equilibrium. In this heterojunction structure, a wide band gap ($E_{g1} = 1.43$ eV) n-GaAs is used as the emitter layer, while a small band gap ($E_{g2} = 0.67$ eV)

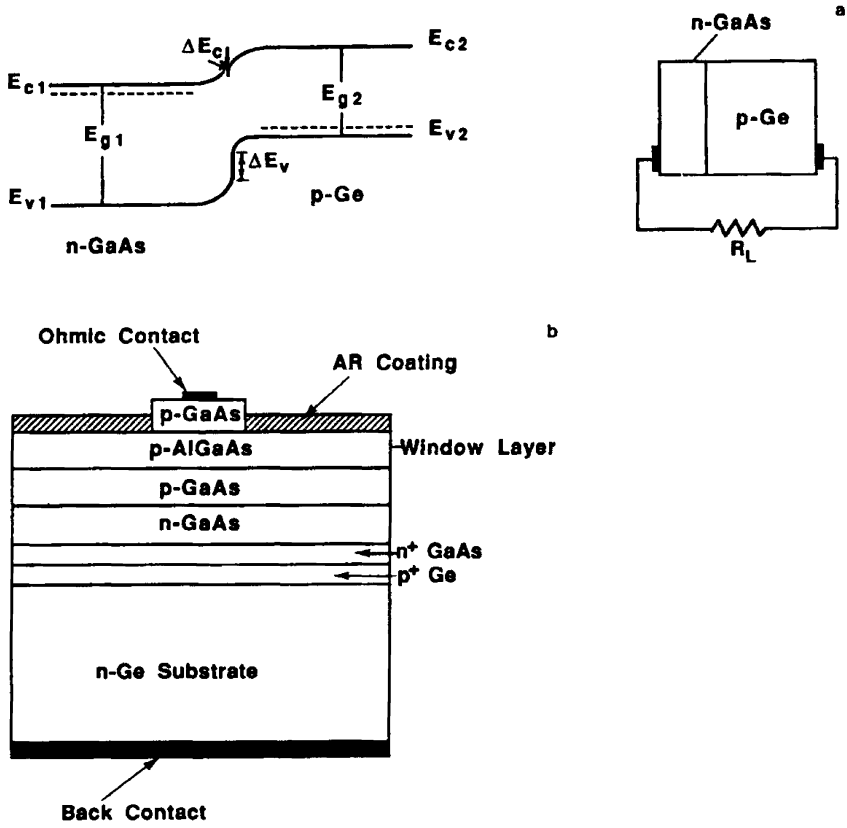


FIGURE 12.12. (a) Energy band diagram of an n-GaAs/p-Ge heterojunction solar cell in equilibrium. (b) Cross-sectional view of a high-efficiency MOCVD-grown GaAs/Ge tandem solar cell. After Tobin et al.,[5] by permission, © IEEE-1988.

p-Ge is used as the base layer. The distinct feature of a p-n heterojunction solar cell lies in its window effect in which photons with energies between E_{g1} and E_{g2} can pass through the wide-band-gap window layer, and are absorbed in the smaller-band-gap base layer. The window layer is usually heavily doped, and has a thickness of a few tenths of a micrometer. Thus, with the addition of a window layer, the sheet resistance of the heterojunction solar cell can be reduced, which is important for reducing the internal power loss of the solar cell. In general, the output power and conversion efficiency of a heterojunction solar cell are determined mainly by the photocurrent produced in the smaller-band-gap base layer. In order to fully utilize the solar spectrum and to increase the conversion efficiency of a solar cell, multijunction solar cell structures have been widely investigated in recent years. For example, a high-efficiency GaAs/Ge heterojunction tandem solar cell grown using the metal-organic chemical vapor deposition (MOCVD) technique has been

reported. This tandem cell structure is shown in Figure 12.12b, which consists of a 4-μm front AR coating, a 0.46-μm p^+-GaAs front contact layer, a 0.03-μm p^+-AlGaAs window layer, a 0.5-μm p-GaAs emitter and a 2.6-μm n-GaAs base layer for the top cell, a 1.7-μm thick n^+-GaAs buffer layer, and a p^+ (1 μm)-n (200 μm) Ge bottom cell. AM0 conversion efficiency of 21.7% and AM1.5G conversion efficiency of 24.3% have been achieved for this tandem solar cell.

The current collection mechanism for a p-n heterojunction solar cell is similar to that of a p-n homojunction solar cell. The main contribution to the photocurrent comes from the base region, with smaller contribution coming from the top emitter layer and the depletion region. The photocurrent for a p-n heterojunction solar cell can be derived in a similar way to that of the p-n homojunction solar cell discussed earlier.

The p-n heterojunction solar cell usually has a better short-wavelength response, lower series resistance, and better radiation tolerance than a conventional p-n homojunction solar cell. In order to obtain maximum short-circuit current, open-circuit voltage, and conversion efficiency, it is essential that materials selected for fabricating the heterojunction cells have good lattice match and compatible thermal expansion coefficients. Energy band discontinuities at the heterointerface of a heterojunction cell must be minimized to avoid barrier formation (spike) at the heterointerface, where photocurrent collection can be severely degraded. Several heterojunction pairs with good lattice matches have been reported in the literature. These include the InGaP/GaAs, GaAs/Ge, AlGaAs/GaAs, GaP/Si, CdTe/CdS, and $CuInSe_2$/CdS material systems.

It is interesting to compare the characteristics of a heterojunction solar cell to a Schottky barrier solar cell. The most striking similarity is that short-wavelength photons can be absorbed within or very near the surface region of both cells, leading to an excellent short-wavelength response. However, the open-circuit voltage of a heterojunction solar cell can be much higher than that of a Schottky barrier solar cell as a result of the use of a larger-band-gap material for the top layer of the solar cell. As a result, higher conversion efficiency can be expected in a heterojunction solar cell. In fact, an $Al_xGa_{1-x}As$/GaAs p-n heterojunction solar cell with AM1 conversion efficiency as high as 21.5% has been reported. It has been shown that a heterojunction solar cell is more radiation-tolerant to low-energy protons and 1-MeV electron irradiation than that of a conventional p-n junction solar cell, because a thicker wide-band-gap window layer is used in a heterojunction solar cell to cut down radiation damage on the cell without losing the short-circuit current and conversion efficiency. Another type of heterojunction solar cell using a wide-band-gap conducting glass such as indium oxide (In_2O_3), tin oxide (SnO_2), or indium tin oxide (ITO) has also been reported in the literature. These highly conducting glasses with band gap energies varying from 3.5 to 3.7 eV are n-type semiconductors, and can be deposited on top of a p-type silicon substrate to form an n^+-ITO/p-Si heterojunction solar cell. The ITO film has a typical thickness of around 400 nm and resistivity of $5 \times 10^{-4}\Omega\cdot$cm. Conversion efficiencies of 12–15% have been reported for the ITO/Si n-p junction solar cells. Heterojunction

structures are used extensively in the fabrication of high-efficiency multijunction tandem solar cells and concentrator solar cells, as will be described later.

12.2.6. Thin-Film Solar Cells

Thin-film solar cells are promising candidates for low-cost, large-scale terrestrial PV power-generation applications. A thin-film solar cell (film thickness $\leq 10\,\mu m$) uses thin absorber layers to form a p-n junction or a Schottky barrier structure on foreign substrates. The absorber layers may be polycrystalline or amorphous films. Recently, a number of semiconductor materials including CdTe, Cu(In,Ga)Se_2(CIGS), and polycrystalline and amorphous silicon (a-Si) materials have been developed for thin-film PV device applications. A wide variety of low-cost substrate materials such as ceramic, soda-lime glass, graphite, aluminum, polymer, stainless steel, and metallurgical grade silicon have been used as substrate materials for the fabrication of thin-film solar cells for space and terrestrial power generation applications. Currently, there are three main competing thin-film PV technologies based on a-Si, CdTe, and CIGS absorbers available for large-scale terrestrial power generation and other consumer electronic applications. These are discussed next.

(i) a-Si thin-film solar cells. Among the various thin-film solar cells, the hydrogenated amorphous silicon (a-Si:H) thin-film solar cells have been developed for a wide variety of consumer electronic uses and for low-cost, large-scale PV power generation. Conversion efficiency around 10–11% AM1.5G for a single-junction a-Si:H solar cell and over 14% AM1.5G for an a-Si/a-SiGe/a-SiGe triple-junction solar cell have been developed for commercial applications. However, problems associated with long-term stability and degradation in a-Si solar cells have yet to be solved before large-scale production of a-Si thin-film solar cell modules can be implemented for terrestrial power generation use.

Most of the a-Si solar cells made today are being used in powering calculators, watches, toys, and cameras. The a-Si solar cells are formed by depositing a 1- to 3-μm thick a-Si thin film by RF glow-discharge decomposition of silane (which produces 10% hydrogenated a-Si) onto metal or ITO (indium–tin oxide) coated glass substrates. The a-Si:H is distinguished from crystalline silicon by the lack of long-range order and the high content of bonded hydrogen (typically around 10%) in device-quality a-Si:H. Hydrogen atoms passivate most of the unsaturated Si dangling bonds and make the a-Si film useful for photovoltaic device applications. Despite the inherent limitations (disorder causes high density of localized defects and the band-tailing effect lowers the value of V_{oc}), the a-Si:H material is a promising candidate for low-cost, large-scale photovoltaic applications. The commercial deposition process of a-Si:H using plasma-enhanced chemical vapor deposition (PECVD) is compatible with large-area deposition and low-temperature processing. This allows the use of a large variety of inexpensive substrate materials. The a-Si:H films can be easily doped by adding phosphorus-or boron-containing gases

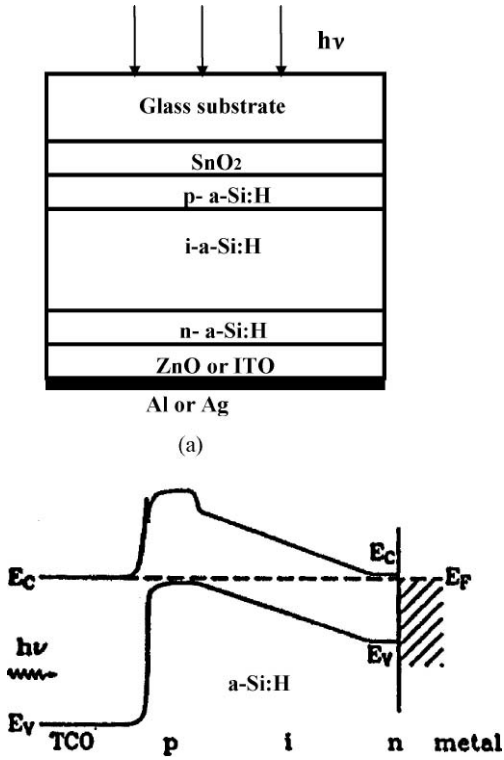


FIGURE 12.13. (a) Layer structure and (b) schematic energy band diagram for an a-Si:H p-i-n solar cell grown on ITO (indium–tin-oxide conducting film) coated glass, and (b) energy band diagram under illumination conditions.

during the deposition process for n- and p-type doping. The optical band gap of a-Si is typically around $E_g \approx 1.7$ eV and can be tuned. For example, the band gap energy of a-Si:H can be increased by alloying with carbon or oxygen, and decreased with incorporation of germanium to form a-Si_xGe_{1-x} ($0 \leq x \leq 1$) films. The energy band gap can also be fine-tuned by changing the hydrogen content using different deposition parameters and methods.

Figures 12.13a and b show the layer structure and the schematic energy band diagram of an a-Si:H p-i-n thin-film solar cell formed on an ITO-coated glass substrate. It is noted that the optical properties for a hydrogenated a-Si (a-Si:H) film resembles a direct band gap material with an energy band gap of $E_g = 1.7$ eV. As a result, a-Si:H film has a much higher optical absorption coefficient at a photon energy of $hv = 1.7$ eV than that of single-crystal silicon. Thus, the short-wavelength response (i.e., the blue-green region) for an a-Si:H solar cell is much better than that of single-crystal silicon solar cell. The most widely used structure for an a-Si:H solar cell is a p-i-n structure deposited on the conducting ITO-coated glass substrate, as shown in Figure 12.13b. In this structure, the typical thickness of p^+ and n^+ layers is around 10–30 nm, and the intrinsic (i) layer thickness may vary between 200 nm and 500 nm. Conversion efficiency over

10% AM1 has been reported for a single-junction a-Si:H solar cell. To achieve higher conversion, a stack of three cells with different band gap energies using an a-Si/a-SiGe/ a-SiGe triple-junction structure has been reported. In the triple stack, the top cell, which captures the blue photons, utilizes a-Si:H with an optical band gap of $E_g = 1.8$ eV for the i-layer. The i-layer for the middle cell is an a-SiGe alloy with 10–15% Ge and a band gap of $E_g = 1.6$ eV, which is ideally suited for absorbing the photons in the green spectral range. The i-layer for the bottom cell uses an a-SiGe with 40–50% of Ge, which has an optical band gap of $E_g = 1.4$ eV, suitable for absorbing the red and infrared photons. Light that is not absorbed in the cells is reflected from the Ag/ZnO back reflector. These three subcells are interconnected through the heavily doped layers that form the tunnel junctions between the adjacent cells. Conversion efficiency over 14% AM1.5G has been achieved in such an a-Si/a-SiGe/a-SiGe triple junction solar cell. The a-Si:H solar cells have been widely used in consumer electronics for powering digital watches, calculators, and other electronic gadgets. In recent years, a-Si:H thin-film solar cell modules have been developed for large-scale low-cost terrestrial power generation, although issues related to long-term stability and degradation problems remain the main concern for such applications. Recently, it has been shown that a two-terminal a-Si:H solar cell stacked with a poly-Si cell structure has achieved a conversion efficiency of 15.04% AM1.5G with $V_{oc} = 1.478$ V, $J_{sc} = 16.17$ mA/cm^2, and FF = 63%. Finally, a prototype four-terminal a-Si:H p-i-n top cell stacked with a poly-Si bottom cell has achieved a total conversion efficiency of 21% AM1.5G with top cell efficiency of 7.25% and bottom cell efficiency of 13.75%.

(ii) Cu(In,Ga)Se$_2$thin-film solar cells. Cu(In,Ga)Se$_2$ (CIGS) is an excellent material for high-efficiency thin-film solar cells because it is a direct band gap semiconductor with suitable energy band gap and high optical absorption coefficient in the visible spectrum of incident sunlight. The absorption coefficient of CIGS films in the visible spectrum is 100 times larger than silicon material. Therefore, a 2-μm thick CIGS absorber film is sufficient to absorb more than 90% of useful sunlight for the conversion of solar energy into electricity. Moreover, the CIGS films can be deposited by physical evaporation deposition (PVD) or spray-paint technique on various inexpensive substrates such as soda-lime glass, stainless steel, and plastic substrates for the production of low-cost solar cell modules. Recently, high-efficiency CIGS solar cells grown on flexible stainless steel substrates have been reported. The device structure consists of MgF$_2$/ITO/ZnO/CdS/CIGS /Mo/stainless steel substrates. A compound layer that contains Na was deposited prior to the formation of the CIGS absorber layer. The maximum efficiency for the best CIGS cell grown on stainless steel substrates is 17% AM1.5G ($V_{oc} = 0.628$ V, $J_{sc} = 37.2$ mA/cm^2, FF = 0.723) with an active area of 0.96 cm^2. This performance is comparable to the CIGS cells fabricated on soda-lime glass substrates, which has a world record efficiency of 19.8% AM1.5G, as reported recently by the NREL (National Renewable Energy Laboratory) research team.

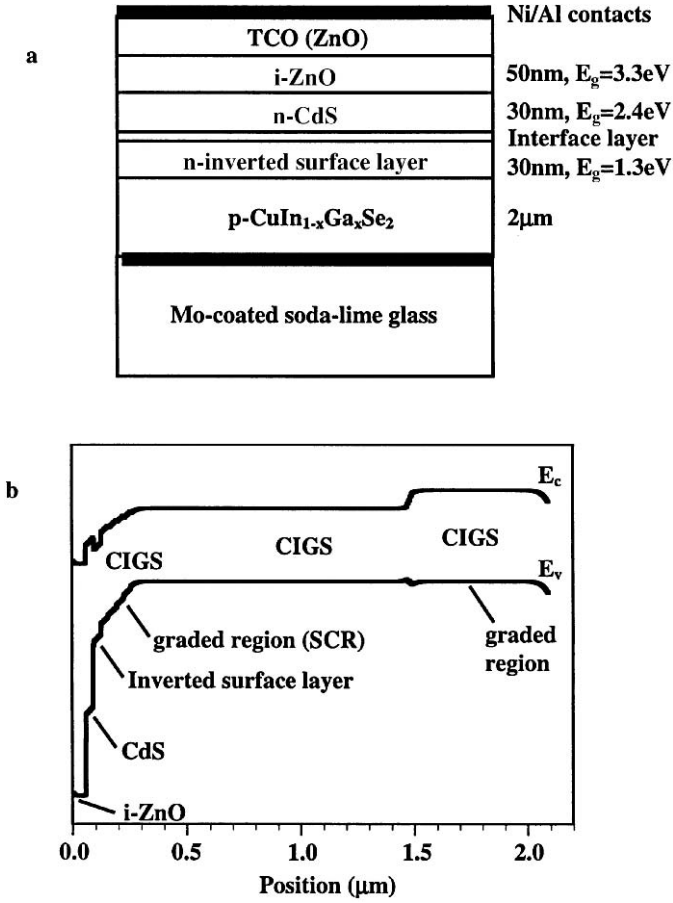


FIGURE 12.14. (a) Layer structure and (b) energy band diagram of a Cds/CIGS thin-film solar cell grown on Mo-coated soda-lime glass substrate.

Figure 12.14 shows (a) the cell structure and (b) the energy band diagram of a CIGS thin-film solar cell grown on Mo-coated soda-lime glass substrate. A typical CIGS thin-film solar cell is fabricated by first depositing a 2-μm thick CIGS absorber layer on an Mo-coated (for bottom contact) soda-lime glass (SLG) substrate using PVD or spray-paint technique, followed by deposition of a 30- to 50-nm CdS buffer layer on top of the CIGS films using chemical bath deposition (CBD), and then the deposition of a thin (50 nm) intrinsic ZnO and a 1-μm thick conducting ZnO (Al-doped) using the sputtering technique, and finally the Al-ohmic contact grids are applied for the top contacts. The $CuInSe_2$ (CIS) film has an energy band gap of 1.04 eV; by adding Ga to CIS to form the $Cu(In_{1-x},Ga_x)Se_2$ (CIGS) alloy, the band gap energy will increase from 1.04 eV for CIS to around

1.67 eV for $CuGaSe_2$(CGS) films. The optimum band gap for the CIGS cells is $E_g = 1.3$ eV with 35% of Ga incorporation in the CIGS films. Figure 12.15a shows a computer-simulated photo-J-V curve using the cell structure and the energy band scheme shown in Figures 12.14a and b. A computer simulation tool AMPS-1D (analysis of microelectronic and photonic structures)[6] was used in the simulation. The results are compared with the published data for a high-efficiency NREL CIGS cell. Excellent agreement was obtained between the simulated and measured data for the NREL high-efficiency CIGS cell (19.8%AM 1.5G). Figure 12.15b shows a comparison of the simulated quantum efficiency (QE) versus wavelength curve with the experimental curve for the same cell shown in Figure 12.15a. Excellent agreement was also obtained in this case in the wavelength (λ) range of 0.5 μm $\leq \lambda \leq 1.2\,\mu$m.[7]

The conversion efficiency could increase to over 21% under high concentration (e.g., 100 suns) of sunlight. Flexible CIGS solar cell modules fabricated on a plastic sheet (e.g., polyamide substrates) with conversion efficiency of over 10% AM1.5G have also been reported by Global Solar Inc. in the United States. Large-scale production of megawatt CIGS solar cell modules for commercial power generation has been achieved in Japan, Germany, and the United States. Further improvement in module efficiency and long-term reliability issues should make the CIGS thin-film solar cells a viable PV technology for producing low-cost photovoltaic systems for terrestrial power generation.

(iii) CdTe thin-film solar cells. Another promising thin-film PV technology is based on CdTe thin-film solar cells. The CdTe is a direct band gap semiconductor with an energy band gap of $E_g = 1.45$ eV, which matches the peak solar irradiance spectrum, and is an ideal semiconductor material for high-efficiency solar cell fabrication. Figure 12.16a shows the schematic energy band diagram of a CdS/CdTe thin-film solar cell. A high-efficiency CdS/CdTe thin-film solar cell typically uses the chemical bath deposition (CBD) and close-space sublimation (CSS) methods for the deposition of CdS and CdTe layers, respectively. The CdS/CdTe is deposited directly onto the TCO (transparent conducting oxides) such as SnO_2- and In_2O_3-coated glass substrates. The CSS process is one of the most efficient methods to control the grain size of the deposited CdTe films, which has a direct effect on the cell performance. However, regardless of the average grain size obtained, CSS CdTe deposition results in a randomized polycrystalline layer. Because of the relatively easy manufacturability and a good absorption coefficient in the visible sun spectrum, CdS/CdTe cell efficiency has improved significantly over the past decade. The maximum conversion efficiency reported for the CdS/CdTe thin-film solar cell is 16.5% AM1.5G by the NREL research group. Figure 12.16b shows the cross-sectional view of a high-efficiency CdS/CdTe thin-film solar cell prepared by a process suitable for large-scale production. The cell is fabricated on a soda-lime glass substrate, which consists of five layers: a 500-nm TCO (In_2O_3) and a 100-nm CdS are first deposited using sputtering, an 8-μm thick CdTe deposited using CSS process, a 150-nm Sb_2Te_3, and a 150-nm Mo contact layer

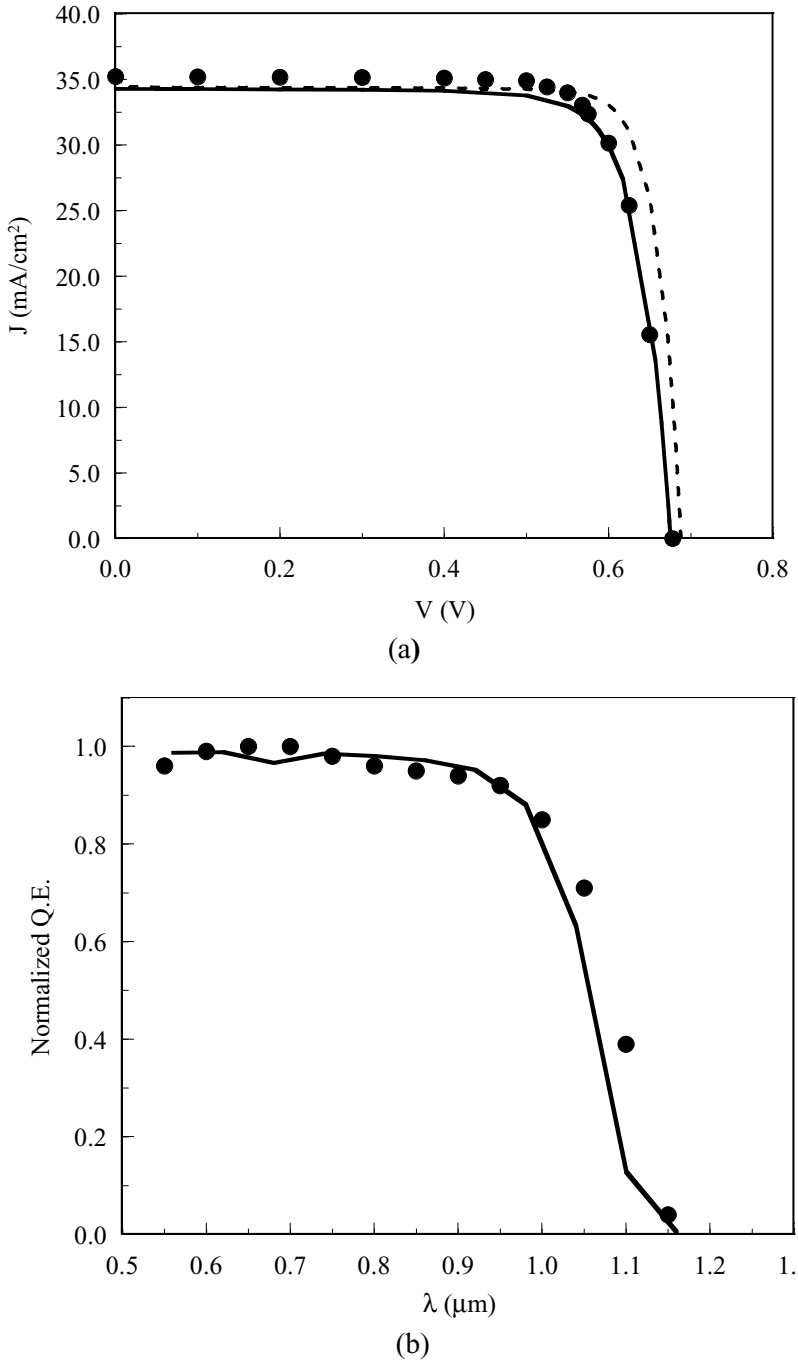


FIGURE 12.15. (a) A comparison of the simulated (using AMPS-1D computer simulation tool) and the measured photo-*I*–*V* curves, and (b) a comparison of the simulated and measured Q.E. versus wavelength curve for a high-efficiency NREL CIGS solar cell.[6,7]

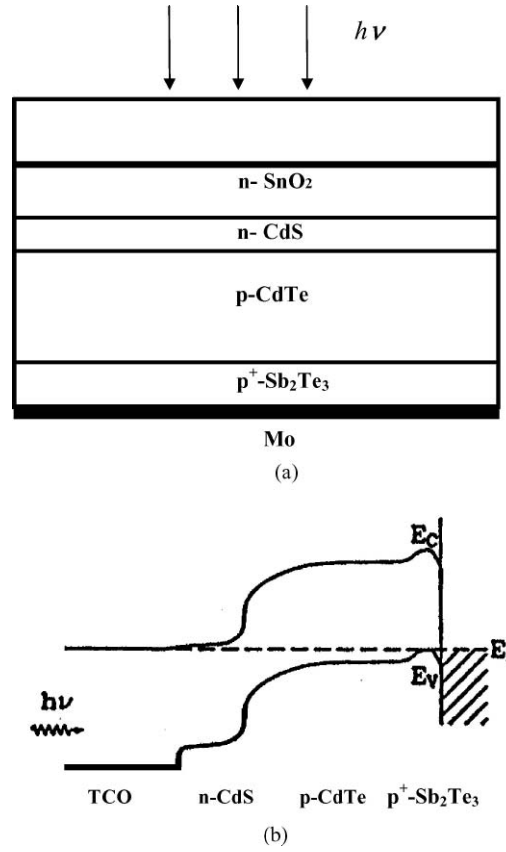


FIGURE 12.16. (a) Layer structure and (b) energy band diagram of a CdS/CdTe thin-film solar cell grown on ITO (Sn_2O) coated glass substrate.

by sputtering technique. Maximum efficiency for this solar cell is 14% AM1.5G with a $V_{oc} = 800$ mV, $J_{sc} = 25$ mA/cm^2, and FF $= 0.66$. Commercial CdTe thin-film solar cell modules are currently in production by First Solar Inc., in the United States.

12.2.7. Multijunction Tandem Solar Cells

Because of the band gap limitation, a single-junction solar cell can only utilize a portion of the sun's spectrum to convert useful sunlight into electricity. In order to further increase the conversion efficiency of p-n junction solar cells, it is necessary to employ the multijunction approach. A high-efficiency multijunction solar cell can be fabricated using semiconductors of different band gaps to form individual p-n junction cells that can absorb photons from different spectral regions of the solar irradiance. These individual cells can be stacked mechanically to form 2-, 3-, or 4-junction cells or interconnected by tunnel junctions for monolithically integrated multijunction tandem cells. For example, a typical triple-junction solar cell is composed of a top cell that is usually formed using a wide-band-gap material

such as GaInP or AlGaAs; the middle cell is formed using a medium-band-gap material such as GaAs, while the bottom cell can be formed using a smaller-band-gap material such as InGaAs or Ge. The triple-junction solar cells can be stacked mechanically on top of each other or interconnected by tunnel junctions to form a 3-junction tandem solar cell.

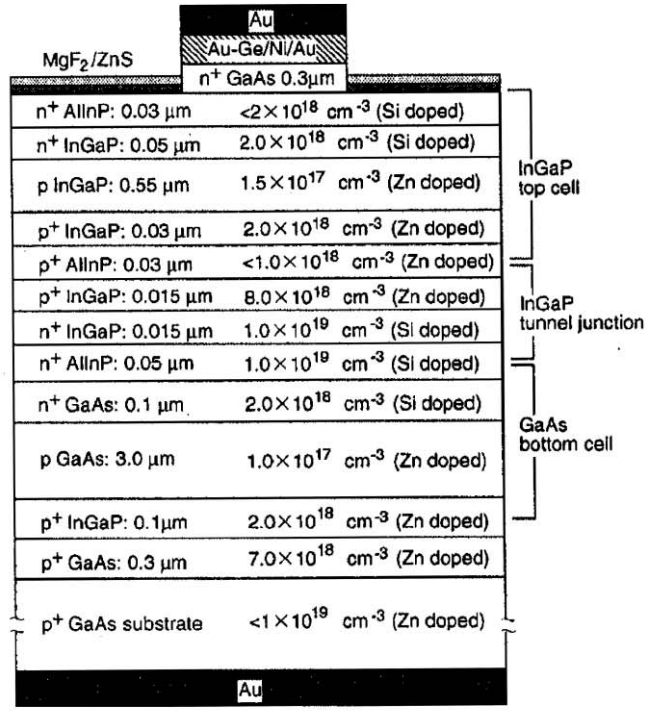
A two-terminal monolithic InGaP/GaAs tandem solar cell with a conversion efficiency of 30.28% under 1-sun AM1.5G conditions has been reported recently by Takamoto et al.[8] The improvements of the tandem cell performance are achieved by using a double-hetero (DH) structure InGaP tunnel junction, in which the InGaP layers are surrounded by wide-band-gap AlInP barriers. The DH structure by the AlInP barriers increases the peak current of the InGaP tunnel junction. The AlInP barrier directly below the InGaP top cell, which takes the part of a BSF layer, was found to be quite effective in reflecting the minority carriers in the top cell. The AlInP BSF layer not only forms a high-potential barrier but also prevents the diffusion of zinc from a highly doped tunnel junction toward the top cell during epitaxial growth. Furthermore, an InGaP tunnel junction reduces the absorption loss, which exists in a GaAs tunnel junction, and increases the photogenerated current in the GaAs bottom cell. Figure 12.17a shows the detailed structure of this InGaP/GaAs 2-junction tandem cell, and Figure 12.17b illustrates the spectral responses for the top and bottom cells and the effect of the tunnel junctions on the spectral response of the top and bottom cells. The InGaP top cell absorbs photons with $h\nu \geq 1.85$ eV, while the GaAs bottom cell absorbs photons with energies of $1.4\text{ eV} \leq h\nu \leq 1.85$ eV. Some optical losses are expected in the tunneling junction connecting the top and bottom cells.

More recently, Yamaguchi et al.[9] reported a mechanically stacked InGaP/GaAs/InGaAs 3-junction solar cell (1 cm^2) with a conversion efficiency of 33.3% under 1-sun AM1.5G conditions. The multijunction solar cell structures have also been applied to other material systems to achieve high conversion efficiency. For example, an a-Si/a-SiGe/mc-SiGe 3-junction thin-film solar cell with conversion efficiency as high as 14% AM1.5G has been reported recently.[10]

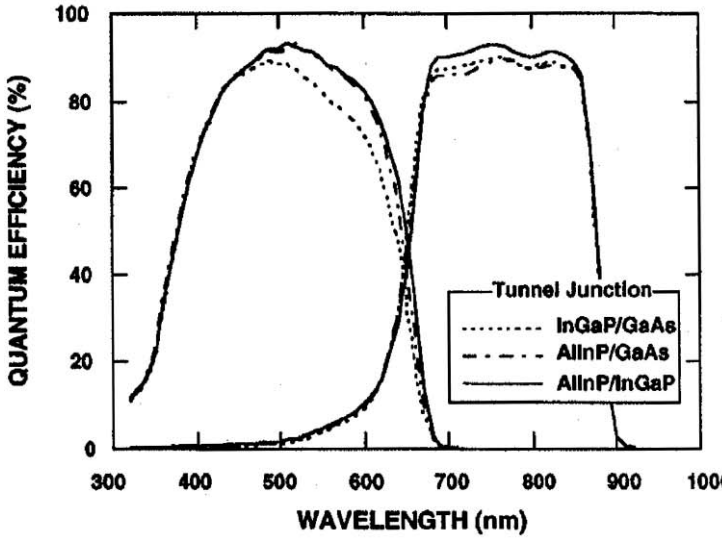
Figure 12.18 shows the evolution of the multijunction solar cell research that covers the state-of-the-art 2- and 3-junction solar cells based on InGaP/GaAs and InGaP/GaAs/Ge (or InGaAs) material systems and a proposed 4-junction solar cell using an InGaP/GaAs/InGaNAs/Ge material system for future development.[11] Theoretical conversion efficiency over 40% AM1.5G is predicted for the proposed 4-junction tandem solar cell. The multijunction solar cells are particularly attractive for space power generation and for terrestrial concentrator cell applications. Table 12.1 summarizes the confirmed conversion efficiencies for the single-junction, thin-film, and multijunction solar cells fabricated from a wide variety of semiconductor material systems.[12]

12.2.8. Concentrator Solar Cells

Single-junction solar cells are typically encapsulated in flat-panel weatherproof modules. Solar cells cover the entire flat-plate module area and are uniformly illuminated under 1-sun conditions. In contrast, a solar concentrator electric

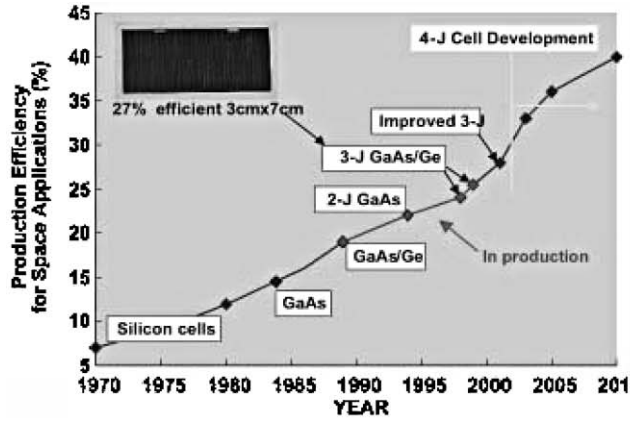


(a)



(b)

FIGURE 12.17. A high-efficiency InGaP/GaAs 2-junction tandem solar cell: (a) layer structure and (b) spectral response for the top and bottom cells, showing the effect of tunnel junction interconnects on spectral response. After Takamoto et al.[8]



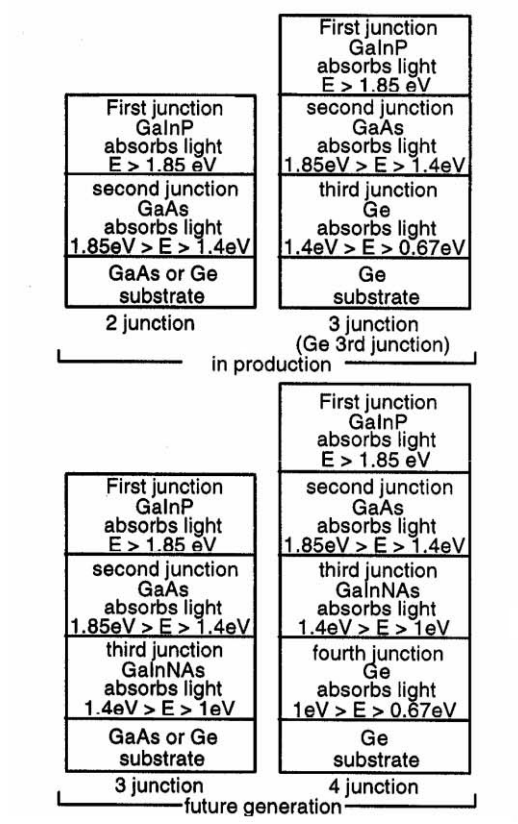


FIGURE 12.18. Evolution of multijunction solar cell structures from the existing InGaP/GaAs 2-junction cell to 3- and 4-junction future generation solar cells by incorporating a 1 eV 3rd junction subcell (using InGaNAs). Tunnel junctions are used to interconnect the individual subcells.[11]

TABLE 12.1. Confirmed terrestrial solar cell and module efficiencies measured under the global AM1.5G spectrum (100 mW/cm^2) for various single-junction, thin-film, and multijunction solar cells.[12]

Classification[a]	Effic[b] (%)	Area[c] (cm^2)	V_{oc} (V)	J_{sc} (mA/cm^2)	FF[d] (%)	Test center[e] (and date)	Description
Silicon cells							
Si (crystalline)	24.7 ± 0.5	4.00 (da)	0.706	42.2	82.8	Sandia (3/99)	UNSW PERL[3]
Si (multicrystalline)	19.8 ± 0.5	1.09 (ap)	0.654	38.1	79.5	Sandia (2/98)	UNSW/Eursolare[3]
Si (supported film)	16.6 ± 0.5	0.98 (ap)	0.608	33.5	81.5	NREL (3/97)	AstroPower (Si-Film)[4]
III-V cells							
GaAs (crystalline)	25.1 ± 0.8	3.91 (t)	1.022	28.2	87.1	NREL(3/90)	kopin, AlGaAs window
GaAs (thin film)	23.3 ± 0.7	4.00 (ap)	1.011	27.6	83.8	NREL(4/90)	Kopin, 5 mm CLEFT[5]
GaAs (multicrystalline)	18.2 ± 0.5	4.011 (t)	0.994	23.0	79.7	NREL(11/95)	RTI, Ge substrate[6]
InP (crystalline)	21.9 ± 0.7	4.02 (t)	0.878	29.3	85.4	NREL(4/90)	Spire, epitaxial[7]
Polycrystalline thin film							
CIGS (cell)	18.4 ± 0.5	1.04 (t)	0.669	35.7	77.0	NREL(2/01)	NREL, CIGS on glass[10]
CIGS (submodule)	16.6 ± 0.4	16.0 (ap)	2.643	8.35	75.1	FhG-ISE(3/00)	U. Uppsala, 4 serial cells[11]
CdTe (cell)	16.4 ± 0.5	1.131 (ap)	0.848	25.9	74.5	NREL (2/01)	NREL, on glass
CdTe (submodule)	10.6 ± 0.3	63.8 (ap)	6.565	2.26	71.4	NREL (2/95)	ANTEC [12]
Amorphous Si							
a-Si (cell)[f]	12.7 ± 0.4	1.0 (da)	0.887	19.4	74.1	JQA (4/92)	Sanyo [13]
a-Si (submodule)[f]	12.0 ± 0.4	100 (ap)	12.5	1.3	73.5	JQA (12/92)	Sanyo[14]
Photochemical							
Nanocrystalline dye	6.5 ± 0.3	1.6 (ap)	0.769	13.4	63.0	FhG-ISE (1/97)	INAP
Nanocrystalline dye (submodule)	4.7 ± 0.2	141.4 (ap)	0.795	11.3	59.2	FhG- ISE (2/98)	INAP
Multijunction cells							
GaInP/GaAs	30.3	4.0 (t)	2.488	14.22	85.6	JQA (4/96)	Japan Energy (monolithic)[15]
GaInP/GaAs/Ge	28.7 ± 1.4	29.93 9 (t)	2.571	12.95	86.2	NERL (9/99)	Spectrolab (monolithic)
GaAs/CIS (thin film)	25.8 ± 1.3	4.00 (t)	–	–	–	NREL (11/99)	Kopin/Boeing (4 terminal)
a-Si/CIGS (thin film)[f]	14.6 ± 0.7	2.40 (ap)	–	–	–	NREL (6/88)	ARCO (4 terminal)[16]

[a] CIGS = GaInGaSe$_2$: a-Si = amorphous silicon/hydrogen alloy.
[b] Effic = efficiency.
[c] (ap) = aperture area; (t) = total area; (da) = designated illumination area.
[d] FF = fill factor.
[e] FhG-ISE = Fraunhofer-Institut for Solare Energiesysteme; GQA = Japan Quality Assurence.
[f] Unstabilized results.

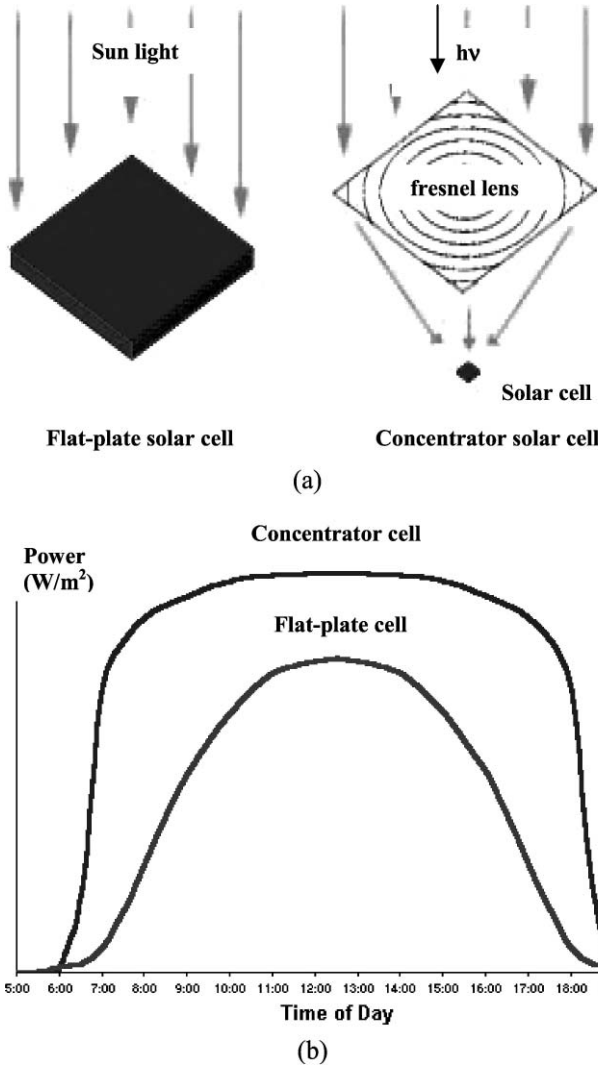


FIGURE 12.19. (a) Comparison of a flat-plate solar cell and a concentrating solar cell, and (b) the power out from a flat-plate cell and a concentrator cell as a function of time from 6:00 AM to 7:00 PM.

system uses a plastic Fresnel lens, similar to those found in overhead projectors, to concentrate sunlight manyfold before it reaches the solar cells. Alternatively, reflective mirrors can be used to concentrate the sunlight. Figure 12.19a shows a comparison of the flat-plate single-junction solar cell and the concentrator solar cell with Fresnel lens, and Figure 12.19b shows a comparison of the power output of the flat-plate solar cell and the concentrator solar cell as a function of the time of day from 6:00 AM to 7:00 PM. When sunlight concentrated 300-fold illuminates

a solar cell, the cell will produce about 300 times more power than it would without concentration. Concentrating lenses and solar cells are assembled together in a plastic housing to form a weatherproof concentrator module. Because of the concentration optics, concentrator modules must be pointed at the sun, so tracking systems are required to follow the sun across the sky during the day. The basic economics of solar concentration are simple: It replaces costly semiconductor solar cell area with lower-cost plastic lenses, which leads to lower overall system cost. For a concentrator module, the solar cell area is a small fraction of the total module area. In addition, the solar concentrator conversion efficiency is significantly higher than flat-plate PV technologies. Because the solar cell is a lower percentage of the overall concentrator system cost, using very high efficiency multijunction solar cells for a concentrator system makes good economic sense. It delivers more kilowatt-hours per day than the flat-panel solar modules. Total energy delivered, not peak power, is what really counts in electricity generation. Flat-plate solar systems are typically rated by peak wattage, when mounted with a fixed orientation they produce this amount of power only once during the day, at solar noon, as shown in Figure 12.19b. Because of solar tracking, the concentrator solar cell produces dramatically more power than a flat-plate cell in early morning and late afternoon hours. In addition, owing to its high efficiency, the concentrator system produces more power than a flat-plate system even at solar noon. The combined benefits of solar tracking and high efficiency allow the concentrator system to produce much more energy per day than the flat-plate system (the areas under the curves in the figure). A concentrator will produce more kilowatt-hours per day (energy) for the same module watt (power) rating. Furthermore, solar tracking also provides much more level power production throughout the day than a fixed-orientation flat-plate system, as seen in Figure 12.19b. This not only leads to higher daily energy production, but also has significant load matching advantages. For example, in sunny climates peak electric load due to air-conditioning occurs late in the afternoon rather than at solar noon. If renewable solar electricity is to make a significant contribution to the world's electricity generation needs in the twenty-first century, an absolutely monumental scale-up of manufacturing capacity will be required. For example, if solar power is to supply 10% of the forecasted world demand for new electric generation capacity, one will have to produce nearly 80 square miles of solar modules per year. Because of concentration, solar concentrators will require about 300 times less solar cell area than competing flat-plate PV technologies. Yamaguchi et al.[9] reported a monolithically integrated InGaP/InGaAs/Ge 3-junction concentrator solar cell with a conversion efficiency of 36% at 100 suns AM1.5G illumination. Spectral Lab has also reported a GaInP/GaInAs/Ge 3-junction concentrator solar cell with a world record efficiency of 36.9±1.8% AM1.5G at 309 suns (30.9 W/cm^2) intensity and 25°C, with $V_{oc} = 2.892$ V, $J_{sc} = 4.608$ A/cm^2, FF = 85.52%, and $V_{mp} = 2.591$ V. Using concentrated sunlight, these 3-junction concentrator solar cells can convert 36.9% of the sun's energy into electricity, a technology capability that could dramatically reduce the cost of generating electricity from solar energy. Table 12.2 summarizes the confirmed conversion efficiencies of the terrestrial concentrator solar cells and

TABLE 12.2. Terrestrial concentrator solar cell and module efficiency measured under AM1.5G spectrum at $T_c = 25°C$ for, 1-, 2- and 3-junction solar cells and modules.[12]

Classification	Effic[a] (%)	Effective area[b] (cm^2)	Actual area[c] (cm^2)	Intensity[d] (suns)	Test center (and date)	Description
Single cells						
GaAs	27.6 ± 1.0	32	0.126 (da)	255	Sandia (5/91)	Spire[17]
GalnAsP	27.5 ± 1.4	13	0.075 (da)	171	NREL (2/91)	NREL, Entech cover
Si	26.8 ± 0.8	154	1.60 (da)	96	FhG-ISE(10/95)	SunPower back-Contact[18]
InP	24.3 ± 1.2	7	0.075 (da)	99	NREL (2/91)	NREL, Entech cover[19]
CIGS (thin film)	21.5 ± 1.5	1	1.02 (da)	14	NREL (2/01)	NREL
2-cell stacks						
GaAs/GaSb (4 terminal)	32.6 ± 1.7	5	0.053 (da)	100	Sandia[e] (10/89)	Boeing, mechanical stack[20]
InP/GalnAs (3 terminal)	31.8 ± 1.6	3	0.063 (da)	50	NREL(8/90)	NREL, monolithic[21]
GalnP/GaAs (2 terminal)	30.2 ± 1.4	19	0.103 (da)	180	Sandia (3/94)	NREL, monolithic[22]
GaAs/Si(large)	29.6 ± 1.5	111	0.317 (da)	350	Sandia[e](9/88)	Varian/Stanford/Sandia, mech, Stack[23]
3-cell stacks						
GalnP/GaAs/Ge	32.4 ± 2.0	42	0.1025 (da)	414	NREL (6/00)	Spectrolab, monolithic[24]
GalnP/GaAs/Ge (large)	30.6 ± 1.5	246	1.050 (da)	234	NREL (9/00)	Spectrolab, monolithic
Submodules						
GaAs/GaSb	25.1 ± 1.4	41	41.4 (ap)	57	Sandia(3/93)	Boeing 3 mech, stack units[25]
GalnP/GaAs/Ge	27.0 ± 1.5	34	34 (ap)	10	NREL (5/00)	ENTECH[26]
Modules						
Si	20.3 ± 0.8	1875	1875 (ap)	80	Sandia (4/89)	Sandia/UNSW/ENTECH(12 cells)[27]
"Notable exceptions"						
GalnP/GaAs/Ge(2 terminal)[f]	34.0 ± 1.5	221	1.05 (da)	210	NREL (9/00)	Spectrolab, Global spectrum
Si (large)	21.6 ± 0.7	220	20.0 (da)	11	Sandia* (9/90)	UNSW laser grooved[28]
GaAs(Si substrate)	21.3 ± 0.8	30	0.126 (da)	237	Sandia (5/91)	Spire[17]
lnP(GaAs substrate)	21.0 ± 1.1	7	0.075 (da)	88	NREL (2/91)	NREL, Entech cover[29]

[a] Effic. = efficiency.
[b] Effective area for cells equals actual area multiplied by intensity (suns).
[c] (da) = designated illumination area; (ap) = aperture area.
[d] One sun corresponds to an intensity of 1000 Wm^{-2}.
[e] Measurements corrected from originally measured values due to Sandia recalibration in January 1991.
[f] Global AM 1.5 rather than direct beam.

modules measured under direct sunlight with AM1.5G spectrum at a cell temperature of 25°C.

12.3. Photodetectors

12.3.1. Introduction

Photodetectors and light-emitting devices (LEDs, LDs) are two important active elements for optoelectronic device applications. Since a large number of semiconductor LDs and LEDs have been developed for use in a broad-wavelength range from UV, to visible, to the IR spectrum, it is equally important to develop a wide variety of photodetectors for detection in the corresponding wavelengths of LDs and LEDs. In fiber-optic communications, the detectors must possess such features as low noise, high responsivity, and large bandwidth. Although high-sensitivity photomultipliers and traveling-wave phototubes are widely used for detecting modulated optical signals at microwave frequencies, recent trends are toward the use of various solid-state photodetectors including Schottky barrier, p-i-n, and APDs fabricated from elemental and compound semiconductors. A GaAs Schottky barrier photodetector with cutoff frequency greater than 100 GHz has been reported recently. High-speed photodetectors are particularly attractive for millimeter-wave fiber-optic links. $In_{0.53}Ga_{0.47}As$ p-i-n photodiodes with bandwidth greater than 30 GHz have been developed for 1.3- to 1.6-μm optical fiber communications. APDs made from InGaAs/InP, InGaAsP/InP, Ge, and GaAs with high internal current gains have also been developed for such applications.

In this section, various photodetectors including p-n junction and p-i-n photodiodes, APDs, Schottky barrier photodiodes, and heterojunction photodiodes are described. Since most of these photodiodes are based on depletion-mode operation (reverse-bias operation), they offer high-speed and high-sensitivity detection. A comparison of different types of solid-state photodetectors reveals that the intrinsic photoconductor has the highest internal gain ($G_p > 10^4$), while Schottky barrier photodiodes have the shortest response time ($\approx 10^{-11}$ s) and largest bandwidth. On the other hand, the APD has the highest-gain bandwidth product among all photodetectors.

A depletion-mode photodiode usually operates under small reverse-bias conditions. Under depletion-mode operation, the reverse saturation current (or dark current) is superimposed by the photocurrent produced by the incident photons in a photodiode. The applied reverse bias is usually not high enough to cause avalanche multiplication, and hence no internal current gain is expected in this operation mode. This is in contrast to an APD, in which an internal current gain is achieved as a result of avalanche multiplication near the breakdown conditions.

12.3.2. Key Physical Parameters and Figures of Merit

In order to evaluate the performance of a photodiode one needs to measure the spectral response, response speed, and noise figure under depletion-mode operation. The spectral response of a photodiode is determined in the wavelength range in

which an appreciable photocurrent can be measured. Key physical parameters affecting the spectral response are optical absorption coefficient, surface recombination velocity, and minority carrier lifetimes of a semiconductor from which the photodiode is fabricated. The cutoff wavelength, λ_c, of a photodiode is determined by the energy band gap of the semiconductor. For example, the energy band gap of silicon is equal to 1.12 eV at room temperature, and hence the cutoff wavelength for a silicon photodiode is around 1.1 μm (i.e., $\lambda_c = 1.24/E_g$ (eV) μm). Germanium has an energy band gap of $E_g = 0.67$ eV at 300 K, and hence its cutoff wavelength is around 1.8 μm. The short-wavelength limit is set by the wavelength at which the absorption coefficient of the semiconductor is in excess of 10^5 cm^{-1}. For wavelengths shorter than this value, the absorption of photons takes place mostly near the surface region of the photodiode, and hence electron–hole pairs generated in this region may recombine right near the surface and not reach the junction. Thus, for photodetectors with large surface recombination velocity, the photocurrent produced by the short-wavelength photons could be greatly reduced.

The important figures of merit for evaluating the performance of a photodetector are quantum efficiency (η), responsivity (R), noise equivalent power (NEP), and detectivity (D^*). These are discussed as follows:

(i) Quantum efficiency. The quantum efficiency, η, is widely used in assessing the spectral response of a photodiode, which can be defined by

$$\eta = \left(\frac{I_{ph}/q}{P_{in}/h\nu}\right) \times 100\% = \left(\frac{I_{ph}}{P_{in}\lambda}\right) \times 124\%, \tag{12.39}$$

where I_{ph} (A) is the photocurrent generated when a light beam with input power P_{in} (watts) and frequency ν falls onto the active area of the photodiode. The quantum efficiency η is determined at low-reverse-bias voltage in which no avalanche multiplication takes place. In (12.39), h is Planck's constant, q is the electronic charge, and λ is the wavelength of the incident photon. Figure 12.20 shows the

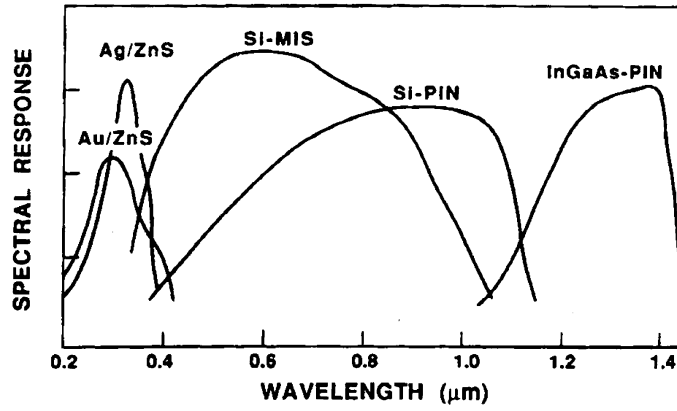


FIGURE 12.20. Relative spectral responses for several selected Schottky barrier, MIS, and p-i-n photodiodes.

relative spectral response curves for some Schottky barrier photodiodes and silicon MIS and p-i-n photodiodes.

(ii) Spectral responsivity. The spectral responsivity, R, is an important figure of merit that relates to the quantum efficiency of the photodetector. The responsivity is defined by the ratio of the photocurrent to the input optical power, which can be expressed by

$$R = \frac{I_{\text{ph}}}{P_{\text{in}}} = \frac{\eta\lambda\,(\mu\text{m})}{1.24} \qquad \text{A/W}, \tag{12.40}$$

where η is the quantum efficiency, and λ is the wavelength of incident photons. Equation (12.40) shows that the responsivity varies linearly with the wavelength of incident photons. For example, if the wavelength of incident photon is $\lambda = 0.62$ μm and the quantum efficiency of the photodetector is $\eta = 100\%$ at this wavelength, then the responsivity of the photodetector is $R = 1.0 \times 0.62/1.24 = 0.5$ A/W, and the responsivity R is equal to 1.0 A/W if the wavelength of incident photons is double (i.e., $\lambda = 1.24$ μm). For IR photodetectors, the responsivity is widely used in evaluating detector performance, since it is directly related to the detectivity D^* of the IR detector.

(iii) Noise equivalent power. Noise equivalent power (NEP) is another figure of merit widely used in assessing the performance of an IR photodetector. By definition, the NEP of an IR detector and its associated amplifier is the RMS (root mean square) value of the sinusoidally modulated optical power falling on a detector that gives rise to an RMS noise voltage referred to the detector terminal at a reference bandwidth of 1 Hz. For a monochromatic radiant flux (ϕ_{p}) with wavelength λ necessary to produce an RMS signal-to-noise ratio (SNR) of 1 at frequency f, the NEP is defined by

$$\text{NEP}(\lambda, f) = \left(\frac{hc}{\lambda}\right) q\phi_{\text{p}} \quad \text{watts}, \tag{12.41}$$

where hc/λ is the incident photon energy in eV, and ϕ_{p} is the RMS photon flux (i.e., photons/s) required to produce the SNR of 1. Equation (12.41) enables one to estimate the value of NEP due to different noise sources in a photodetector. It should be noted that a major limitation of NEP is due to two additional parameters, namely, the noise bandwidth (Δf) and the detector area (A_{d}), which must be given. Both are related to noise considerations—different Δf gives different noise values and smaller areas collect less power. In addition to the spectral NEP defined above, one can also consider the blackbody NEP, which is defined as the blackbody radiant flux (i.e., due to the background blackbody radiation) necessary to produce an RMS SNR of 1 at frequency f. As an example, consider a background-noise-limited detector performance, which is common in the IR spectral region. The spectral NEP for a background-limited infrared photodetector (BLIP) can be expressed by

$$\text{NEP}(\lambda, f) = \left(\frac{hc}{\lambda}\right) q\phi_{\text{p}} = \left(\frac{hc}{\lambda}\right) \left(\frac{2\phi_{\text{p}}^{\text{BG}} \Delta f}{\eta}\right)^{1/2}, \tag{12.42}$$

where ϕ_p^{BG} is the background photon flux (photons/s) that falls on the detector, Δf is the noise bandwidth, and η is the quantum efficiency. Values of NEP may vary from 10^{-8} to 10^{-13} W for a wide variety of IR detectors reported in the literature.

(iv) Spectral detectivity. Another figure of merit commonly used in an IR detector is known as the spectral detectivity $D^*(\lambda, f)$, which is defined as the SNR normalized per unit area per unit noise bandwidth, and is given by

$$D^*(\lambda, f) = \frac{A_d^{1/2}(\Delta f)^{1/2}}{\text{NEP}} \quad (\text{cm} \cdot \text{Hz}^{1/2})/\text{W}. \tag{12.43}$$

The spectral detectivity D^* is usually used for comparing the signal-to-noise performance of a photodiode having different active areas and operating at different noise bandwidths. Values of D^* have been found to vary from 10^8 to 10^{14} cm-$\text{Hz}^{1/2}$/W for various semiconductor photodetectors reported in the literature. In general, both D^* and NEP are frequently used in evaluating the performance of an IR detector, while the quantum efficiency and responsivity are often used in assessing the spectral response of a photodiode operating in the visible to near-IR spectral ranges. Using (12.42) and (12.43), the spectral detectivity under BLIP conditions can be expressed as

$$D^*_{BG}(\lambda, f) = \left(\frac{\lambda}{hc}\right)\left(\frac{\eta A_d}{2\phi_p^{BG}}\right)^{1/2}. \tag{12.44}$$

Figure 12.21 shows the plot of detectivity D^* as a function of wavelength for various photoconductors and photodiodes. The dashed curves are the theoretical ideal D^* at 77 and 300 K with a 360° field of view (FOV).

The response speed of a photodetector is discussed next. In general, the response speed of a p-n junction photodiode depends on three key factors, namely, the carrier diffusion time in the n and p quasineutral regions, the carrier drift transit time across the depletion layer, and the RC time constant of the detector system. In a depletion-mode photodiode, excess electron–hole pairs are generated inside the depletion region and quasineutral regions of the photodiode, and are collected as photocurrent across the junction of the photodiode. Since the minority carrier diffusion in the quasineutral regions is usually slower than the drift of excess carriers in the depletion region, high-speed detection is achieved by generating excess carriers inside the depletion region or close to the junction so that the diffusion time of the excess carriers is comparable to the transit time across the depletion region. For most semiconductors, the saturation-limited velocity ($v_s = (3k_BT/m^*)^{1/2}$) of the excess carriers generated inside the junction space-charge region of the photodiode is about 1 to 2×10^7 cm/s. Since the depletion layer width for most p-n junction photodiodes is only a few micrometers or less, the carrier transit time in the picosecond range can be readily achieved in a depletion-mode photodiode. Since the response speed or the bandwidth of a depletion-mode photodiode is determined by the three time constants discussed above, the 3-dB cutoff frequency of a p-n junction or Schottky barrier photodiode can be calculated

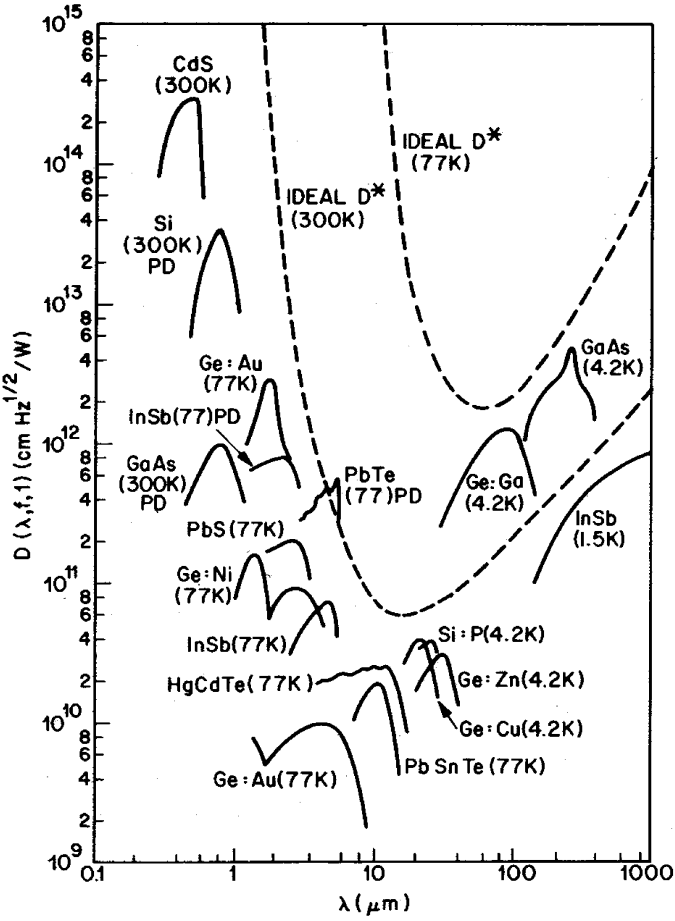


FIGURE 12.21. Detectivity (D^*) versus wavelength for various photoconductors and photodiodes. The dashed lines are theoretical ideal D^* at 77 K and 300 K with FOV $= 2\pi$. After Sze.[13]

using the expression

$$f_c = \frac{0.35}{(t_{tr}^2 + t_{dif}^2 + t_{RC}^2)^{1/2}}, \tag{12.45}$$

where

$$t_{tr} = \frac{W_d}{2.8 v_s}, \tag{12.46}$$

$$t_{dif} \approx \frac{W_p}{2.43 \tau_n}, \tag{12.47}$$

$$t_{RC} = \frac{1}{RC}. \tag{12.48}$$

Note that t_{tr} is the carrier transit time across the depletion layer region of width W_d, v_s ($\approx 10^7$ cm/s) is the saturation velocity, t_{dif} and W_p are the electron diffusion time constant and the width of the p-base region, τ_n is the electron lifetime, and t_{RC} is the RC time constant.

In a practical detector system, however, the cutoff frequency of a photodetector is usually lower than that predicted by (12.45) because of the finite load resistance and stray capacitances from the load resistance and the amplifier circuit. Fast photodiodes may be fabricated using a planar structure on a semi-insulating substrate with small active area (e.g., diameter less than 10 μm) to keep the diode capacitance and series resistance (or RC time constant) low. The point-contact Schottky barrier photodiode has the highest response speed and bandwidth among all photodetectors discussed in this section.

For a photodetector, the ultimate limitation on its performance is the noise generated in the detector. In general, the noises generated in a photodiode under reverse-bias conditions consist of the shot noise, $1/f$ noise (or flicker noise), and thermal noise (or Johnson noise). Shot noise is created by reverse leakage current flowing through the photodiode and is given by

$$i_s^2 = 2qI_D\Delta f, \tag{12.49}$$

where I_D is the dark current of the photodiode and Δf denotes the noise-equivalent bandwidth. For frequencies below 1 kHz, the noise of a photodiode is usually dominated by $1/f$ noise ($I_f^2 = BI_{dc}\Delta f/f$), which has a current-dependent power spectrum inversely proportional to the signal frequency. The origin of the flicker noise can be attributed to fluctuation associated with generation-recombination of excess carriers in a photodiode. In the intermediate frequency range (1 kHz $< f <$ 1 MHz), the generation-recombination noise becomes the dominant component. At high frequencies ($f >$ 1 MHz) the photodetector is dominated by white noise (i.e., independent of frequency) that includes shot noise, thermal noise, and generation-recombination noise. It should be noted that the break points of the frequencies for each of these noise sources vary from material to material.

Thermal noise is usually generated by random motion of carriers through the series resistance of the photodetector and load resistance. For photodiodes using a guard-ring structure, the channel resistance must also be included. The thermal noise of a photodiode can be calculated using the expression

$$i_{th}^2 = 4k_BTG\Delta f, \tag{12.50}$$

which shows that the thermal noise of a photodiode varies with the square root of the product of temperature (T), noise bandwidth (Δf), and diode conductance (G).

12.3.3. p-n Junction Photodiodes

In this section, the basic principles and general characteristics of a p-n junction photodiode are discussed. Figure 12.22a shows the schematic diagram of a p-n junction photodiode under reverse-bias conditions. Electron–hole pairs are generated by the

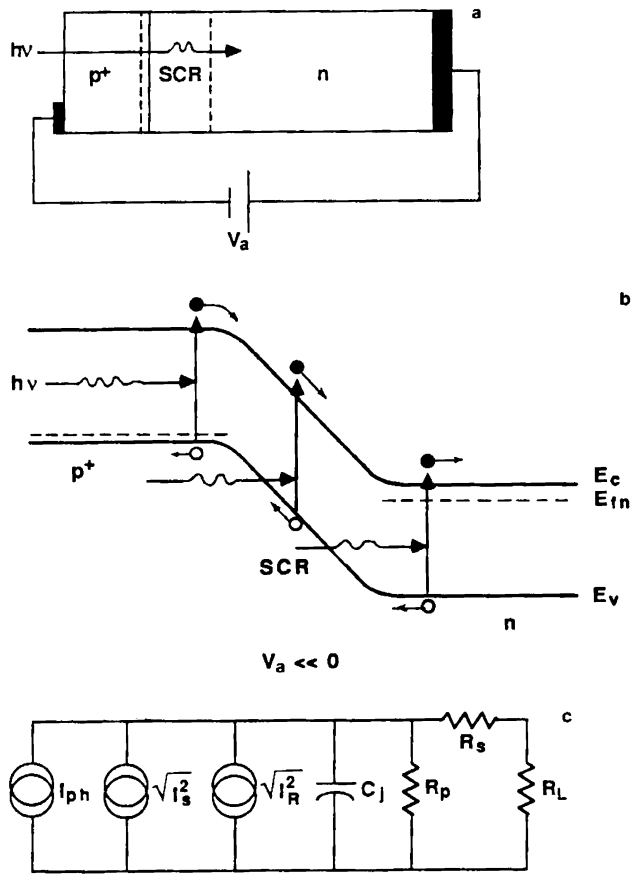


FIGURE 12.22. (a) Schematic representation of a p-n junction photodiode, (b) energy band diagram under illumination and reverse-bias conditions, and (c) equivalent circuit diagram: I_{ph} is the photocurrent; I_s the shot-noise current source; I_R the thermal-noise current source; C_j the junction capacitance; R_s and R_p denote the series and shunt resistances, and R_L the load resistance.

internal photoelectric effect in the photodiode to a depth on the order of $1/\alpha$, where α is the optical absorption coefficient at the wavelength of interest. As shown in Figure 12.22b, under reverse-bias conditions, the photogenerated electron–hole pairs are separated in the depletion region by the built-in electric field and collected as photocurrent in the external circuit. The small-signal equivalent circuit for a p-n junction photodiode is shown in Figure 12.22c, which consists of a photocurrent source I_{ph}, junction transition capacitance C_j, series resistance R_s, and shunt resistance R_p. The shunt resistance is usually very large and can be neglected for a typical photodiode operating in the visible spectral range, but is included to account for the possible leakage current path (i.e., low R_p) in a photodiode fabricated from small-bandgap semiconductors. The symbols $\sqrt{I_s^2}$ and $\sqrt{I_R^2}$ shown in

Figure 12.22c are the equivalent noise current sources due to the shot noise and thermal noise of the photodiode, respectively.

A high-speed p-n junction photodiode is usually constructed in such a way that most of the photons are absorbed in the p-emitter region. The junction is placed as deep as possible so that efficient separation of photogenerated electron–hole pairs can be achieved. This ensures that most of the photocurrent is carried out by electrons whose speed, either by diffusion or drift, is always faster than that of holes. The conditions for achieving excellent low-frequency response in a p-n junction photodiode are that $sW_p/D_n < W_p/(D_n\tau_n)^{1/2} < 1$ and $W_d/(v_s\tau_n) < 1$, where s is the surface recombination velocity, W_p is the width of the p region, D_n is the electron diffusion constant, τ_n is the electron lifetime in the p region, W_d is the depletion layer width, and v_s is the saturation velocity of electrons in the depletion region. The diffusion time constant for the photogenerated electron–hole pairs in the p region is given by (12.47), which is valid for $\alpha W_p < 1$ and for uniform doping in the p region. If an impurity concentration gradient is present in the p region, then faster detection can be expected owing to the built-in drift field created by the impurity concentration gradient. Since a large impurity concentration gradient is difficult to obtain in the thin diffused p region, the maximum transit time reduction by the field-assisted diffusion is about a factor of 5–10. The drift transit time governed by the electric field in the depletion region is given by (12.46).

The power available from a p-n junction photodiode may be characterized by the power available in a conjugate matched load. Using the equivalent circuit shown in Figure 12.22c, this can be written as

$$P(\omega_m) = \frac{I_{ph}^2 R_p}{4\left(1 + R_s/R_p + R_s R_p C^2 \omega_m^2\right)}, \tag{12.51}$$

where I_{ph} is the photocurrent and ω_m is the frequency at which the photodiode is conjugately matched. For high-frequency operation, a match of the photodiode parameters with load impedance is normally required at frequencies $\omega_m \geq 1/C(R_p R_s)^{1/2}$, so that the maximum power output is given by

$$P(\omega_m) = \frac{I_{ph}^2}{4R_s C_j^2 \omega_m^2}, \tag{12.52}$$

where C_j, R_p, and R_s are the junction capacitance, shunt resistance, and series resistance of the photodiode shown in Figure 12.22c.

12.3.4. p-i-n Photodiodes

The p-i-n photodiode is the most commonly used detector structure in the visible to near-IR spectral range. Silicon p-i-n photodiodes are widely used in the 0.4–1.06 μm spectral range, while InGaAs/InP p-i-n photodiodes can extend the detection wavelengths to the 1.3–1.55 μm wavelength range. A p-i-n photodiode consists of a highly doped p^+-emitter layer, a wide undoped intrinsic layer (i region), and a highly doped n^+-base layer. Figure 12.23a shows the schematic diagram of a p-i-n

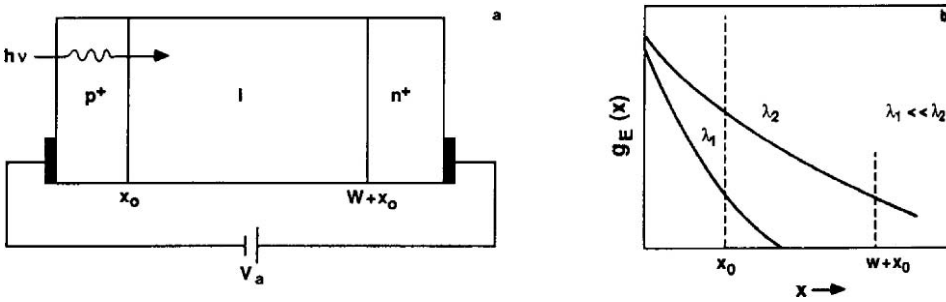


FIGURE 12.23. (a) Schematic diagram of a p-i-n photodiode, and (b) generation rate versus distance for two different wavelengths, where λ_1 denotes the short-wavelength photons, and λ_2 the long-wavelength photons.

photodiode, and Figure 12.23b shows the photogeneration rates versus distance for two different wavelengths. The reason that a p-i-n photodiode is so popular is that the spectral response near the cutoff wavelength region can be tailored to meet specific detection requirements. As shown in Figure 12.23, the long-wavelength spectral response of a p-i-n photodiode can be greatly improved by increasing the thickness (W) of the undoped i layer. In addition to the p-i-n (p^+-π-n^+) structure shown in Figure 12.23a, other p-i-n photodiode structures such as n^+-ν-p^+, n^+-π-p^+, and p^+-ν-n^+ junctions can also be fabricated (note that ν denotes n^-, and π is for p^-).

A p-i-n photodiode usually operates under the depletion-mode condition in which a sufficiently large reverse bias is applied to the photodiode such that the entire i region is fully depleted. When photons with energies greater than the bandgap energy of the semiconductor (i.e., $h\nu \geq E_g$) are impinging on the photodiode, a small fraction of the short-wavelength photons will be absorbed in the p^+ region, while the majority of photons are absorbed in the i region. Excess electron–hole pairs generated in the i region are swept out by the high electric field created by the applied reverse-bias voltage across the photodiode. The photogenerated electron–hole pairs are then collected at the ohmic contacts of the photodiode. The carrier transit time across the i region can be calculated using the expression

$$t_r = \frac{W}{v_s}, \tag{12.53}$$

where W is the thickness of the i layer and v_s is the thermal velocity of the excess carriers ($v_s = (3k_B T/m^*)^{1/2}$) in the i region.

Photocurrents generated in each region of the p-i-n photodiode and the spectral responses are derived next. As shown in Figure 12.23a, the thickness of the p layer is denoted by x_0, and W is the i-layer thickness ($W \gg x_0$). When monochromatic light is impinging on a p-i-n photodiode at $x = 0$, the rate of generation of excess carriers is given by

$$g_E(x) = \alpha\phi_0(1 - R)\, e^{-\alpha x}, \tag{12.54}$$

where ϕ_0 is the photon flux density, R is the reflection coefficient at the surface of the p region, and α is the absorption coefficient. Under steady-state conditions, the total photocurrent density $J_{\rm ph}$ is equal to the sum of electron and hole current components produced by the incident photons at a given plane along the x-direction of the photodiode. This can be expressed by

$$J_{\rm ph} = J_{\rm n}(x_0) + J_{\rm p}(x_0) = J_{\rm n}(x_0) + J_{\rm p}(W) + J_{\rm i}, \tag{12.55}$$

where

$$J_{\rm i} = J_{\rm p}(x_0) - J_{\rm p}(W) \tag{12.56}$$

is the photocurrent density due to the hole generation in the i region.

The spectral dependence of the quantum yield η can be obtained by solving $J_{\rm n}(x_0)$, $J_{\rm p}(W)$, and $J_{\rm i}$ as functions of λ, α, W, and x_0. To derive expressions for $J_{\rm n}$, $J_{\rm p}$, and $J_{\rm i}$ in the three regions of a p-i-n photodiode, it is assumed that (1) the photogenerated excess carrier densities are small compared to the majority carrier density in both the n^+ and p^+ regions (i.e., $\Delta p \ll n_0$ and $\Delta n \ll p_0$), (2) the reverse-bias voltage is not large enough to cause avalanche multiplication in the i region, (3) the surface recombination velocity is very high at the illuminated surface such that $\Delta n(0) = 0$, (4) the excess carrier density at the edge of the i region is small enough so that the boundary condition $\Delta n(x_0) = 0$ holds at $x = x_0$, (5) the effect of the built-in electric field in the emitter region (i.e., the p region) is neglected, and (6) recombination of excess carriers in the depletion region is negligible. Assumption (3) is valid for most silicon p-n junction photodiodes, because impurity concentration at the surface of the p-emitter region is usually several orders of magnitude higher than that of the i region. As a result, the carrier lifetime at the surface is also expected to be much shorter than that of the bulk. Therefore, excess carriers generated at the surface will usually recombine before they are able to diffuse to the junction.

The photocurrents produced by the absorbed incident photons in the three regions of a p-i-n photodiode can be derived as follows:[14]

(i) The p region ($0 < x \leq x_0$). In the p-emitter region, the contribution of photocurrent is mainly due to the electron diffusion current generated in the p region, which can be evaluated at $x = x_0$. This photocurrent component can be derived by solving the continuity equation of excess electrons in the p region, given by

$$D_{\rm n}\frac{{\rm d}^2\Delta n}{{\rm d}x^2} - \frac{\Delta n}{\tau_{\rm n}} = -\alpha\phi_0\,(1-R)\;{\rm e}^{-\alpha x}. \tag{12.57}$$

The general solution of (12.57) is given by

$$\Delta n\,(x) = A\sinh\left(\frac{x_0 - x}{L_{\rm n}}\right) + B\cosh\left(\frac{x_0 - x}{L_{\rm n}}\right) - \frac{\alpha\phi_0\,(1-R)\,\tau_{\rm n}\,{\rm e}^{-\alpha x}}{\left(\alpha^2 L_{\rm n}^2 - 1\right)}. \tag{12.58}$$

Constants A and B in (12.58) can be determined using the boundary conditions $\Delta n(x) = 0$ at $x = 0$ and $x = x_0$, which yield

$$A = \frac{\alpha\phi_0\,(1-R)\,\tau_{\rm n}[1-\cosh(x_0/L_{\rm n})\,{\rm e}^{-\alpha x_0}]}{(\alpha^2 L_{\rm n}^2-1)\sinh(x_0/L_{\rm n})} \tag{12.59}$$

and

$$B = \frac{\alpha\phi_0\,(1-R)\,\tau_{\rm n}\,{\rm e}^{-\alpha x_0}}{(\alpha^2 L_{\rm n}^2-1)}. \tag{12.60}$$

The electron diffusion current density in the p-emitter region of the photodiode is obtained by substituting (12.59) and (12.60) into (12.58) and evaluating at $x = x_0$, assuming that $\alpha L_{\rm n} \gg 1$. This yields

$$\begin{aligned} J_{\rm n}(x_0) &= q D_{\rm n} \left.\frac{{\rm d}\Delta n\,(x)}{{\rm d}x}\right|_{x=x_0} \\ &= q\phi_0(1-R)\left\{{\rm e}^{-\alpha x_0} - \frac{1}{\alpha L_{\rm n}\sinh\left(\frac{x_0}{L_{\rm n}}\right)}\left[1-\cosh\left(\frac{x_0}{L_{\rm n}}\right){\rm e}^{-\alpha x_0}\right]\right\}, \end{aligned} \tag{12.61}$$

which shows the functional dependence of $J_{\rm n}(x_0)$ on the absorption coefficient α, the electron diffusion length $L_{\rm n}$, and the junction depth x_0.

(ii)] The i region ($x_0 \le x \le W$). In the undoped (i) layer, the drift current density contributed by the photogenerated excess carriers in this region is given by

$$J_{\rm i} = q\int_{x_0}^{x_0+W} g_{\rm E}\,(x)\,{\rm d}x = q\phi_0(1-R)({\rm e}^{-\alpha W} - {\rm e}^{-\alpha x_0}). \tag{12.62}$$

In (12.62), it is assumed that $W \gg x_0$ and $W + x_0 \approx W$; $g_{\rm E}(x) = \alpha\phi_0(1-R)\,{\rm e}^{-\alpha x}$ is the photon generation rate inside the photodiode. Figure 12.23b shows the photogeneration rate $g_{\rm E}(x)$ as a function of distance x for two different wavelengths.

(iii) The n region ($x \ge W + x_0$). In this region, the photogenerated excess carriers (i.e., holes) contribute to the hole diffusion current. The hole current density can be derived by solving the continuity equation for the excess hole density in the n region, which is given by

$$D_{\rm p}\frac{{\rm d}^2\Delta p}{{\rm d}x^2} - \frac{\Delta p}{\tau_{\rm p}} = -\alpha\phi_0(1-R)\,{\rm e}^{-\alpha x}, \tag{12.63}$$

where $D_{\rm p}$ and $\tau_{\rm p}$ denote the hole diffusion coefficient and hole lifetime, respectively. The hole diffusion current density $J_{\rm p}(x)$ can be obtained by solving (12.63) for $\Delta p(x)$ using the boundary conditions

$$\begin{aligned} \Delta p(x) &= -p_{\rm n0} \quad \text{at } x = W + x_0, \\ &= 0 \quad \text{for } x \to \infty. \end{aligned} \tag{12.64}$$

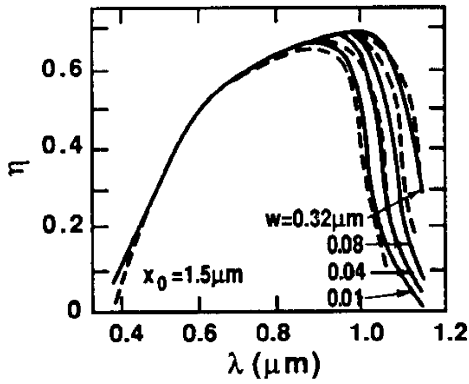


FIGURE 12.24. Quantum yield versus wavelength for a silicon p-i-n photodiode with different depletion layer widths. The solid lines are calculated from (12.62) and (12.63), and the dashed lines are experimental data. Li and Lindholm.[14]

The hole current density is obtained by evaluating $J_{\mathrm{p}}(x)$ at $x = W + x_0$, which yields

$$J_{\mathrm{p}}(W) = -qD_{\mathrm{p}}\frac{\mathrm{d}\Delta p(x)}{\mathrm{d}x}\bigg|_{x=x_0+W} = \frac{-q\phi_0(1-R)\alpha L_{\mathrm{p}}\,\mathrm{e}^{-\alpha W}}{(1+\alpha L_{\mathrm{p}})}. \tag{12.65}$$

Thus, the total photocurrent density produced in a p-i-n photodiode is equal to the sum of (12.61), (12.62), and (12.65), which is given by

$$J_{\mathrm{ph}} = q\phi_0(1-R)\left\{\frac{1}{\alpha L_{\mathrm{n}}\sinh\left(\frac{x_0}{L_{\mathrm{n}}}\right)}\left[1-\cosh\left(\frac{x_0}{L_{\mathrm{n}}}\right)\mathrm{e}^{-\alpha x_0}\right]-\frac{\mathrm{e}^{-\alpha W}}{(1+\alpha L_{\mathrm{p}})}\right\}. \tag{12.66}$$

The quantum efficiency η, defined as the number of electron–hole pairs generated per absorbed photon, can be expressed by

$$\eta = \frac{J_{\mathrm{ph}}}{q\phi_0}\times 100\%, \tag{12.67}$$

where J_{ph} is given by (12.66).

For a silicon p-i-n photodiode, the p region is usually very thin ($\leq 1.5\,\mu$m), while the i region is much wider (from a few μm to hundreds of μm). As a result, excess carriers generated by the short-wavelength photons are confined mainly in the p region. The quantum efficiency for a silicon p-i-n photodiode can be calculated using the optical absorption coefficient data and carrier diffusion lengths for silicon. Figure 12.24 shows the quantum yield versus photon wavelength for a silicon p-i-n photodiode with i-layer width W as a parameter and $x_0 = 1.5\,\mu$m. Note that solid lines are the calculated results using (12.67) for $W = 0.32, 0.08, 0.04$, and 0.01 cm, respectively. The reflection coefficient R for silicon is assumed equal to 0.3 in these calculations.

The sharp decrease of quantum yield in the short-wavelength regime can be explained as follows. As can be seen from (12.58) and Figure 12.23, the excess carriers generated by the short-wavelength photons are confined mainly in

the p region. Therefore, the photocurrent due to photon excitations in the short-wavelength regime must come from the excess carriers generated in the p region. However, only those excess carriers generated near the junction will diffuse toward the i region. In fact, only a fraction of the excess carriers generated by the short-wavelength photons in the p region will be collected and will contribute to the photocurrent $J_n(x_0)$. To improve the short-wavelength (or UV) response, the surface recombination loss must be minimized to preserve the excess carriers generated near the surface region so that photocurrent generated by the short-wavelength photons can be collected across the junction.

An estimate of the maximum cutoff frequency for a p-i-n photodiode operating under the condition limited by the load impedance R_L can be obtained using the equivalent circuit shown in Figure 12.21c. It is noted that the excess carriers generated in the i region that are separated by the built-in electric field can be represented by a current source I_{ph}, which is in parallel with the junction capacitance C_j. The series resistance is denoted by R_s, and R_L is the load resistance. The junction capacitance of the p-i-n photodiode is given by

$$C_j = \frac{A\varepsilon_0\varepsilon_s}{W}, \tag{12.68}$$

where A is the cross-sectional area of the junction, ε_s is the dielectric constant of the semiconductor, and ε_0 is the free space permittivity. The maximum cutoff frequency for a p-i-n photodiode can be calculated using the expression

$$f_c = \frac{2.4}{2\pi\tau_{tr}} \approx 0.4\alpha v_s, \tag{12.69}$$

where v_s is the average thermal velocity of electrons in the i region and α is the optical absorption coefficient. For a Ge p-i-n photodiode with $v_s = 6 \times 10^6$ cm/s, $\varepsilon_s = 16$, $A = 2 \times 10^{-4}$ cm^2, and $R_L = 10\Omega$, the cutoff frequency f_c was found to be 41.84 GHz.

If the light modulation frequency approaches that of the transit time of excess carriers across the entire i region, a phase shift between the photon flux and the photocurrent will appear in the photodiode. This effect is severe for the case in which incident photons are absorbed very close to the outer edge of the depletion layer or near the surface of the photodiode. However, for most semiconductors the absorption coefficients vary between 10 and 10^5 cm^{-1}; the optical absorption takes place quite deep inside the photodiode. If the modulation frequency is around 10^9 Hz, then the depletion layer width must be a few microns. This means that the excess carriers are generated throughout the entire volume of the depletion layer, and hence have a distribution of transit times. The phase-shift effect in this case is less severe than in the case of surface generation.

12.3.5. Avalanche Photodiodes

Avalanche photodiodes (APDs) are high-gain and high-speed photodetectors that have been extensively investigated for a wide variety of applications. An APD

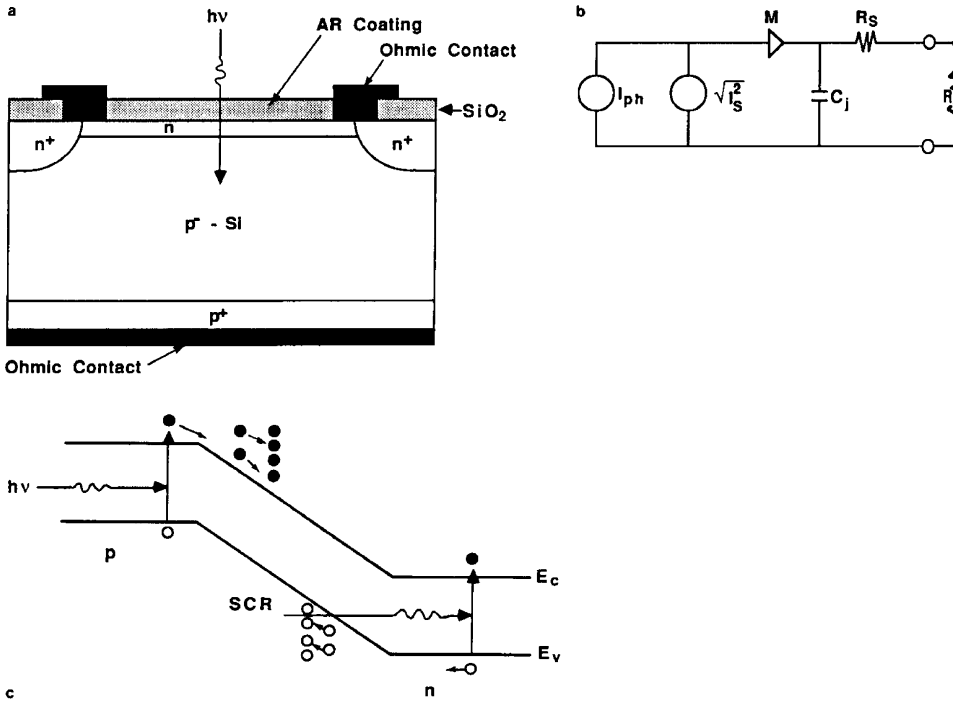


FIGURE 12.25. (a) Cross-sectional view of a planar silicon avalanche photodiode (APD) with a guard-ring structure, (b) the equivalent circuit, and (c) the energy band diagram under illumination and large-reverse-bias conditions showing avalanche multiplication in the space-charge region.

can have the dual function of serving as a photodetector for detecting incident photons or optical signals, and as an amplifier (with internal current gain via avalanche multiplication) for the excess carriers generated by the incident photons. An APD is known to produce the highest-gain bandwidth product among all the solid-state photodetectors discussed in this chapter. APDs can be fabricated from a wide variety of semiconductors using different device structures. Besides the conventional germanium and silicon APDs, various APDs fabricated from III-V compound semiconductors have also been reported, in particular the long-wavelength (1.3 and 1.55 μm) APDs using InGaAs/InP and InGaAsP/InP material systems. More recently, APDs fabricated from wide-band-gap materials such as GaN/AlGaN and GaP have also been reported for UV detection.

Figure 12.25a shows the cross-sectional view of a silicon planar APD. In this structure, the (p^+-n or n^+-p) guard-ring structure around the edge of the APD is employed to prevent the occurrence of microplasmas (i.e., small regions where the breakdown voltage is lower than that of the p-n junction as a whole) around the edge of the active region of the APD. The occurrence of bulk microplasmas can be reduced using detector-grade semiconductor materials with low defect densities.

Figure 12.25b shows the equivalent circuit of an APD, where M denotes the multiplication factor, i_s is the shot-noise current source, and R_s and C_j denote the series resistance and junction capacitance, respectively.

The basic principle underlying the operation of an APD is discussed next. If the reverse-bias voltage across a photodiode is smaller than the breakdown voltage under dark conditions, a small reverse leakage current will flow through the APD. The reverse leakage current I_D is due to thermally generated carriers in the quasineutral regions of the diode, and should be kept as low as possible. If the reverse-bias voltage continues to increase, the electric field in the depletion region will eventually become strong enough (i.e., $\mathcal{E} \geq \mathcal{E}_c$; $\mathcal{E}_c$ is the critical field) to cause both thermally generated electrons and holes to gain sufficient kinetic energy, and impact ionization will occur when electron–electron or hole–hole collisions take place at electric field strength $\mathcal{E} \geq \mathcal{E}_c$. These energetic electrons and holes will undergo further impact ionizations, which produce more electron–hole pairs and more impact ionization, resulting in a runaway condition limited only by the series resistance and the external circuitry of the APD. This runaway condition is called avalanche breakdown. Thus, it is understandable that the current flow in an APD during avalanche breakdown, I_{MD}, may be orders of magnitude larger than the initial thermally generated dark current I_D. Figure 12.25c shows the avalanche multiplication that occurs inside the junction space-charge region of a silicon APD. When an APD is illuminated by light under a small-reverse-bias condition, the primary current flowing through the APD is denoted by I_p, which consists of the dark current I_D and the photocurrent I_{ph} (due to electron–hole pairs generated optically within the depletion region). If the reverse-bias voltage is increased, avalanche multiplication of the primary current I_p occurs, but the onset of avalanche multiplication is usually gradual. This is clearly illustrated by the photo-I–V characteristics of a Ge APD shown in Figure 12.26, with prime current I_p as a parameter.[15]

In general, the photo-I–V characteristics of an APD can be predicted using an empirical formula

$$I = MI_p, \tag{12.70}$$

where

$$I_p = I_D + I_{ph}. \tag{12.71}$$

Here I_p is the primary current of the APD before the onset of avalanche multiplication, I_{ph} is the primary photocurrent produced by the incident photons in the APD, and I_D is the dark current given by

$$I_D = \left(q\sqrt{\frac{D_p}{\tau_p}}\frac{n_i^2}{N_D} + \frac{qn_iW}{\tau_e} \right) A_j. \tag{12.72}$$

Equation (12.72) consists of two terms: the first term is the thermal generated current from the n-quasineutral region of the APD, and the second term is the generation current produced in the space-charge region of the APD under

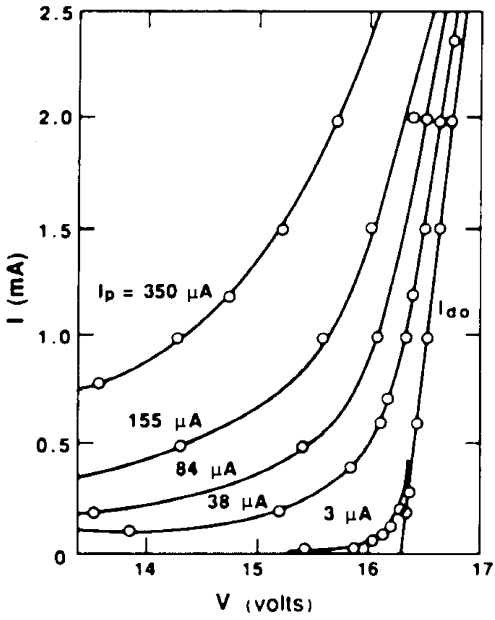


FIGURE 12.26. Current–voltage (*I–V*) characteristics of a germanium avalanche photodiode with different primary currents. After Melchior and Lynch,[15] by permission, © IEEE-1966.

reverse-bias conditions. The multiplication factor given in (12.70) can be described by an empirical formula given by

$$M = \frac{1}{\left\{1 - \left[\frac{(V - I_p R)}{V_B}\right]^n\right\}}, \tag{12.73}$$

where $R = R_s + R_c + R_T$ is the total resistance of the APD, R_s is the series resistance of the contacts and the bulk material, R_c is the resistance due to carrier drift through the depletion layer, and R_T is the thermal resistance that heats the junction and increases the diode breakdown voltage V_B. The factor n in the exponent depends on the semiconductor material used, the doping profile in the junction, and the wavelength of incident photons. A low value of n corresponds to high avalanche multiplication at a given bias voltage.

Although a large internal current gain can be achieved using an avalanche multiplication process, the shot noise also increases rapidly with the multiplication process in an APD. It is noted that a significant improvement in overall sensitivity has been achieved in both the Si and Ge APDs with wide instantaneous bandwidth. Figure 12.27 shows the signal and noise power outputs of a Ge APD operating at 1.5 GHz and $\lambda = 1.15\mu$m.[15] Silicon APDs are commercially available for fiber optic and very low light level applications. For example, Perkin Elmer type C30902E APD utilizes a silicon detector chip fabricated with a double-diffused "reach-through" structure. This structure provides high responsivity between 0.4 and 1.0 μm as well as extremely fast rise and fall times at all wavelengths. The responsivity of the device is independent of modulation frequency up to about

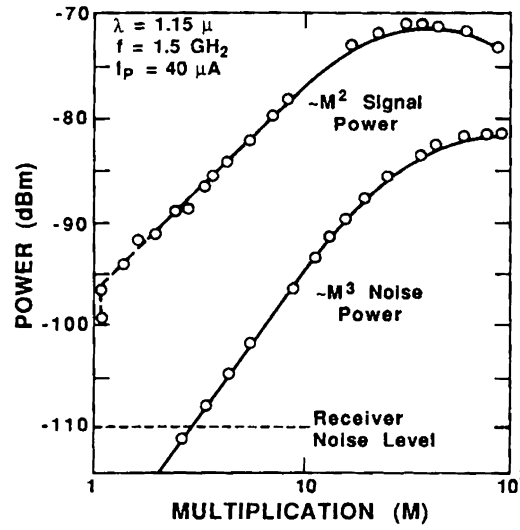
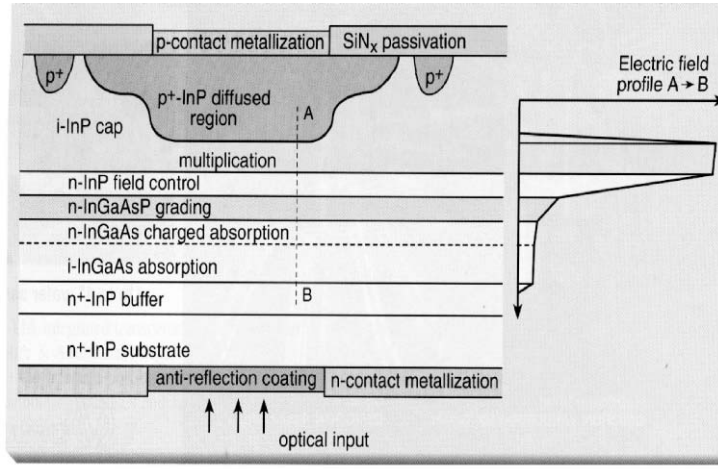


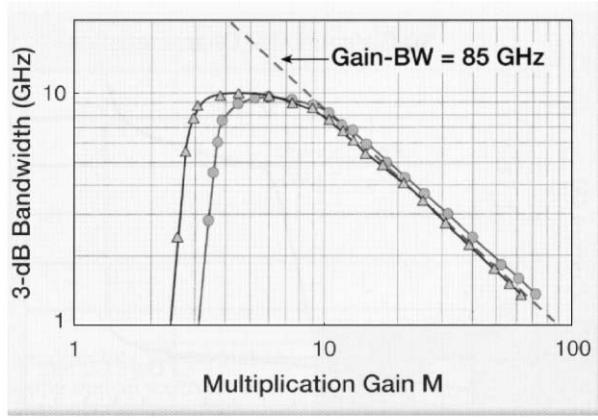
FIGURE 12.27. Signal and noise power outputs of a germanium avalanche photodiode. Both are measured with an input light intensity of 1 mW. After Melchior and Lynch,[15] by permission, © IEEE-1966.

800 MHz. Typical breakdown voltage for this type of APD is $V_B = 225$ V, $M = 150$–250, current responsivity $R_A = 65$–109 A/W, quantum efficiency $\eta = 60\%$ at 900 nm and 77% at 830 nm.

Figure 12.28 shows a novel InGaAs/InP SAM (Separate Absorption and Multiplication) APD grown by the molecular beam epitaxy (MBE) technique for a 1.3-μm fiber optic receiver application.[10] A SAM-APD uses a smaller-band-gap $In_{0.53}Ga_{0.47}As$ ($E_g = 0.74$ eV) active layer to absorb long-wavelength photons ($\lambda = 1.3\,\mu m$) and a wider-band-gap InP p-n junction grown on top of an $In_{0.53}Ga_{0.47}As$ absorber layer to achieve avalanche multiplication. Long-wavelength photons impinging on the AR coating layer of a SAM-APD will pass through the top layer of the wider-band-gap InP p-n junction, and absorb in the smaller-band-gap $In_{0.53}Ga_{0.47}As$ active layer. Electron–hole pairs generated in the InGaAs region by the absorbed photons will move into the upper InP p-n junction, where a large reverse-bias voltage is applied to produce avalanche multiplication by impact ionization. It should be noted that doping density and thickness of each layer in the SAM-APD must be calculated so that the electric field at the heterointerface remains sufficiently small to avoid a significant tunneling current, but is large enough to deplete the entire absorber region. Precise control of the device parameters for a planar SAM-APD is very important for achieving optimum performance. The high-performance InGaAs/InP SAM-APD is developed primarily for 10-Gbps fiber optic receiver applications.[16] As shown in Figure 12.28a, the APD employs a widely used SAM layer structure to achieve both high optical absorption and internal gain. The epitaxial layers are grown lattice-matched to the InP substrate. The back-illuminated planar device geometry minimizes junction capacitance and dark current for a given optical coupling diameter. The electron–hole pairs are generated in the InGaAs absorber layer and multiplication takes place in the InP high-field region with an internal gain factor M of around 10



(a) InGaAs/InP SAM APD



(b) 3-dB bandwidth versus multiplication gain M

FIGURE 12.28. (a) Cross-sectional view of InGaAs/Si SAM APD structure; (b) bandwidth as a function of the multiplication gain; the experimental data points are also shown.[16]

or higher. This SAM-APD has a gain bandwidth product of 85 GHz and a peak bandwidth of about 10 GHz at 1.55 μm wavelength, as shown in Figure 12.28b.

The conventional near-IR InGaAs/InP APDs discussed above are limited in performance by a small ionization coefficient ratio, which results in a low-gain bandwidth product and high excess noise. At shorter wavelengths, the silicon APD is used extensively for applications where high sensitivity and high-gain bandwidth product are required. Silicon is an ideal material for APDs because of its very high ionization coefficient ratio, which results in a high-gain bandwidth product and very low excess noise. Unfortunately, silicon has a very low absorption coefficient

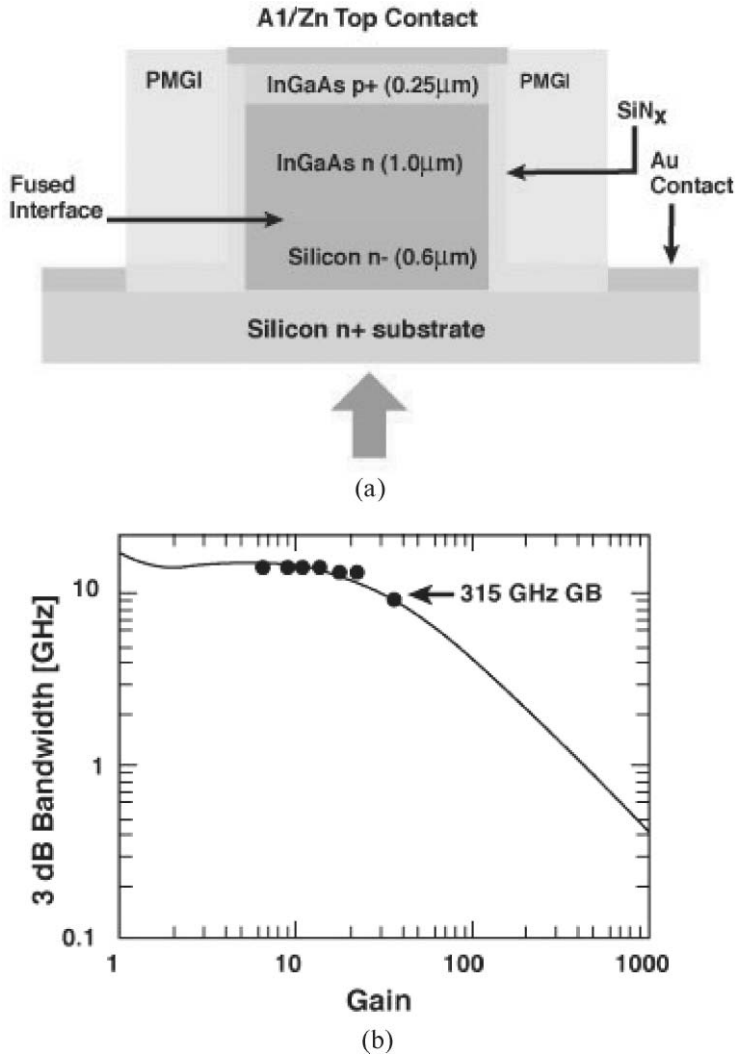


FIGURE 12.29. (a) A passivated planar InGaAs/InP SAM-APD structure grown on n^+-InP substrate and the electric field profile. The APD used the InP p-i-n structure for multiplication and i-InGaAs layer for absorption, (b) the 3-dB bandwidth versus multiplication gain.[17]

especially at the fiber-optic and free space optical communications wavelengths of 1.3 and 1.55 μm. On the other hand, InGaAs is an excellent optical absorber in the wavelengths of 1.3 and 1.55 μm.

A new InGaAs/Si SAM-APD has been demonstrated recently with separate absorption and multiplication regions as shown in Figure 12.29a. The device active area diameter is 23 μm with an InGaAs absorption layer thickness of 1μm and a Si multiplication thickness of 0.6 μm. This SAM-APD was fabricated by wafer fusing

an InGaAs absorption region with a silicon multiplication region, taking advantage of the high IR absorption capability of InGaAs and the high multiplication efficiency of silicon. The enormous advantage of using silicon for the multiplication region is due to its high ionization coefficient ratio, which results in much higher sensitivity, higher-gain bandwidth product, lower noise, and higher temperature and voltage stability than any near-IR APD previously fabricated. The wafer fusion technology employed in the new APD involves the heterogeneous fusion of two dissimilar semiconductor materials (InGaAs and Si). The resulting advantage is the ability to utilize the inherent optical and electrical characteristics of each material to optimize photodetector detectivity at these important wavelengths. Measurements demonstrated a bandwidth of 13 GHz and a gain bandwidth product of 315 GHz as shown in Figure 12.29b. According to calculations, an optimally designed InGaAs/Si SAM-APD could extend the gain-bandwidth product to beyond 500 GHz. This device has substantially higher optical sensitivity, higher speed, lower noise, and higher temperature and voltage stability than currently available APDs operating in this wavelength range. This detector is particularly effective at eye-safe wavelengths.

In addition to the APDs depicted above for the visible to near-IR detection, GaP APDs have also been reported recently for UV photon detection. Although APDs are typically made from silicon, GaP offers three advantages over silicon: GaP is a widely available large-band-gap ($E_g = 2.26$ eV) material, and it has a low intrinsic carrier concentration ($n_i = 1$ cm^{-3} at 300 K) with extremely low reverse leakage current and a band gap amenable to solar blind UV detection. A GaP APD using a guard-ring p-i-n structure has recently achieved a current gain of $M = 1000$ at a reverse bias of around 21 V. The device has applications in a variety of areas, from protein tagging to CD data storage. Finally, APDs using GaN-based wide-band-gap material systems have also been developed for UV detection and chemical sensor applications. Commercial applications of GaN-based UV detectors include environmental monitoring, automobile engine combustion sensing, solar UV monitoring, burner monitoring in gas turbines, and flame detection.

12.3.6. Schottky Barrier Photodiodes

Schottky barrier photodiodes (SBDs) are particularly attractive for high-speed detection. The basic detection principles for an SBD were described in Chapter 10. Depending on the modes of detection, the SBD may be used to detect photons or optical signals with wavelengths extending from UV to IR spectral ranges. When an SBD is operating in the depletion mode under reverse-bias conditions, electron–hole pairs are generated by incident photons with energy greater than the band gap energy of the semiconductor (i.e., $h\nu \geq E_g$). In this case, the cutoff wavelength of the SBDs is determined by the band gap energy of the semiconductor ($\lambda_c = 1.24$ eV/E_g). Depletion-mode Sods fabricated from larger-band-gap semiconductors such as SiC, GaN, GaP, and ZnS are used primarily to detect shorter-wavelength photons (e.g., from UV to visible), while Sods and p-i-n photodiodes fabricated

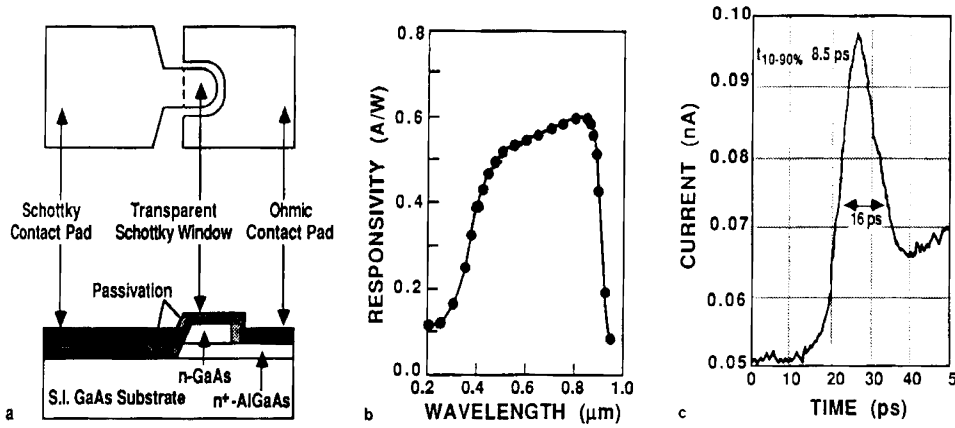


FIGURE 12.30. (a) Top and cross-sectional views of a Au/GaAs/AlGaAs heterostructure planar Schottky barrier photodiode, (b) external quantum efficiency and responsivity curves, and (c) impulse response of the photodiode showing a rise time of 8.5 ps and an FWHM of 16 ps, which corresponds to a cutoff frequency of 45 GHz. After Lee et al.[18]

from smaller-band-gap semiconductors such as Ge, InGaAs, InAs, and CdHgTe are used mainly for mid- to long-wavelength infrared (MWIR to LWIR) detection. The SBD may be fabricated by depositing different metals on various semiconductors to cover the wavelengths from UV to IR spectral ranges. For example, an Ag/ZnS SBD, which has a peak photoresponse at 0.3 μm, is mainly used for UV light detection. An Au/GaAs SBD, which has a peak response at 0.85 μm, is used for visible to near-IR detection. IR detectors using Schottky barrier structures such as a Au/p-$In_{0.53}Ga_{0.47}As$/p^+-InP SBD for 1.3–1.5 μm and PtSi on p-Si Schottky photodiode operating at 77 K have been developed for 3–5 μm IR imaging array applications. Figure 12.30a shows the top and cross-sectional views of a high-speed Au/n-GaAs/n^+-GaAlAs planar SBD. The detector is capable of detecting modulating optical signals up to 45 GHz at $\lambda = 0.8$ μm. Figure 12.30b displays the spectral response of such an SBD, and Figure 12.30c is the impulse response for this detector. A rise time of 8.5 ps and an FWHM of 16 ps have been measured for this detector using the sampling/correlation technique. This corresponds to a 3-dB cutoff frequency of 45 GHz.[18]

SBDs can also be operated in the hot electron detection mode. In this case, photons are absorbed inside the metal film, and photogenerated hot electrons in the metal film are then swept across the Schottky barrier and injected into the semiconductor side. The cutoff wavelength under this detection mode is determined by the barrier height (i.e., $\lambda_c = 1.24$ eV/$q\phi_{Bn}$). Therefore, under this detection mode, an SBD with small barrier height can be used for LWIR detection. A typical example for this type of photodetector is the PtSi/p-type silicon SBD, which has a barrier height of $q\phi_{Bp} = 0.2$ eV and a cutoff wavelength of $\lambda_c = 5.6$ μm. Large-format (1024 × 1024) PtSi/p-Si SBD focal plane arrays have been widely used for 3–5 μm IR image sensor applications.

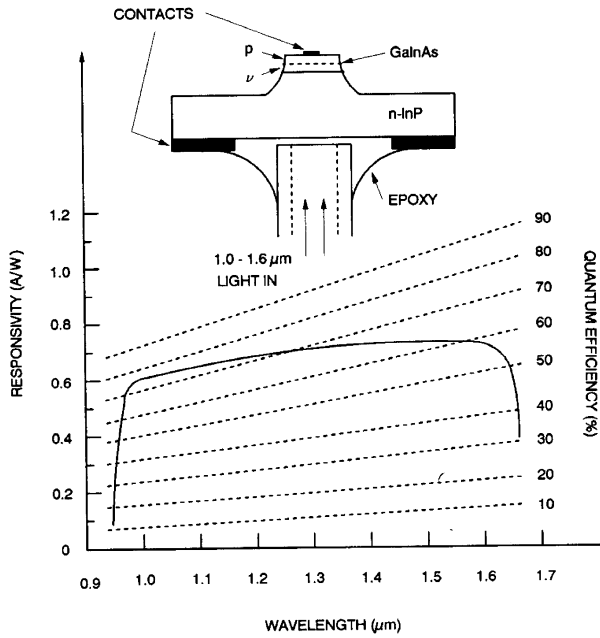


FIGURE 12.31. A p-$In_{0.53}Ga_{0.47}As$–ν-$In_{0.53}Ga_{0.47}As$–n-InP heterojunction p-i-n infrared photodetector for 1.1- to 1.6-μm detection: the cross-sectional view, and quantum efficiency and responsivity curves. After Lee et al.,[19] by permission, © IEEE-1981.

Recently, there has been shown a strong interest in wide-band-gap semiconductor (e.g., GaN, SiC, AlGaN) based gas and liquid sensors for applications including fuel leak detection in spacecraft and automobiles as well as chemical and bioagent sensing for the defense industry and homeland security applications. Wide-bandgap semiconductor devices are capable of operating in harsher environments and at much higher temperatures than the conventional semiconductors such as Si and GaAs. Simple Schottky diode or field-effect transistor structures fabricated on GaN (or SiC) are sensitive to a number of gases, including hydrogen, carbon monoxide, and hydrocarbons. One additional attractive attribute of GaN and SiC is that gas sensors based on these materials can be integrated with high-temperature electronic devices on the same chip. ZnO-based wide-bandgap devices are currently attracting attention for application to UV light emitters, transparent high-power electronics, surface acoustic wave devices, and piezoelectric transducers. Pt/ZnO Schottky diode-based hydrogen sensors with detection limits as small as 5 ppm of H_2 in N_2 have been reported recently.

12.3.7. Point-Contact Photodiodes

A point-contact photodiode may be constructed as a Schottky barrier or as a p-i-n photodiode, depending on the device structures. The active area of a point-contact

photodiode is usually very small, and as a result, both the junction capacitance and transit time are extremely small. The point-contact photodiode is used mainly to detect optical signals at very high modulation frequencies. As an example, consider a Ge point-contact photodiode. The photodiode is formed using a p-type epitaxial layer 8 μm thick grown on a p^+-Ge substrate, and a 0.6-mil arsenic-doped gold foil is alloyed for Schottky contact on the p region using a short current pulse. The alloy region is approximately 4 μm in diameter and depth. Light impinges on the surrounding area of the p region. Using a 4-μm deletion layer width, the carrier transit time for this photodiode is less than 4×10^{-11} s. The junction capacitance and series resistance for such a diode are 4.5×10^{-4} pF and 10 Ω, respectively, which yield an RC time constant equal to 4.5×10^{-15} s. Therefore, the bandwidth (or cutoff frequency) for this photodiode is limited by the transit time, which is on the order of 10^{-11} s. A cartridge-type point-contact photodiode capable of responding to the modulation optical signals with frequencies up to 30 GHz has been reported in the literature. Point-contact photodiodes are particularly attractive for applications in fiberoptic communications in which incident light is confined inside the optic fiber with a light spot a few microns in diameter.

12.3.8. Heterojunction Photodiodes

A heterojunction photodiode is formed with two types of semiconductor materials with different energy bandgaps and opposite dopant impurities. To reduce the dark current and noise of a heterojunction photodiode, the lattice constants for both semiconductors should be chosen as closely matched as possible. There are several semiconductor pairs with good lattice match that can be used to fabricate heterojunction photodiodes. These include InGaP/GaAs, AlGaAs/GaAs, GaAs/Ge, and InGaAs/InP. For example, a heterojunction photodiode made from a wide-bandgap n-type GaAs on a narrow-bandgap p-type Ge can be used to detect IR radiation in the 1.1- to 1.8-μm wavelength regime. The detector is designed in such a way that long-wavelength photons can pass through the top wide-bandgap n-GaAs layer and absorb in the depletion region of the bottom narrow-bandgap p-Ge layer. Carrier generation takes place as a result of absorption of long-wavelength photons in the p-Ge base layer. The optical absorption coefficients for GaAs and Ge are around 10 cm^{-1} and 2.4×10^4 cm^{-1} at 1.6 μm, respectively. This implies that less than 1% of the incident photons are absorbed in the n-GaAs layer, while more than 99% of the incident photons are absorbed within 1 μm from the depletion edge of the p-Ge layer. For this n-GaAs/p-Ge heterojunction photodiode, a narrow-peak spectral response will occur at $h\nu = 1.38$ eV. Since germanium is an indirect bandgap material, the quantum efficiency and responsivity for an n-GaAs/p-Ge heterojunction photodiode are usually low. A superior IR detector using a p-$In_{0.53}Ga_{0.47}As$/n-InP p-i-n heterojunction structure can produce excellent spectral response, high quantum efficiency, and high responsivity for wavelengths between 1.0 and 1.6 μm. Figure 12.31 shows the cross-sectional view and spectral response for such a detector.[12] Responsivity greater than 0.5 A/W and quantum efficiency between 55 and 70% are obtained for this detector in the wavelength

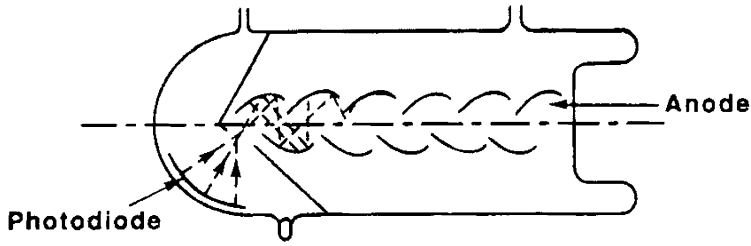


FIGURE 12.32. A typical photomultiplier electrode arrangement showing the linear configuration of the dynode section, with a partitioned transparent photocathode.

range from 1.0 to 1.6 μm. In this photodetector, back-illumination (i.e., incident photons impinging from the wide bandgap InP substrate into the narrow bandgap InGaAs active layer) is used so that most of the long-wavelength photons are absorbed in the undoped ν-$In_{0.53}Ga_{0.47}As$ active layer. Since most of the incident photons are absorbed in the i region, a p-i-n heterojunction photodiode is relatively insensitive to the surface condition. It should be noted that the transit time of the photogenerated carriers across the i region is usually smaller than the RC time constant of the detector. As a result, the frequency response for such a photodiode is usually limited by the RC time constant rather than by the transit time of the detector.

12.3.9. Photomultipliers

The photomultiplier is another type of photodetector and is known as the most sensitive detector available in the visible spectral region. This type of photodetector is particularly useful for photon-counting applications. Figure 12.32 shows a partition-type electron multiplier in which the photocathode is transparent and mounted at the end of the tube. The incident light falls on the front face of the photocathode, and electrons emitted from the cathode surface are multiplied by nine dynode stages. In a conventional P2 phototube, the photocathode is of the S4 class, and the sensitivity is about 40 μA/lm. The overall gain for the nine dynode stages is 2×10^6, corresponding to an average gain of about 5 per stage. Conventional dynode materials for such a photomultiplier include Cs_3Sb, Mg–Ag, and Be–Cu. It is screened from the secondary-emission dynode electrodes by a partition and an aperture that provides convenient separation during activation. Other electron multiplier structures such as Venetian screen, box-type, cross-field, and diode arrangements are used in different photomultiplier applications.

Since the gain per stage is only about 5 for a conventional secondary emitter, the statistical fluctuation in the number of secondary electrons emitted by the first dynode usually limits the tube performance. In order to provide discrimination between signals representing the emission of one or two photoelectrons, it is necessary that the first dynode provide a gain of 15–20. Even higher gains are needed

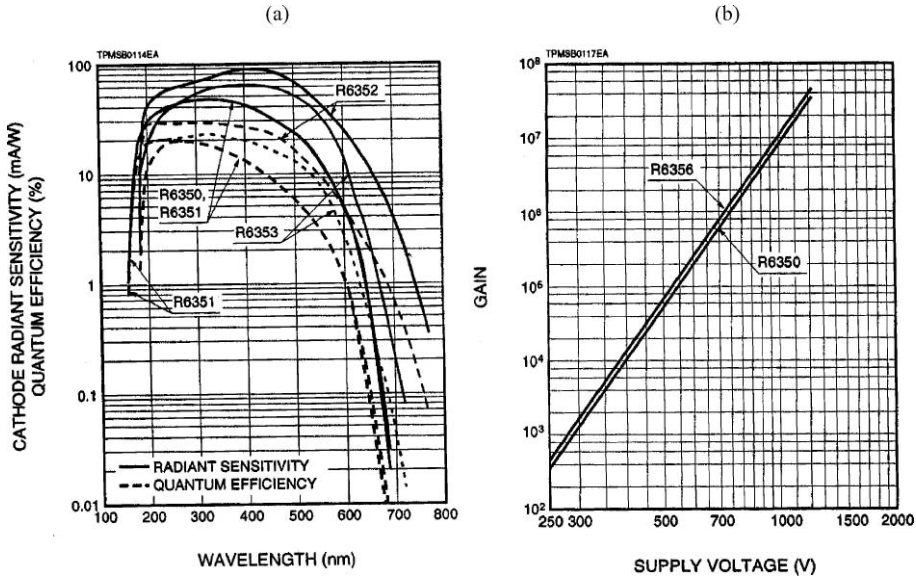


FIGURE 12.33. (a) Spectral responses of cathode radiant sensitivity and quantum efficiency for several new compact-type photomultiplier tubes (PMTs) by Hammamatsu, and (b) internal current gain versus applied voltage for two of these PMTs.

to distinguish between n and $(n + 1)$ photoelectrons, where n is greater than 1. For a cesium-coated GaP photocathode with the first dynode operating at 600 V, gain of 20–40 can be achieved. New compact-type nine-stage photomultiplier tubes (PMTs) developed by Hammamatsu using cathode materials such as Sb–Cs, BA (bialkali), LBA (low dark current bialkali), and MA (multialkali) can cover the wavelengths from 185 nm to around 900 nm and provide a cathode sensitivity up to 100 mA/W and internal current gain of up to 10^7 at $V_a = 1000$ V. Figure 12.33a shows the typical spectral responses for the BA-, LBA-, and Sb–Cs-coated PMTs by Hammamatsu, which have detection wavelengths from the UV (185 nm) to the near-IR (700–900 nm) spectral range. The dashed line is for the quantum efficiency and the solid line denotes the radiant sensitivity of the PMT. Figure 12.33b shows the current gain versus applied voltage for the Sb–Cs-and MA-coated PMTs by Hammamatsu.

PMTs provide high sensitivity and fast response speed for wavelengths from UV to near-IR spectral ranges ($\lambda = 185$ to 830 nm). The PMTs are primarily used in detection and measurements of very-low-light-level scintillations. Some practical applications of the PMTs include emission spectroscopy (e.g., ICP, direct reader), environmental monitoring (NO_x, SiO_2, etc.), fluorescence immunoassay, hygiene monitor (bioluminescence), X-ray phototimer, fluorometer, and laser-scanning microscope. New photocathode materials using III-V compound semiconductors such as $InAs_xP_{1-x}$ and $In_xGa_{1-x}As$ can extend the useful detection wavelengths to the 1–2 μm IR spectral range.

12.3.10. Infrared Photodetectors

Because of the increasing use of IR technologies for various IR image sensor systems and optical communications, a wide variety of IR photodetectors covering the wavelengths in the 1–2 μm, 3–5 μm, 8–14 μm, and even longer wavelengths have been extensively investigated and developed in the last two decades using different device structures and material systems. For example, InGaAs/InP and InGaAsP/InP p-i-n photodiodes are used mainly for detection in the 1.3 and 1.55 μm range, PtSi/Si Schottky barrier photodetectors have been widely used in the 3–5 μm image sensor arrays, and the extrinsic photoconductors using impurity (Cu, In, Hg) doped Ge and (In, Ga) doped Si photoconductors have been used for detection in the 8–40 μm wavelength range. $Hg_xCd_{1-x}Te$ (MCT) is the most widely used IR material for the 3–5 μm and 8–12 μm IR image sensor applications. $Hg_xCd_{1-x}Te$ (MCT) photoconductive and photovoltaic detectors have also been developed for the 3–5 μm MWIR and 8–14 μm LWIR focal plane arrays (FPAs) for IR camera applications. In order to reduce the dark current in the LWIR detectors, these IR detectors are operated at cryogenic temperatures (i.e., at 77 K or 40 K depending on the detection wavelength). In spite of tremendous efforts in the development of various IR detectors, further improvement in the quality of CdHgTe materials and the development of new IR detectors are needed for LWIR FPAs applications. Although the CdHgTe material system can cover a wide range of wavelengths (1–30 μm), the quality of this material needs further improvement; in particular, uniformity of Hg and Cd alloy composition across the entire wafer remains a serious problem. Figure 12.34 shows the energy bandgap versus cutoff wavelength for various impurity-doped IR semiconductor materials.

The development of new and improved long-wavelength (8–12 μm) IR photodetectors for FPA technology is an important step toward meeting the challenges and needs of future IR applications including remote sensing, forward-looking infrared (FLIR) image sensors, highly sensitive staring IR sensor systems, medical imaging, atmospheric optical communication, environmental studies, and space exploration. Extending detection to longer wavelengths offers several advantages. These include the following: (i) IR radiation in the 8–12 μm atmospheric spectral window can travel a longer distance through the atmosphere with small attenuation, (ii) enabling the use of an existing powerful CO_2 laser ($\lambda \approx 10.6$ μm) and maturing technology, (iii) reducing the interference radiation reflected from the background and eliminating susceptibility of false signals triggered by sunlight and other background radiations, and (iv) enabling detection and tracking of cooler targets, such as satellites.

12.3.11. Quantum-Well Infrared Photodetectors

Recent advances in III-V semiconductor epitaxial layer growth using MBE and MOCVD techniques have made it possible to grow a wide variety of novel semiconductor heterostructures. Significant progress has been made in quantum wells and superlattice optoelectronic devices using these growth techniques. The quantum

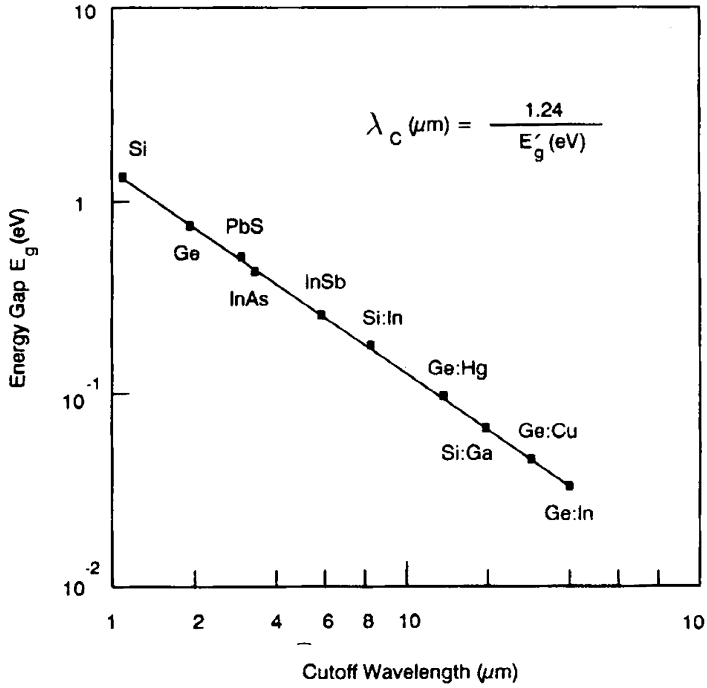


FIGURE 12.34. Energy band gap versus cutoff wavelength for some important impurity-doped semiconductor materials (extrinsic photoconductors).

well is formed using an ultrathin layer of narrow-bandgap semiconductor (e.g., GaAs) sandwiched between two thin wider-band gap semiconductor (e.g., AlGaAs) barrier layers. The thickness of the quantum well is typically smaller than the electron mean free path such that the motion of the carriers perpendicular to the layers becomes quantized so that localized two-dimensional (2-D) subbands of quantized states are formed inside the quantum well.

West and Eglash[20] first reported the observation of intersubband absorption in GaAs/AlGaAs quantum well in 1985. Levine et al.[21] demonstrated the first GaAs/AlGaAs quantum-well infrared photodetector (QWIP) in 1987. Since then, QWIPs based on the bound-to-bound (B-B) state, bound-to-continuum (B-C) state,[12] and bound-to-miniband (B-M)[22] state transitions have been widely investigated for 3–5 μm MWIR and 8–14 μm LWIR FPA applications.[23] Additionally, there is considerable interest in the development of multistack QWIP structures for multicolor FPA applications in the MWIR and LWIR atmospheric spectral windows.[24]

One major difference between the QWIP and other IR detectors is that QWIPs use intersubband transitions either in the conduction band (n-type) or in the valence band (p-type) to detect IR radiation. The existence of a large-dipole matrix element between the subbands of the quantum well makes such a structure extremely

attractive for LWIR detection and modulation, especially in the 8–12 μm atmospheric spectral window. The basic intersubband transition schemes for n-type QWIPs include (a) B-B state, (b) B-C state, (c) B-M state, and (d) bound-to-quasibound (B-QB) state transitions. By using different well widths and barrier heights, the detection wavelengths of QWIPs can be varied from 3 to 20 μm and beyond. Depending on where the upper excited states are located and the barrier layer structure, the intersubband transitions in a QWIP can be based on the B-B, B-M, B-QB, and B-C state transitions. Among the various types of QWIPs reported, GaAs/AlGaAs and InGaAs/AlGaAs QWIP structures using B-C or B-M state transitions are the most widely used material systems and structures for the fabrication of large-format FPAs for LWIR and MWIR imaging applications. Large-format (640 × 480), highly uniform FPAs using GaAs/AlGaAs QWIPs have been developed for 8–12 μm IR imaging camera applications. Multicolor QWIPs using a multistack of InGaAs/AlGaAs (for MWIR) and GaAs/AlGaAs (for LWIR) QWIPs have been developed recently. The multistack QWIP structures are widely used to obtain multicolor detection in the MWIR and LWIR atmospheric spectral windows.

Responsivity and detectivity are two key figures of merit commonly used in comparing the performance of QWIPs. Since the performance of QWIPs depends on the dark current and responsivity, reducing dark current and enhancing responsivity are key to improving the detectivity of a QWIP. Although reducing dark current and increasing responsivity can be achieved by optimizing the quantum-well structure and material parameters, considerable effort has been directed toward the design of efficient light-coupling schemes in n-type QWIPs for FPA applications.

The spectral responsivity of a QWIP is defined by

$$R_{\mathrm{p}} = \frac{q}{h\nu}\eta g p_{\mathrm{e}}, \tag{12.74}$$

where $h\nu$ is the photon energy, η is the absorption quantum efficiency, p_{e} is the escape probability of a hot electron from the quantum-well region, and g is the photoconductive (PC) gain. Note that the value of p_{e} depends on how easy an electron can escape out of the quantum-well region after absorbing photons; its value usually increases with increasing bias voltage.

In a B-C state transition QWIP, p_{e} is large even at very small bias voltage because the excited state is above the conduction band edge of the barrier layer. In general, the B-C QWIP has a relatively large PC gain because of the easy electron transport in the continuum states above the barrier. The responsivity of a B-C QWIP is usually larger than that of the B-QB and B-M QWIPs under the same bias conditions, while the B-M QWIP has the smaller responsivity and lower dark current compared to B-QB and B-C QWIPs owing to the lower electron mobility in the miniband.

For a given QWIP structure, reducing the number of quantum wells can increase the responsivity and PC gain. However, reducing the number of quantum wells will also reduce the absorption quantum efficiency if the light paths and doping

density are fixed. If the absorption is kept constant using effective optical coupling schemes, then the responsivity will increase with decreasing number of quantum wells. Certain light-coupling schemes give effective coupling with a thinner active region. For example, the enhancement QWIP (E-QWIP) reported by Dodd and Claiborn[25] employed a diffractive resonant optical cavity in the place of optical gratings. The cavity requires a relatively thin active layer to resonate at the desired wavelength. This grating structure gives effective light coupling with fewer quantum wells.

The spectral response of a QWIP can be measured using a monochrometer and a blackbody IR source, or an FTIR system. The absolute value of the responsivity can be determined using a calibrated blackbody source. For a single n-type QWIP device, the IR radiation is usually incident through a 45° facet on the edge of the substrate. Either front- or backside illumination can be used depending on the contact metal geometry.

The detectivity D^* is an important figure of merit for IR detectors. Values of D^* can be determined from the measured responsivity and noise of the detector. In general, D^* is a function of operating temperature, detector bias, and cutoff wavelength. Therefore, a typical detector characterization should include measurements of both the responsivity and noise as a function of temperature, bias, and wavelength. The noise current consists of two components: one is due to the device noise current, and the other is due to the background photon noise current. In general, there are two main noise sources in a QWIP device: one is the generation–recombination noise (in PC mode operation) and the other is the Johnson noise (in PV mode operation). Since the majority of QWIPs are operating under PC mode detection, the QWIP device noise is due primarily to the dark-current-related shot noise at high temperatures ($T > 60$ K).

The peak detectivity of a QWIP can be calculated using

$$D_{\mathrm{p}}^* = \frac{R_{\mathrm{i,p}}\sqrt{A_{\mathrm{d}}\Delta f}}{i_{\mathrm{n}}}, \tag{12.75}$$

where $R_{\mathrm{i,p}}$ is the peak current responsivity, Δf is the noise spectral bandwidth, A_{d} is the device area, and i_{n} is the overall root-mean-square noise current of a QWIP.

The background limited performance (BLIP) peak detectivity can be determined using

$$D_{\mathrm{BLIP}}^* = \frac{1}{2}\sqrt{\frac{\eta}{h\nu I_{\mathrm{BG}}}}, \tag{12.76}$$

where η is the absorption quantum efficiency and I_{BG} is the intensity of the incident background photons, which is given by

$$I_{\mathrm{BG}} = \sin^2\left(\frac{\Omega}{2}\right)\cos(\theta)\int_{\lambda_1}^{\lambda_2} W(\lambda)\,\mathrm{d}\lambda, \tag{12.77}$$

where Ω is the solid angle, θ is the angle between the incident IR radiation and the normal to the quantum-well plane, and $W(\lambda)$ is the blackbody spectral density

given by

$$W(\lambda) = \frac{2\pi c^2}{\lambda^5} \frac{1}{(e^{hc/\lambda k_B T_{BG}} - 1)}, \tag{12.78}$$

where c is the speed of light, λ is the wavelength, h is Planck's constant, k_B is Boltzmann's constant, and T_{BG} is the background temperature.

The D^* value also depends on the detector structure. For example, a B-M QWIP could have a higher D^* at high temperatures when the miniband width is very narrow and resonant with the excited state. The miniband structure could lower the dark current more effectively than the photocurrent in this situation, and hence give rise to a higher signal-to-noise ratio. At low operating temperatures ($T < 50$ K at 10 μm cutoff), the dark current is low, and D^* varies linearly with the spectral responsivity. The B-C QWIP usually has a higher value of D^* at lower temperatures and smaller bias voltages owing to its larger responsivity compared with other types of QWIPs.

Figure 12.35a shows the energy band diagram of a standard GaAs/AlGaAs QWIP, which uses the B-C state transition to achieve charge transport and IR detection. A typical device structure for a standard GaAs/AlGaAs QWIP consists of 50 periods of GaAs (width of 3–5 nm) quantum wells doped to 1×10^{18} cm^{-3} and an undoped $Al_{0.3}Ge_{0.7}As$ (thickness of 40–50 nm) barrier layer; the heavily doped (2×10^{18} cm^{-3}) GaAs buffer layer and GaAs cap layer are deposited on the bottom and top of the active quantum-well layers for ohmic contacts. As shown in Figure 12.35a, in the B-C QWIP structure only one bound ground state (E_1) exists in the quantum well that is filled with electrons, and the next empty band is the continuum band states located slightly above the conduction band edge of the AlGaAs barrier layer. The conduction band offset (equal to the potential barrier height) for the GaAs/AlGaAs quantum well is about 190 meV, and the energy separation between the bound state, E_1, and the continuum states, E_c, is about 120 meV, which corresponds to a peak detection wavelength of around 10 μm ($\lambda_p = 1.24$ eV/0.120 eV ≈ 10 μm). A standard QWIP device operating in PC detection mode uses the B-C states' intersubband transitions. Since the dark current in this QWIP is controlled by the thermionic emission across the barrier from the ground states in the quantum well to the continuum states, the detector is required to cool down to 77 K or lower in order to reduce the dark current. Excellent responsivity and detectivity have been obtained in GaAs/AlGaAs QWIPs. Detectivity greater than 10^{10} (cm $\cdot$ Hz$^{1/2}$)W has been achieved for the GaAs/AlGaAs B-C QWIP at $\lambda_p = 10$ μm and $T = 77$ K.

Another type of QWIP aimed at reducing the dark current of the standard B-C QWIP has been reported by Yu and Li.[22] Figure 12.35b shows a B-M transition GaAs/AlGaAs QWIP. In the B-M QWIP structure the AlGaAs bulk barrier used in the standard GaAs/AlGaAs QWIP is replaced by a short-period (5 periods) AlGaAs (5.8 nm)/GaAs (2.9 nm) superlattice barrier layer, and the width of the GaAs quantum well is increased to 8.8 nm. The B-M QWIP differs from the standard B-C QWIP in that (a) the potential barrier is increased to 300 meV, and (b) a global miniband superimposed with the first excited state (E_2) in the quantum

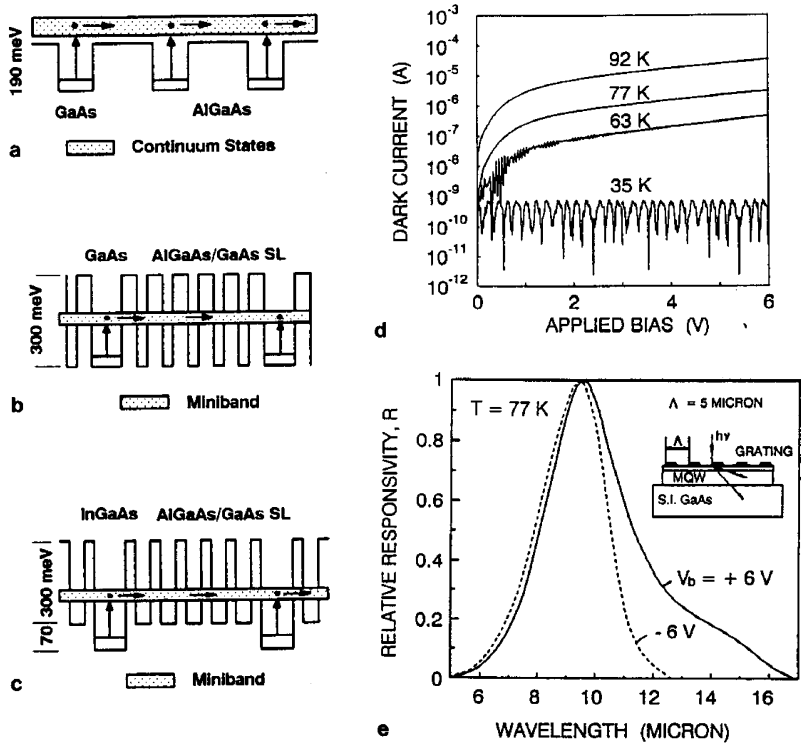


FIGURE 12.35. Energy band diagrams for (a) a standard GaAs/AlGaAs (40/480 Å) quantum-well infrared photodetector (QWIP) using bound-to-continuum (B-C) state transition, (b) a GaAs (88Å QW)/AlGaAs-GaAs (58/29Å superlattice) QWIP using bound-to-miniband (B-M) transition, (c) an InGaAs (106 Å QW)/AlGaAs-GaAs (58/29 Å SL) QWIP using step-bound-to-miniband (SBM) transition, (d) dark current versus bias voltage for QWIP shown in (c) with temperature as a parameter, and (e) relative responsivity versus wavelength for QWIP shown in (c), measured at $V_B = 6$ V and $T = 77$ K.

well of the superlattice barrier layer is formed inside the quantum well to facilitate intersubband IR detection. The energy separation between the ground state and the miniband determines the spectral bandwidth of the IR detection (typically $E_m - E_1 = 90$ to 120 meV). The current conduction mechanism in this miniband transport is based on the thermionic-assisted resonant tunneling from the ground state to the global miniband states as shown in Figure 12.35b. As a result, the dark current in such a B-M QWIP is expected to be lower than a standard B-C QWIP. Excellent detectivity and responsivity have been obtained for the GaAs/AlGaAs B-M QWIP. To further reduce the dark current in a B-M QWIP, an enlarged InGaAs (with less than 10% of In) quantum well (width of 10.6 nm) is introduced to replace the GaAs quantum well, and the resulting structure is shown in Figure 12.35c. This modified B-M QWIP is also referred to as the step-bound-to-miniband (SB-M) QWIP. The dark current in an SB-M QWIP is generally lower than a

B-M QWIP and a standard B-C QWIP. Figure 12.35d shows the dark current versus bias voltage for an SB-M QWIP with temperature as a parameter. The results show that thermionic-assisted tunneling current dominates at 77 K, while resonant tunneling current prevails for temperatures below 50 K. A typical spectral response curve for an SB-M QWIP is displayed in Figure 12.35e, which shows a peak response around 10 μm. It is interesting to note that the spectral response peak wavelength for a B-M QWIP is usually voltage-tunable, as is evidenced in Figure 12.35e.

Another advantage of QWIPs is the flexibility and ease of fabricating multicolor IR detectors using MBE growth of multiquantum well layer structures on GaAs substrates. Figure 12.36a shows a schematic conduction band diagram of a two-stack, two-color QWIP. An MWIR QWIP stack consisting of 20 periods of InGaAs/AlGaAs QWIP structure with peak wavelength at 4.3 μm was first grown on the GaAs substrate, followed by the deposition of a thin highly doped ohmic contacting layer, and the LWIR QWIP stack consisting of 20 periods of GaAs/AlGaAs QWIP structure with peak wavelength at 9.5 μm was then grown on top of this intermediate contact layer, and finally, an n^+ GaAs cap layer was grown on top of the LWIR QWIP stack. Figure 12.36b shows the dark *I–V* characteristics for the MWIR and LWIR QWIP stacks, and Figure 12.36c shows the spectral responsivity of the MWIR and LWIR QWIP stacks, respectively. Excellent performance has been achieved for these two-stack, two-color MWIR and LWIR QWIPs.[26]

Although detectivity and responsivity for the QWIPs discussed above are generally lower than for the HgCdTe IR detectors, the GaAs/AlGaAs QWIP device has the advantages of low noise, high uniformity, and extremely high number of operable pixels (>99%), which results in excellent imaging performance with a noise-equivalent temperature difference ($NE\Delta T$) of 10 mK. In fact, low-noise large-format (640 × 480) GaAs/AlGaAs QWIP FPA for staring IR (10 μm) sensor systems has been developed for IR imaging camera applications in recent years. QWIP devices can be fabricated using the mature GaAs growth and processing technology, which could produce highly uniform and 99.99% operable pixels for IR FPA applications. In addition, QWIP technology also offers the benefits of wavelength selectivity, multiple-band sensitivity, compatibility for hybridization with silicon and GaAs IC read-out electronics, and the possibility of full optical and electronic monolithic integration. QWIP FPAs are comparable in complexity to existing GaAs devices and are expected to be producible and low-cost. One drawback of QWIP technology is related to the fact that owing to the quantum-mechanical selection rule, n-type QWIP requires metal or dielectric grating structures to couple the normal-incidence IR radiation into the quantum wells for normal-incident absorption. The grating structures will make the fabrication of QWIP FPAs more complicated than the conventional p-n junction or Schottky barrier photodetectors. Figure 12.37 shows the spectral detectivity (D^*) versus wavelength for both n- and p-type QWIPs published in the literature.[27] The best fit to the experimental data for the spectral detectivities of n-and p-type QWIPs with a 45° polished incident

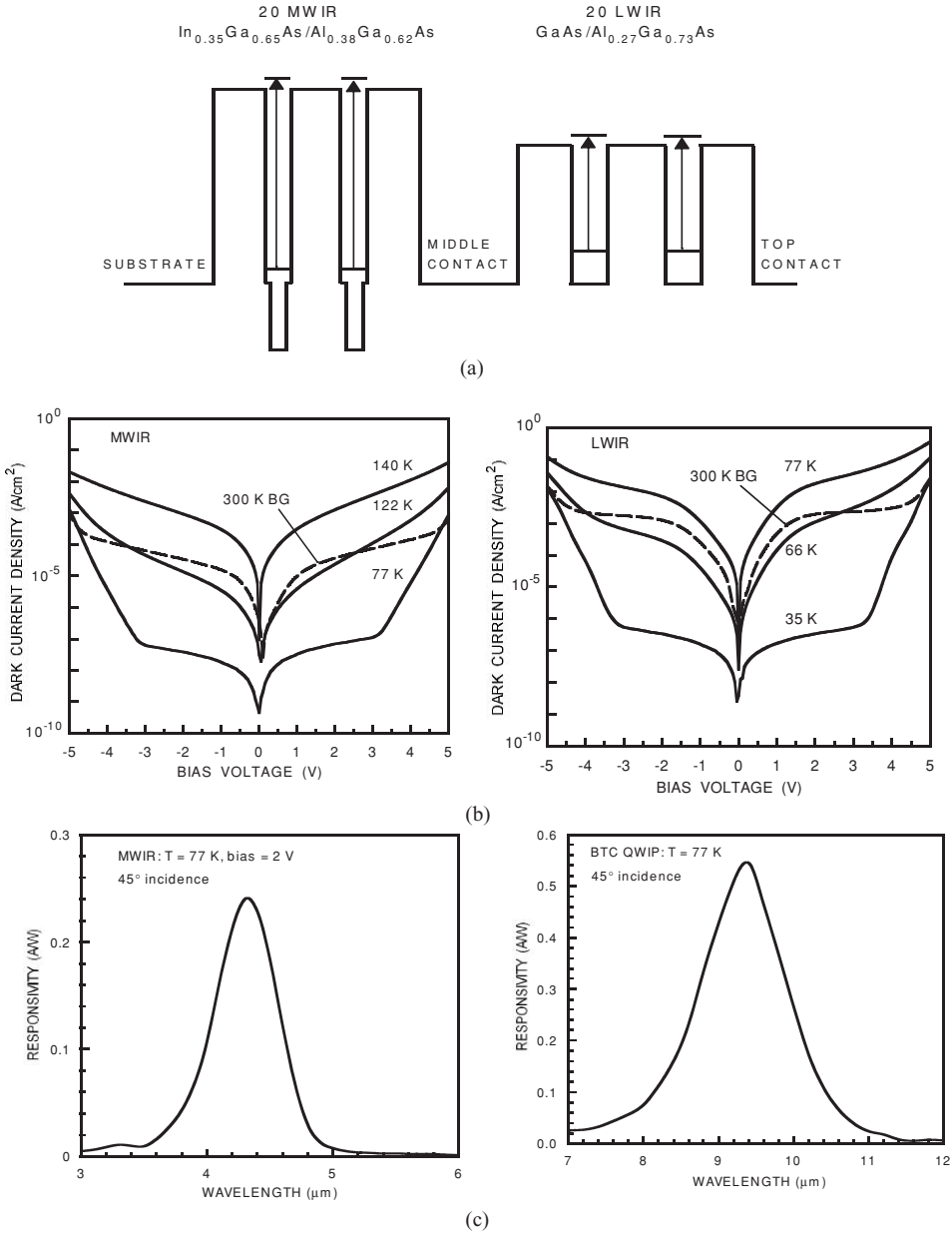


FIGURE 12.36. (a) Schematic conduction band diagram of a two-stack, two-color InGaAs/AlGaAs and GaAs/AlGaAs QWIP for MWIR and LWIR detection, (b) dark *I–V* characteristics for the MWIR and LWIR QWIP, and (c) the responsivity versus wavelength for the MWIR and LWIR QWIP stacks.[26]

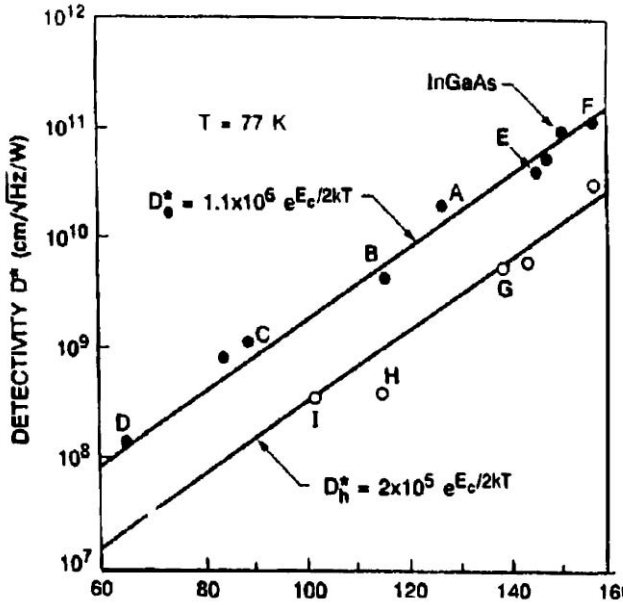


FIGURE 12.37. Peak detectivity (D^*) versus cutoff energy E_c ($= 1.24/\lambda_c$) for n-type QWIPs (solid circles) and p-type QWIPs (open circles). Straight lines are the best fits to the measured data.[27]

facet are given by

$$D_n^* = 1.1 \times 10^6 \exp(E_c/2k_BT)\,\text{cm} \cdot \text{Hz}^{1/2}/\text{W} \quad \text{for n-type QWIPs,} \quad (12.79)$$
$$D_p^* = 2.0 \times 10^5 \exp(E_c/2k_BT)\,\text{cm} \cdot \text{Hz}^{1/2}/\text{W} \quad \text{for p-type QWIPs,}$$

where $E_c = 1.24/\lambda_c$ is the cutoff energy in eV, and λ_c is the cutoff wavelength in μm. Values of D^* for n-type QWIPs are found to be one to two orders of magnitude higher than those of p-type QWIPs because of the higher electron mobilities and the use of grating structures to enhance responsivity and absorption quantum efficiency in n-type QWIPs. As a result, all QWIP FPAs used in IR cameras are made from n-type QWIPs. Table 12.3 lists some of the photodetectors for UV, visible, near-IR, MWIR, and LWIR applications.

12.3.12. Quantum-Dot Infrared Photodetectors

Quantum-dot infrared photodetectors (QDIPs) have been widely studied in the past decade when the self-assembled growth technique was applied to form quantum dots in III-V semiconductors. By controlling the barrier characteristics, dot size, and doping level in semiconductor quantum-dot systems, QDIPs operating in the 3.5–14 μm spectral range have been reported in single-stack QDIPs.[28,29]. The QDIPs have emerged as promising devices for 3–5-μm

TABLE 12.3. Photodetectors for UV, visible, near-IR, MWIR, LWIR, and VLWIR detection.

Detector type and materials	Spectral range	λ_{max}	Responsivity λ_{max}(A/W)	Remarks: D^* (cm $\cdot$ Hz$^{1/2}$W^{-1}), spectral range
SiC	210–380 nm	275 nm	0.13	Solar blind, UV
$Al_xGa_{1-x}N$	250–325 nm	75–350 nm	0.04–0.10	Solar blind, UV
GaN	250–365 nm	365 nm	0.10	UV
GaP	250–400 nm	365 nm	0.075	UV
Si-PIN-PD	250–1100 nm	250 nm, 900 nm	0.12, 0.5	UV, visible
Si-APD	400–1100 nm	830–900 nm	45–128	Visible to near-IR
Ge-APD	800–1500 nm	1300 nm	0.63–0.84	SWIR at $M = 1$, $V_B = 60-70$ V
InGaAs PIN-PD	900–1700 nm	1600 nm	0.85–0.95	SWIR D$^* = 1.6 \times 10^{12}(-20°C)$
InGaAs APD	900–1700 nm	1300/1550 nm	8.4/9.4	SWIR, at $M = 1$, $V_B = 30$ V.
InP/InGaAs APD	1200–1630 nm	1300/1550 nm	0.6,4.5	SWIR, at $M = 1, 10$, $V_B = 30$ V
PbS	1.0–3.5 μm	2.4 μm	1–8$\times 10^5$ V/W	SW-MWIR $D^* = 1 \times 10^{11}(23°C)$
PbSe	1–5 μm	4.4 μm	–	MWIR $D^* = 1.9 \times 10^9(-20°C)$
InAs	1.0–3.8 μm	3.5 μm	1.5	MWIR $D^* = 3 \times 10^{10}$, $(-20°C)$
InSb	1.0–5.5 μm	5.0 μm	3.0	MWIR $D^* = 1 \times 10^{11}$
PtSi/p-Si SBD	3–5 μm	5.3 μm	0.5	MWIR
$Hg_xCd_{1-x}Te^*$	2.0–5.5 μm	4.0 μm	2×10^5 V/W	MWIR $D^* = 4 \times 10^{10}$, $-40°C$
$Hg_xCd_{1-x}Te^*$	2–12 μm	11 μm	0.4–1 $\times 10^5$ V/W	MW-LWIR $D^* = 5 \times 10^{10}$, 77 K
$Hg_xCd_{1-x}Te^*$	2–22 μm	16 μm	30–800 V/W	MW-VLWIR $D^* = 0.6$ to 1×10^{10}, 77 K
GaAs-based QWIPs	3–30 μm	—	—	MW-VLWIR $D^* = 10^9$ to 10^{11}

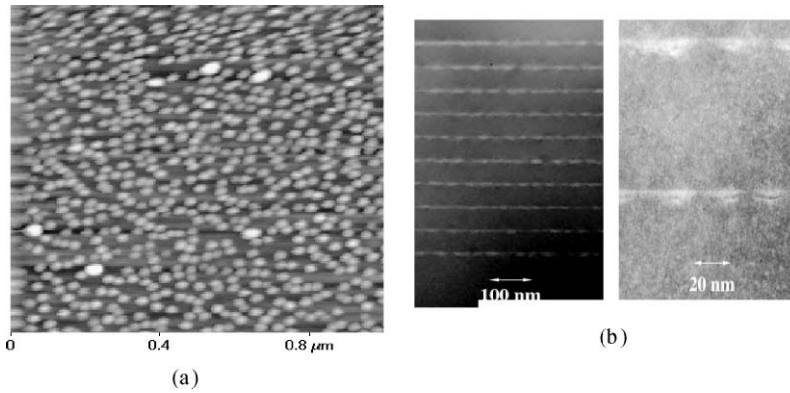
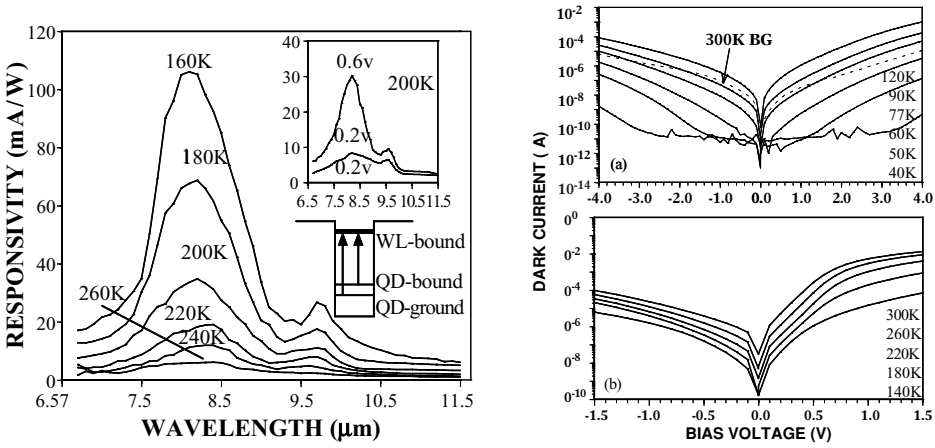


FIGURE 12.38. (a) The AFM photo of $In_{0.6}Ga_{0.4}As$ QDs grown on the GaAs substrates, and (b) cross-sectional TEM micrographs of an InGaAs/GaAs QDIP.

MWIR and 8–12-μm LWIR detection because of their potential advantages over conventional QWIPs. The advantages of QDIPs include (1) intrinsic sensitivity to normal incident IR light, (2) longer carrier lifetime due to greatly suppressed electron–phonon scattering, and (3) potential for very low dark current. These unique properties arise from 3-D carrier confinement of quantum dots (QDs). Although several earlier studies showed that the performance of QDIPs was still inferior to that of QWIPs, recent studies have shown that InAs/GaAs QDIPs could achieve higher operating temperature (> 100 K) by using a large-band-gap material such as AlGaAs or InGaP as the current-blocking barrier to reduce the device dark current.[28,29] Figure 12.38 shows a highly sensitive $In_{0.6}Ga_{0.4}As$/GaAs QDIP operating in the 6.7–11.5 μm spectral range, and photoresponse up to 260 K is demonstrated. The BLIP detectivity at $V_b = -2$ V, $T = 77$ K, and $\lambda_p = 7.6$ μm was found to be 1.1×10^{10} $(\text{cm} \cdot \text{Hz}^{1/2})/\text{W}$, with a corresponding responsivity of 0.25 A/W.[28]

The QDIP sample with self-assembled $In_{0.6}Ga_{0.4}As$ QDs was grown on a semi-insulating GaAs substrate using the Stranski–Krastanov (S-K) growth mode by the solid-source MBE technique. Before the growth of QDIP structure, a 0.5-μm GaAs buffer layer was grown on the GaAs substrate. The active region consists of 10 periods of $In_{0.6}Ga_{0.4}As$/GaAs QDs, and a 60-nm GaAs spacer is added to each QD cell. The QD active layers were sandwiched by a 500-nm n-type GaAs top contact layer and a 1-μm bottom contact layer. These contact layers were doped with Si to $2.0 \times 10^{18}\text{cm}^{-3}$. The nominal thickness for the $In_{0.6}Ga_{0.4}As$ QDs is 5 ML, and the QDs are Si-doped to $8.0 \times 10^{17}\text{cm}^{-3}$. The $In_{0.6}Ga_{0.4}As$ QD growth rate is 0.5 ML/s; the GaAs spacer is controlled at a growth rate of 1 μm/h. The growth temperature is 580°C for the GaAs buffer and contact layers, and 520°C for the $In_{0.6}Ga_{0.4}As$ QDs and GaAs spacers. Figure 12.38a shows the atomic force microscopy (AFM) of $In_{0.6}Ga_{0.4}As$ QDs grown on GaAs substrates. The AFM images reveal that the average QD density is 1.2×10^{10} cm^{-2}, and the average



(a) (b)

FIGURE 12.39. (a) Responsivity versus wavelength for the InGaAs/GaAs QDIP for 160 < T K < 260 K, and (b) dark current versus bias voltage for 40 K < T < 300 K.[28]

size of the QDs is 26 nm in diameter and 6 nm in height. Figure 12.38b shows the cross-sectional transmission electron microscopy (TEM) of this QDIP structure.

Figure 12.39a shows the spectral responsivities for 160 K < T < 260 K. As shown in this figure the maximum responsivity obtained at 160 K and $V_b = -0.8$ V was 0.11A/W at $\lambda_p = 8.1$ μm. The peak responsivity was 6.1 mA/W at $T = 260$ K, $V_b = -0.55$ V, and $\lambda_p = 8.4$ μm. Values of FWHM ($\Delta\lambda/\lambda_p$) vary from 14.3 to 14.8%. The narrow absorption bandwidth of this QDIP reveals that the optical transition is due to the B-B state intersubband transition (i.e., the absorption is due to the QD first and second bound states to wetting-layer bound state transitions). Since the size and shape of the dots that affect the electronic levels cannot be accurately determined, and the strain tensor of the dots is complicated, it is difficult to obtain a true band structure for this QDIP. Based on the calculated values for a typical InGaAs/GaAs QDs (i.e., 15–28 nm in base width, 3–7 nm in QD height), the energy difference from the ground bound state of QDs to the wetting-layer bound state was estimated in the 100–200 meV range, and the energy separation between the electron states in QDs was in the 30–80 meV range. The photon absorption peaks (127–163 meV) and the energy separation of the electron states in QDs ($\approx$ 36 meV) are in reasonably good agreement with the estimation. Another key factor for high-temperature operation of this QDIP is attributed to the intrinsic property of the 0-D device: the large electron relaxation time from the excited states to the ground state of the QDs, which makes photoexcited electrons difficult to recapture by QDs, and hence can increase the signal-to-noise ratio and allow for higher operating temperature. Figure 12.39b shows the dark I–V curves along with the 300 K window current with a 180° FOV measured at different temperatures. In the top figure, the dark current was measured at $T = 40, 50, 60, 77, 90, 120$ K with

bias voltages varying from -4.0 to $+4.0$ V. The BLIP conditions for this QDIP are obtained for $-2.2\ \text{V} \leq V_\text{b} \leq 0\ \text{V}$ at 90 K, and for $0\ \text{V} \leq V_\text{b} \leq 1.64\ \text{V}$ at 77 K. The bottom figure shows the dark current density measured at $T = 140, 180, 220$, 260, and 300 K for $1.5\ \text{V} \leq V_\text{b} \leq 1.5\ \text{V}$. As shown in this figure, at $T = 260$ K and $V_\text{b} = -0.5$ V, the dark current density is $3.75 \times 10^{-3}\ \text{A/cm}^2$, which is comparable to the dark current density of an LWIR InGaAs/GaAs QWIP operating at 90 K ($J_\text{D} = 10^{-3}\text{A/cm}^2$, $V_\text{b} = -1\ \text{V}$).

Problems

12.1. Using (12.59) and (12.60) calculate the quantum yield versus photon wavelength for a Si p-i-n photodiode for $x_0 = 0.4, 0.8$, and 1.2 μm, and $W = 0.01$ cm. Assume $R = 0.3$, $L_\text{n} = 6 \times 10^{-2}$ cm, and $L_\text{p} = 4 \times 10^{-3}$ cm.

12.2. (a) Describe some key factors that need to be considered in the design of a photodetector.

(b) Which of the following detectors would you choose (and explain why) for detection in the specific wavelength shown below: (i) a Si p-i-n photodiode, (ii) a Au/n-type Si Schottky barrier photodiode, and (iii) an n-type GaAs/AlGaAs QWIP.

(i) For 10.6 μm detection.

(ii) Maximum sensitivity needed for detection of 1.06 μm wavelength.

(iii) Low-noise, high-speed detection in the visible spectral range.

12.3. Design a silicon p-n junction solar cell using your own device and material parameters that could produce a conversion efficiency of 19% under AM1.5G conditions (i.e., $P_\text{in} = 100\ \text{mW/cm}^2$). What are the short-circuit current density, open-circuit voltage, and fill factor (FF) for such a solar cell. What is the key difference in the performance characteristics between a p-n and an n-p junction solar cell?

12.4. The conversion efficiency of a Schottky barrier solar cell can be expressed by

$$\eta_c = V_\text{mp}^2 I_0 (q/k_\text{B}T)\ \exp(qV_\text{mp}/k_\text{B}T)(P_\text{in}A_\text{j})^{-1}, \tag{1}$$

where V_mp is the voltage at maximum power output, $P_\text{in} = 100\ \text{mW/cm}^2$ for AM1.5G sunlight, A_j is the cell area,

$$I_0 = A_\text{j} A^{**} T^2\ \exp(-q\phi_\text{Bn}/k_\text{B}T), \tag{2}$$

where I_0 is the reverse saturation current; V_mp is related to I_0 and I_ph via the following relation:

$$\left(1 + \frac{qV_\text{mp}}{k_\text{B}T}\right) \exp(qV_\text{mp}/k_\text{B}T) = \left(1 + \frac{I_\text{ph}}{I_0}\right). \tag{3}$$

Equation (3) can be solved iteratively for V_mp. By substituting (2) into (1), the conversion efficiency can be calculated as a function of ϕ_Bn.

Calculate the conversion efficiency η_c for an Al-n-Si Schottky barrier solar cell. Given: $\phi_{Bn} = 0.71$ eV, $A^{**} = 110A/(\text{cm}^2 \cdot \text{K}^2)$, $A_j = 4$ cm^2, and $I_{ph} = 140$ mA. Repeat for a Au-n-type Si Schottky barrier solar cell with $\phi_{Bn} = 0.81$ eV.

12.5. Draw the energy band diagram for an AlGaAs/GaAs p-n junction solar cell shown in Figure 12.7. If x in the $Al_xGa_{1-x}As$ window layer (top layer) varies from $x = 0, 0.3, 0.5, 0.7$, to 0.9, plot the relative spectral responses for these solar cells. Given: $E_g = 1.43$ eV for $x = 0$ and $E_g = 2.1$ eV for $x = 0.9$; assume that a linear relation exists between E_g and x.

12.6. Consider an InGaP/GaAs/Ge triple-junction solar cell. If the short-circuit current densities produced in each subcell are given by $J_{sc1} = 16.3$ mA/cm^2 for the InGaP top cell, $J_{sc2} = 17.5$ mA/cm^2 for the GaAs middle cell, and $J_{sc3} = 16.3$ mA/cm^2 for the Ge bottom cell, the open-circuit voltage (V_{oc}) for this 3-junction cell is equal to 2.56 V and the FF is 84.6%. (a) Calculate the conversion efficiency of this cell under AM0 conditions ($P_{in} = 135.3$ mW/cm^2 at 28°C), (b) draw the relative spectral response curves for the top, middle, and bottom cells, and (c) plot the photo-I–V curve using the data given above for this 3-junction cell.

12.7. Using the equations for dark currents in an ideal Schottky barrier diode and a p-n junction diode given in the text calculate the dark current for an Al-n-Si Schottky barrier solar cell and a Si p-n junction solar cell. Given: $\phi_{Bn} = 0.71$ V, $N_D = 10^{16}$ cm^{-3}, $N_A = 5 \times 10^{18}$ cm^{-3}, $L_n = 100\,\mu$m, $L_p = 20\,\mu$m, $n_i = 1.4 \times 10^{10}$ cm^{-3}, $A_j = 4$ cm^2, $\mu_n = 1000$ cm^2(V·s), and $\mu_p = 100$ cm^2/(V·s). If the photocurrents generated in both cells are assumed the same ($I_{ph} = 140$ mA), what are the open-circuit voltages for both cells?

12.8. (a) Show that the short-circuit current $J_p(\lambda)$ generated in the n-base region of a metal-n-type semiconductor Schottky barrier solar cell is given by (12.34), assuming that $d \gg L_p$ and $s_n = \infty$ at $x = d$.

(b) Using (12.32) to (12.36) calculate the quantum efficiency ($= I_{ph}/q\phi_0$) versus wavelength for a Au-n-type Si Schottky barrier solar cell for different depletion layer thicknesses: $W = 0.01$, 0.05, and 0.1 cm, assuming $L_p = 0.05$ cm and $R = 0.3$.

12.9. Design a Au-GaAs Schottky barrier photodiode for detecting a 20-GHz modulation optical signal at 0.84 μm (select your own design parameters: diode area, dopant density, AR coating, etc.). If the bandwidth of the detector is increased to 100 GHz for the Au-GaAs Schottky diode, what device parameters need to be modified in this photodiode in order to meet this specification?

12.10. An $In_{0.53}Ga_{0.47}As$ p-i-n photodiode is used to detect 1.3-μm IR radiation. If the dopant density is 1×10^{16} cm^{-3} in the n region and 2×10^{18} cm^{-3} in the p^+ region, the diode area is 50 μm^2, and the n-layer is 1.5 μm thick, calculate R_s, C_j, and the RC time constant of this photodiode. What is the maximum cutoff frequency for this photodetector?

12.11. Draw the schematic energy band diagrams for n-type GaAs/AlGaAs QWIPs using (a) bound-to-bound (B-B) state, (b) bound-to-miniband (B-M) state, (c) bound-to-continuum (B-C) state, and (d) bound-to-quasibound (B-QB) state intersubband transitions. Compare the differences in the spectral response bandwidth, the dark current, and the responsivity of QWIPs based on these four different transition schemes.

References

1. M. P. Thekaekara, *Suppl. Proc. 20th Annual Meeting Inst. Environ Sci.,* (1974). p. 21
2. J. J. Wysocki and P. Rappaport, *J. Appl. Phys.* **31**, 571 (1961).
3. J. Lindmayer and J. F. Allison, *Conf. Record, 9th IEEE Photovoltaic Spec. Conf.*, New York, p. 83 (1972).
4. S. S. Li, *Solid-State Electron*. **21**, 435 (1978).
5. S. P. Tobin, S. M. Vernon, C. Bajgar, V. E. Haven, and L. M. Geoffroy, *IEEE Electron. Device. Lett.* **9**, 256 (1988).
6. Zhu H., Kalkan A. K., Hou J., Fonash S. J. AIP Conf. Proceedings, 462, 309–314. (1999).
7. J. Song, S. S. Li, C. H. Huang, O. D. Crisalle, and T. J. Anderson, Solid State Electronics, **48**, pp. 73–79 (2004).
8. T. Takamoto, E. Ikeda, H. Kurita, and M. Ohmori, *Appl. Phys. Lett*. **70**, 381 (1997).
9. M. Yamaguchi and K. Araki, Japanese R & D activities of multijunction and concentrator solar cells, in www.dlnet.vt.edu. (2005).
10. X. Deng, *Proc. of 31th IEEE Photovoltaic Specialist Conference*, Jan. (2005).
11. R. A. Sherif, R. R. King, H. L. Cotal, C. Fetzer, K. Edmondson, D. Law, G. Kinsey, H. Yoon, and N. H. Karam, *Proc. of DOE Solar Program Review Meeting*, p. 194, Jan. (2004).
12. M. A. Martin, K. Emery, D. L. King, S. Igari, and W. Warta, *Prog. Photovolt: Res. Appl.* **13,** 387 (2005). (Published online in Wiley Interscience (www.interscience.wiley.com).DOI:10.1002/pip.651.
13. S. M. Sze, *Physics of Semiconductor Devices*, 2nd edition, p. 748 Wiley, New York, (1981).
14. S. S. Li and F. A. Lindholm, *Phys. Status Solidi A* **15**, 237 (1973).
15. H. Melchior and W. T. Lynch, *IEEE Trans. Electron. Dev.* **ED-13**, 829 (1966).
16. K. K. Loi and M. Itzler, Compound Semiconductor, **6**(3), p. 44 (2000).
17. K. Nishida, K. Taguchi, and Y. Matsumoto, *Appl. Phys. Lett.* **53**, 251 (1979).
18. D. H. Lee, S. S. Li, and N. Paulter, Proc. *Int. Conf. on Solid State Devices and Materials*, Tokyo Japan, (1988).
19. T. P. Lee, C. A. Burrus, and A. G. Dentai, *IEEE J. Quantum Electron*. **17**, 232 (1981).
20. L. C. West and S. J. Eglash, *Appl. Phys. Lett*. **46**, 1156 (1985).
21. B. F. Levine, *J. Appl. Phys*. **74**, R1 (1993).
22. L. S. Yu, and S. S. Li, *Appl. Phys. Lett.* **59**(11), 1332 (1991).
23. S. S. Li, *Int. J. of High-Speed Electronics and Systems*, **12** (3), pp. 761–801 (2002).
24. M. Sandaram, T. Faska, M. Taylor, R. Williams, A. Reisinger, and S. Wang, *Int. Symp. On Advanced Luminescent Materials and Quantum Confinement,* edited by M. Cahay, Electrochemical Society (**ECS**) Proc. vol. **PV-99-22**, pp. 459–465 (1999).
25. M. A. Dodd and L. T. Claiborn, *Proc. of 5th Int. Symp. On Long Wavelength Infrared Photodetectors and Array*, **ECS-97-33**, pp. 22–31 (1997).

26. M. Z. Tidrow, J. C. Chiang, S. S. Li, and K. Bacher, *Appl. Phys. Lett.*, **70** (1991) 859.
27. S. S. Li and M. Z. Tidrow, *Handbook of Nanostructured Materials and Technology*, Chapter 9: Quantum Well Infrared Photodetectors, editor: H. S. Nalwa, **vol. 4**, pp. 561–619 (2000).
28. L. Jiang, S. S. Li, N. T. Yeh, J. I. Chyi, C. E. Ross, and K. S. Jones, *Appl. Phys. Letters.*, **82** (12), pp. 1986–88 (2003).
29. L. Jiang, S. S. Li, N. T. Yeh, J. I. Chyi, and M. Z. Tidrow, *Electronics Letters*, **38**(22), (2002) 1374.

Bibliography

A. A. Bergh and P. J. Dean, *Light-Emitting Diodes*, Clarendon Press, Oxford (1976).
F. Capasso, "Multilayer Avalanche Photodiodes and Solid State Photomultipliers," *Laser Focus/Electro-Optics*, July (1984).
H. C. Casey, Jr. and M. B. Panish, *Heterojunction Lasers*, Academic Press, New York (1978).
L. Figueroa and C. W. Slayman, "A Novel Heterostructure Interdigital Photodetector (HIP) with Picosecond Optical Response," *IEEE Electron Dev. Lett.* **EDL-2**, No. 8, Aug. (1981).
K. Gillessen and W. Shairer, *Light Emitting Diodes*, Prentice-Hall, New York (1987).
H. J. Hovel, in: *Solar Cells, Semiconductors and Semimetals*, Vol. 11 (R. K. Willardson and A. C. Beer, eds.), Academic Press, New York (1975).
S. Kim, H. Mohseni, M. Erdtmann, e. Michel, C. Jelen, and M. Razeghi, *Appl. Phys. Lett.* 73, 963, (1998).
H. Kressel, in: *Fundamentals of Optical Fiber Communications* (M. K. Barnoski, ed.), 2nd ed., Chap. 4, Academic Press, New York (1981).
C. H. Lee, *Picosecond Optoelectronic Devices*, Academic Press, New York (1984).
S. Maimon, F. Finkman, G. Bahir, S. E. Schacham, J. M. Garcia, P. M. Petroff, *Appl. Phys. Letts*., **73**, 2003 (1998).
H. Melchior, M. P. Lepselter, and S. M. Sze, "Metal–Semiconductor Avalanche Photodiode," IEEE Device Research Conf., Boulder, Colo., June 17–19 (1968).
H. Melchior, A. R. Hartman, D. P. Schinke, and T. E. Seidel, "Planar Epitaxial Silicon Avalanche Photodiode," *Bell Syst. Tech. J.* **57**, 1791 (1978).
R. J. McIntyre, "The Distribution of Gains in Uniformly Avalanche Photodiodes: Theory," *IEEE Trans. Electron Dev.* **ED-19**, 703 (1972).
J. Muller, "Photodiodes for Optical Communication," *Adv. Electron. Electron*. Phys. **55**, 189 (1981).
L. D. Partain, M. S. Kuryla, R. E. Weiss, R. A. Ransom, P. S. McLeod, L. M. Fraas, and J. A. Cape, "26.1% Solar Cell Efficiency for Ge Mechanically Stacked under GaAs," *J. Appl. Phys.* **62**, 3010 (1987).
J. Philips, P. Bhattacharya, S. W. Kennerly, D. W. Beekman, and M. Dutta, *IEEE J. of Quantum Electron.* **35**, 936, (1999).
G. E. Stillman, L. W. Cook, N. Tabatabaie, G. E. Bulman, and V. M. Robbins, "InGaAsP Photodiodes," *IEEE Electron Dev.* **30**, 364 (1983).
W. T. Tsang, "Lightwave Communication Technology: Photodetectors," in: *Semiconductors and Semimetals*, Vol. 22-D, Academic Press, New York (1985).
S. Y. Wang, S. D. Lin, H. W. Wu, C. P. Lee, *Infrared Physics & Technology* 42, 473, (2001).
R. K. Willardson and A. C. Beer, *Infrared Detectors, Semiconductors and Semimetals*, Vol. 12, Academic Press, New York (1976).

13
Light-Emitting Devices

13.1. Introduction

Photonic devices play an important role in a wide variety of applications in areas of optical communications, optical computing and interconnects, data transmission and signal processing, optical storage, sensors and optical imaging, solid-state lamps, and displays. Recent advances in III-V compound semiconductor growth and processing technologies have enabled these applications to become a reality. As a result, various photonic devices such as light-emitting diodes (LEDs), laser diodes (LDs), modulators, and photodetectors using III-V semiconductors have been developed for a wide variety of commercial applications. The LEDs are p-n junction diodes made from III-V and II-VI compound semiconductors that emit incoherent light under forward-bias conditions, while the LDs are p-n junction diodes with higher doping densities that emit coherent light for use in space and fiber-optic communications and data transmission, laser printers, CDs, and DVDs. In addition to high performance, low cost, and reliability, another factor in favor of LEDs is their compatibility with modern electronic devices as well as the increasingly important applications in visual displays. Low power, low operating voltage, small size, fast switching speed, and long life are some attractive features of LEDs. It is noted that the manufacturing technology for LEDs is compatible with silicon-integrated circuit technology. Depending on the complexity of visual tasks, LEDs are being used as solid-state lamps, symbolic and picture displays, data transmission, and in optical communications. LEDs coupled with silicon photodiodes can be used as optically isolated switches and sensing elements. With the technology breakthrough and cost reduction, the flat panel picture display using LEDs will soon become a reality for commercial applications. In this chapter the basic device physics and structures, operation principles, and general characteristics of various LEDs and LDs fabricated from III-V and II-VI compound semiconductors are discussed.

Section 13.2 describes the basic device physics and structures, the injection and recombination mechanisms, and the electrical and optical characteristics of an LED. An LED can emit incoherent light from the minority carrier injection in a forward-biased p-n junction diode followed by radiative recombination in

the p and n regions of the device. The basic mechanism of an LED involves the spontaneous emission of photons via the radiative recombination of electron–hole pairs, which converts the electrical energy into optical radiation. Random emission of incoherent optical radiation from a standard LED leads to a broad emission spectral line width of 10 nm or greater (on the order of $k_B T$). If a resonant cavity (RC) is added to the LED structure, then a very narrow spectral line width similar to that of an LD could be achieved in a RCLED. Section 13.3 presents different types of LEDs fabricated from III-V and II-VI compound semiconductor materials, such as GaN, InGaN, GaP, GaAsP, AlGaInP, AlGaAs, GaAs, InGaAsP, InP, SiC, and ZnSe. The emission spectra for LEDs fabricated from these materials cover the wavelengths from UV (ultraviolet), visible (violet, blue, green, yellow, amber, and red), to the near IR (infrared) spectral range. Ultrabright white, blue, green, yellow, orange, and red color LEDs have been developed in recent years using GaN- and AlGaInP-based material systems for a wide variety of commercial applications such as commercial lighting and display, traffic signals, automotive lighting and signals, instrument panel displays, and solid-state lamps for home and office uses.

Section 13.4 describes the basic device physics and structures, electrical and optical characteristics, and the performance parameters for semiconductor LDs. The conditions for population inversion and oscillation, the threshold current density, and the slope efficiency for an LD are described in this section. Section 13.5 presents recent developments of various edge-emitting single- and double-heterostructure (DH) LDs, vertical cavity surface-emitting laser (VCSEL) diodes, tunable LDs, and quantum-well (QW) lasers fabricated from III-V semiconductor materials. The commercial applications for these LDs such as solid-state light source, optical storage device (DVD, CD) and laser printers, optical networking, data transmission, and optical fiber communications are also discussed in this section.

13.2. Device Physics, Structures, and Characteristics of LEDs

In this section the basic device physics and structures, operating principles, and electrical and optical characteristics of an LED are presented. An LED is a semiconductor p-n junction light-emitting device, which under proper forward-biased conditions can emit spontaneous optical radiation in the wavelengths from the UV, visible, to IR regions of the electromagnetic spectrum. Depending on the semiconductor material used in the light-emitting layers (active layers), the wavelength of the emitted light can vary from the UV to the IR spectral range. Most commercially available LEDs are made from III-V compound semiconductors, while some LEDs are fabricated from ZnSe and SiC materials. An LED may be considered an electroluminescent device. The emission of light in such a device is accomplished by applying a sufficiently large forward-bias voltage across the p-n junction, followed by the minority carrier injection and radiative recombination taking place in the quasineutral n and p regions of the diode. Table 13.1 lists the most widely used

TABLE 13.1. Energy band gap, emitting wavelength, and color for some LED materials.

Material	Energy band gap (eV)	Wavelength (μm)	Color of LEDs
GaN	3.3	0.375	UV, blue
InGaN	0.67–3.3	0.375–1.85	Blue, green, red
AlGaN	3.3–6.2	0.2–0.375	UV, blue
SiC	2.86	0.435	Blue
ZnSe	2.70	0.46	Blue
GaP	1.98	0.63	Green, red
AlGaInP	1.35–1.98	0.63–0.92	Yellow, orange, amber, red
AlGaAs	1.43–2.19	0.57–0.83	Orange, red
GaAsP	1.41–1.95	0.63–0.88	Yellow, orange, red
GaAs	1.43	0.84	IR emitter
InGaAs	0.34–1.43	1.3	IR

LED material systems. These include a majority of the III-V compound semiconductors as well as some II-VI and IV-IV compound semiconductors such as ZnSe and SiC.

13.2.1. Injection Mechanisms

The basic requirement for radiative recombination to take place in a semiconductor is the injection of minority carriers into the bulk semiconductor. To explain the injection mechanism in an LED, Figure 13.1a shows the schematic drawing and

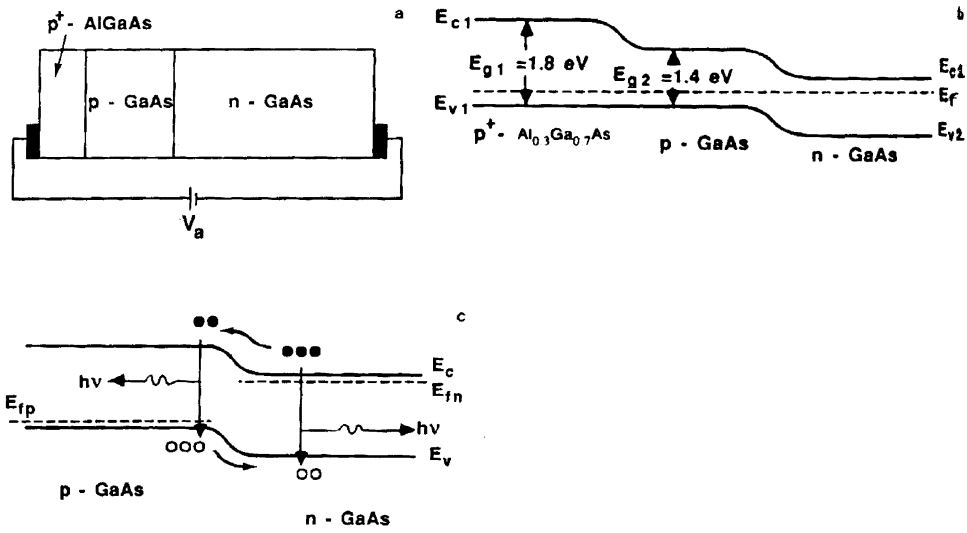


FIGURE 13.1. (a) Schematic drawing of a GaAs IR emitter with an AIGaAs window layer, (b) the energy band diagram in equilibrium, and (c) the energy band diagram under forward bias condition showing light emission from both p and n regions.

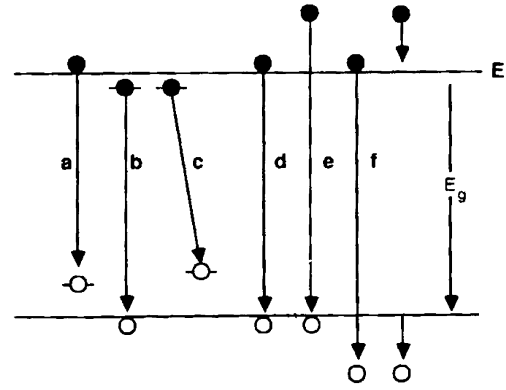


FIGURE 13.2. Possible electronic transitions that lead to radiative recombination in a semiconductor. (a) Conduction band to acceptor states, (b) donor states to valence band, (c) donor to acceptor states (pair emission), (d) conduction band to valence band (intrinsic emission), (e) hot carrier or avalanche emission, and (f) intraband transition. After Ivey,[1] by permission, © IEEE–1981.

Figure 13.1b the energy band diagram of an AlGaAs/GaAs p-n junction IR emitter under thermal equilibrium. The wide-band-gap p^+-$Al_{0.3}Ga_{0.7}As$ window layer is employed to reduce the surface recombination velocity and to increase the luminescent efficiency of the GaAs IR emitter. Figure 13.1c shows the energy band diagram of the same LED under forward-bias conditions. It is seen that under forward-bias conditions, electrons are injected from the n region into the p region while holes are injected from the p into the n region of the junction. For a direct band gap material such as GaAs, if the band-to-band radiative recombination is dominated on both sides of the junction, then the emission of optical radiation can be readily achieved.

13.2.2. Electronic Transitions

Figure 13.2 shows the possible electronic transitions in a semiconductor due to external excitations. These transitions may lead to either radiative or nonradiative recombination processes, which include (a) conduction band to acceptor states, (b) donor states to valence band, (c) donor to acceptor states (pair emission), (d) conduction band to valence band (intrinsic emission), (e) hot carrier or avalanche emission, and (f) intraband transition. For an efficient luminescent material, the radiative transition usually dominates the nonradiative process. In a direct-band-gap semiconductor such as GaAs, the emission of optical radiation is due mainly to the band-to-band radiative recombination, as shown by process (d) in Figure 13.2. The emission spectrum for such a transition is given by

$$I(h\nu) = \nu^2 (h\nu - E_g)^{1/2} \exp[-(h\nu - E_g)/k_B T], \tag{13.1}$$

which shows that the peak intensity occurs near the band gap energy of the semiconductor, and the theoretical full width at half maximum (FWHM) line width for the emission spectrum of an LED is $\Delta E = 1.8\, k_B T$, as shown in Figure 13.3.

If the electronic transition is from the band edge of one energy band to the impurity level near the opposite band (see process (a) and (b) in Figure 13.2), then the energy of the emitted photons will be slightly smaller than the band gap

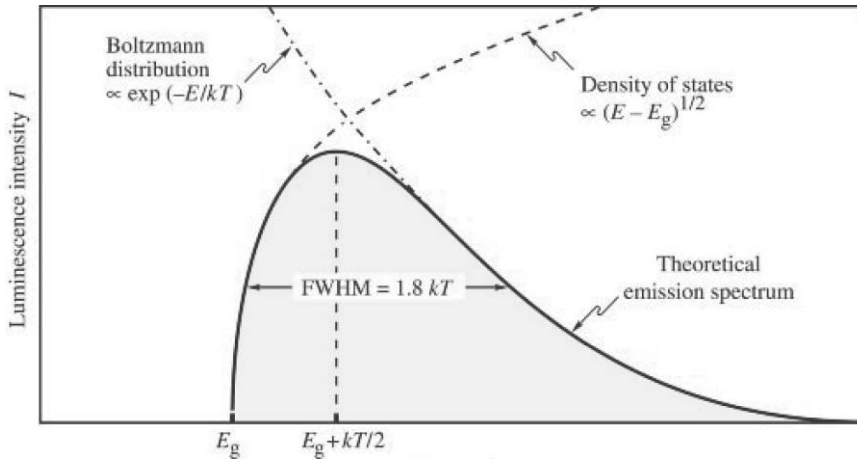


FIGURE 13.3. The emission spectrum of an LED due to band-to-band radiative recombination in a direct band gap semiconductor (see process "d" in Figure 13.2). After E. F. Schubert.[21]

energy of the material. The emission spectrum for the electronic transition from the conduction band edge to the acceptor level near the valence band is given by

$$I(h\nu) = \nu^2(h\nu - E_g + E_a)^{1/2}\{\exp[(h\nu - E - E_g + E_{fn})/k_B T] + 1\}^{-1}, \tag{13.2}$$

where E_a is the ionization energy of the acceptor level. The peak intensity occurs near $(E_g - E_a)$, and the width of the emission spectrum is also proportional to $k_B T$.

The peak emission wavelength of an LED for band-to-band radiative recombination can be calculated using the formula

$$\lambda_p = \frac{hc}{E_g} = \frac{1.24}{E_g}\,\mu\text{m}, \tag{13.3}$$

where E_g is the band gap energy of the semiconductor. For a GaAs IR emitter, with $E_g = 1.43$ eV, the peak emission wavelength is $\lambda_p = 0.873\,\mu$m at 300 K.

13.2.3. Luminescent Efficiency

The luminescent efficiency of an LED is defined as the ratio of total optical radiation output power associated with the radiative recombination process to the total input power. For a given input power, the radiative recombination process is in direct competition with the nonradiative processes such as the Auger and Shockley–Read–Hall (SRH) recombination processes occurring inside the LED. Therefore, in order to increase the luminescent efficiency it is important to increase

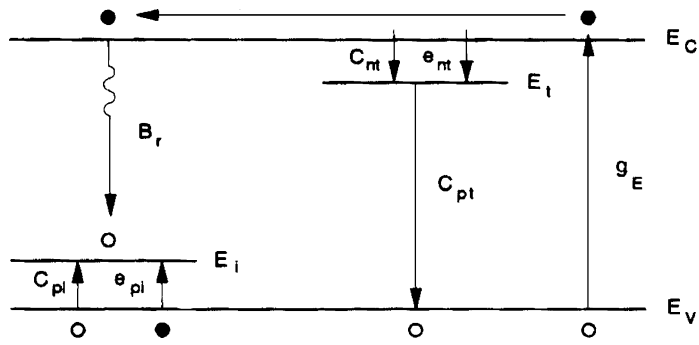


FIGURE 13.4. Energy band diagram showing emission and capture processes via a single electron trap and a luminescence center in the forbidden energy band gap of an LED. After Ivey,[1] by permission.

the radiative recombination process and in the meanwhile reduce the nonradiative recombination processes in the LED material.

To derive the luminescent efficiency of an LED, consider the radiative and nonradiative transition processes for a typical LED material, as shown in Figure 13.4. In the present case, it is assumed that only one electron trap level with activation energy E_t and density N_t exists below the conduction band edge. Furthermore, it is assumed that there is only one luminescent center with density N_l and activation energy E_l above the valence band edge. It should be noted that the recombination of electron–hole pairs via the electron trap E_t is nonradiative, while recombination via the luminescent center E_l is radiative. Under steady-state conditions, the rate equations for electrons in the conduction band and in the trap level are given, respectively, by

$$\frac{dn}{dt} = g_E - C_{nt} n (N_t - n_t) - B_r n p_l + e_{nt} n_t = 0, \tag{13.4}$$

$$\frac{dn_t}{dt} = C_{nt} n (N_t - n_t) - C_{pt} n_t p - e_{nt} n_t = 0, \tag{13.5}$$

where g_E is the external generation rate, C_{nt} is the electron capture rate at the E_t trap center, n_t is the electron density at the E_t trap level, B_r is the radiative capture rate at the E_l luminescent center, e_{nt} is the electron emission rate from the E_t trap, C_{pt} is the hole capture rate of the E_t trap, p_l is the hole density in the E_l, and N_l is the density of E_l center.

Similar rate equations can also be written for holes in the valence band and in the luminescent center. Solving (13.4) and (13.5) under steady-state conditions yields the external generation rate as

$$g_E = B_r n p_l + C_{pt} n_t p. \tag{13.6}$$

The first term on the right-hand side of (13.6) is due to radiative recombination, while the second term is attributed to the nonradiative recombination process.

Thus, the luminescent efficiency can be obtained using (13.6), which yields

$$\eta_{\mathrm{l}} = \left(\frac{B_{\mathrm{r}} n p_{\mathrm{l}}}{g_{\mathrm{E}}}\right) \times 100\% = \frac{1}{\left(1 + C_{\mathrm{pt}} n_{\mathrm{t}} p / B_{\mathrm{r}} n p_{\mathrm{l}}\right)} \times 100\%. \tag{13.7}$$

For the low-injection case, one can assume that the principle of detailed balance prevails between the electron trap level and the conduction band as well as between the luminescent center and the valence band. Thus, under thermal equilibrium condition, (13.4) and (13.5) become

$$e_{\mathrm{nt}} n_{\mathrm{t}} = C_{\mathrm{nt}} n_0 (N_{\mathrm{t}} - n_{\mathrm{t}}), \tag{13.8}$$

$$e_{\mathrm{pl}} p_{\mathrm{l}} = C_{\mathrm{pl}} p_0 (N_{\mathrm{l}} - p_{\mathrm{l}}). \tag{13.9}$$

Furthermore, it is assumed that the Fermi level is located between E_{t} and E_{l} such that

$$(N_{\mathrm{t}} - n_{\mathrm{t}}) \approx N_{\mathrm{t}}, \tag{13.10}$$

$$(N_{\mathrm{l}} - p_{\mathrm{l}}) \approx N_{\mathrm{l}}. \tag{13.11}$$

From Chapter 6, the relationships between e_{nt} and C_{nt} and between e_{pl} and C_{pl} are given, respectively, by

$$e_{\mathrm{nt}} = n_1 C_{\mathrm{nt}}, \tag{13.12}$$

$$e_{\mathrm{pl}} = p_1 C_{\mathrm{pl}}, \tag{13.13}$$

where

$$n_1 = n_0 \exp[-(E_{\mathrm{t}} - E_f)/k_{\mathrm{B}} T], \tag{13.14}$$

$$p_1 = p_0 \exp[-(E_f - E_l)/k_{\mathrm{B}} T]. \tag{13.15}$$

Solving (13.8) through (13.15), one obtains

$$\frac{n_{\mathrm{t}}}{p_{\mathrm{l}}} = \left(\frac{N_{\mathrm{t}}}{N_{\mathrm{l}}}\right) \exp\left[-\left(E_{\mathrm{f}} - E_{\mathrm{l}}\right)/k_{\mathrm{B}} T\right]. \tag{13.16}$$

Substituting (13.16) for $n_{\mathrm{t}}/p_{\mathrm{l}}$ into (13.7) yields the luminescent efficiency

$$\eta_{\mathrm{l}} = \frac{1}{1 + \left(C_{\mathrm{pt}} p N_{\mathrm{t}} / B_{\mathrm{r}} n N_{\mathrm{l}}\right) \exp\left[-\left(E_{\mathrm{t}} - E_{\mathrm{l}}\right)/k_{\mathrm{B}} T\right]}. \tag{13.17}$$

From (13.17) it is noted that the luminescent efficiency can be enhanced by increasing the density of luminescent centers or by decreasing the operating temperature. The luminescent efficiency of an LED can also be increased by reducing the energy separation between the luminescent level E_{l} and the valence band edge E_{v} (i.e., E_{l} should be as close to the valence band edge as possible), and the density of electron trap N_{t} must be kept as low as possible.

The minority carrier injection efficiency is another important parameter that governs the internal quantum efficiency of an LED. This parameter is directly related to the radiative recombination current, which is the dominant current component in an LED. Depending on the impurity profile and the external applied bias voltage, there are four current components in an LED that should be considered under

forward-bias conditions, namely, the electron diffusion current in the p quasineutral region, the hole diffusion current in the n quasineutral region, the recombination current in the depletion region, and the tunneling current across the junction barrier. The tunneling current is important only in a heavily doped p-n junction under small-forward-bias conditions and can be neglected in an LED operating under a moderate forward-bias condition. Since most of the luminescence is usually produced by the electron diffusion current inside the p quasineutral region, one can define the current injection efficiency of an LED as

$$\gamma = \frac{I_{\mathrm{n}}}{(I_{\mathrm{n}} + I_{\mathrm{p}} + I_{\mathrm{r}})}, \tag{13.18}$$

where

$$I_{\mathrm{n}} = \left(\frac{q D_{\mathrm{n}} n_{\mathrm{i}}^2}{L_{\mathrm{n}} N_{\mathrm{A}}}\right) A \left[\mathrm{e}^{qV/k_{\mathrm{B}}T} - 1\right], \tag{13.19}$$

$$I_{\mathrm{p}} = \left(\frac{q D_{\mathrm{p}} n_{\mathrm{i}}^2}{L_{\mathrm{p}} N_{\mathrm{D}}}\right) A \left[\mathrm{e}^{qV/k_{\mathrm{B}}T} - 1\right], \tag{13.20}$$

$$I_{\mathrm{r}} = \left(\frac{q n_{\mathrm{i}} W}{2\tau_0}\right) A \, \mathrm{e}^{qV/2k_{\mathrm{B}}T}. \tag{13.21}$$

The hole-diffusion current component given in (13.20) is usually small compared to the electron current component in a practical LED due to the high electron–hole mobility ratio, and hence (13.18) can be further simplified. The overall internal quantum efficiency of an LED is equal to the product of (13.17) and (13.18), which is given by

$$\eta_{\mathrm{i}} = \eta_{\mathrm{l}} \gamma. \tag{13.22}$$

13.2.4. External Quantum Efficiency

The single most important physical parameter for assessing the performance of an LED is the external quantum efficiency η_{E}. Even though the internal quantum efficiency η_{i} given by (13.22) can be quite high (e.g., $\eta_{\mathrm{i}} \geq 80\%$), the external quantum efficiency is usually lower than the internal quantum efficiency. This is due to significant losses of internal absorption and reflection taking place during light emission from the LED. A simple expression relating the external quantum efficiency to the internal quantum efficiency is given by

$$\eta_{\mathrm{E}} = \frac{\eta_{\mathrm{i}}}{\left(1 + \bar{\alpha} V / A \bar{T}\right)} = \frac{\eta_{\mathrm{i}}}{\left(1 + \bar{\alpha} x_{\mathrm{j}} / \bar{T}\right)}, \tag{13.23}$$

where $\bar{\alpha}$ is the average absorption coefficient and $\bar{T}$ is the total light transmitted within the critical angle θ_{c}, which is related to the transmissivity T by

$$\bar{T} = T \sin^2 (\theta_{\mathrm{c}}/2). \tag{13.24}$$

Here θ_c is the critical angle defined by Snell's law, which can be expressed by

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right). \tag{13.25}$$

The transmitivity T is given by

$$T = \frac{4n_1n_2}{(n_1 + n_2)^2}, \tag{13.26}$$

where n_1 and n_2 denote the refractive indices of the semiconductor and the ambient, respectively; V and A are the volume and the active area of the LED and x_j is the junction depth. For most LEDs, n_1 varies between 3.3 and 3.8, and $n_2 = 1$ for air. From (13.23), it is noted that the external quantum efficiency can be increased by reducing the junction depth x_j or by increasing $\bar{T}$. However, reducing the junction depth to less than one minority carrier diffusion length will increase the number of minority carriers diffused toward the surface. This may not be desirable, since a high surface recombination loss will reduce the internal quantum efficiency. For example, to reduce the surface recombination loss in a GaAs IR emitter, it is a common practice to incorporate a wider-band-gap AlGaAs window layer in the GaAs LED structure. Since the band gap energy of AlGaAs is larger than that of GaAs, the AlGaAs window layer is transparent to light emitting from the GaAs active region. It is worth noting that the interface state density between the AlGaAs and GaAs layers is usually much lower than that at the GaAs surface due to excellent lattice match between these two material systems. Therefore, the junction depth of an AlGaAs/GaAs LED can be greatly reduced. Reduction in the absorption of emitting photons in a GaAs IR emitter can be achieved by shifting the luminescence peak beyond the absorption edge of GaAs with photon energies $hv < E_g$. Higher external quantum efficiency can be obtained for this case, since the emitted photons fall beyond the absorption edge of the semiconductor where the absorption coefficient in the LED is very small. Figure 13.5a shows the emission spectra at 295 K and 77 K, and Figure 13.5b shows the external quantum efficiency versus temperature for a GaAs IR emitter reported earlier.

Other loss mechanisms, which may reduce the number of emitted photons and the external quantum efficiency, include the absorption loss within the LED, Fresnel loss, and critical angle loss. For example, the absorption loss for a GaAsP LED grown on the GaAs substrate could be quite large, since GaAs is opaque to visible light and can absorb about 85% of the photons emitted from a GaAsP LED. However, for a GaAsP LED grown on the GaP substrate, the absorption loss can be greatly reduced. In fact, only about 25% of the photons emitted from the active region of the GaAsP LED are absorbed by the GaP substrate. Therefore, the external quantum efficiency for such an LED can be greatly improved. Fresnel loss arises from the fact that when photons emit from a medium with a higher index of refraction (e.g., for GaAs, $n_1 = 3.66$) to a medium with a low index of refraction (e.g., $n_2 = 1$ for air), a portion of the light is reflected back to the medium interface (i.e., $R = (n_1 - n_2)^2/(n_1 + n_2)^2$). Finally, critical angle loss is caused by

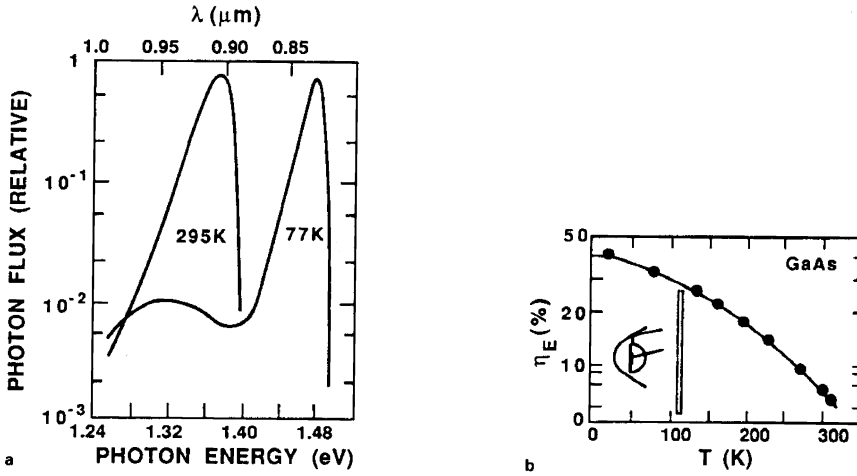


FIGURE 13.5. (a) Emission spectra at 295 K and 77 K and (b) the external quantum efficiency versus temperature for a GaAs infrared (IR) emitter. After Carr[4], by permission, © IEEE–1965.

total internal reflection of incident photons impinging on the surface of an LED at an angle greater than the critical angle θ_c, defined by (13.25).

Recent advances in LED technologies have greatly reduced the losses described above and as a result of technical breakthroughs the external quantum efficiency has been increased to over 30% for the InGaN/GaN-based blue/blue-green LEDs, and greater than 50% for the ultrabright AlGaInP/GaP red LEDs. Figure 13.6 shows the external quantum efficiency versus forward current for a red ($\lambda_p = 650$ nm) AlGaInP/GaP truncated inverted-pyramid (TIP) LED and a conventional large-junction (LJ) AlGaInP/GaP LED in power lamp packages. This AlGaInP/GaP TIP red LED exhibits a 1.4-fold improvement in extraction efficiency as compared to the LJ AlGaInP/GaP LED, resulting in a peak external quantum efficiency of 55% at $I_F = 100$ mA.

13.2.5. Device Structures and Electrical Characteristics

The device structures commonly used in an LED include the p-n homojunction, p-n heterojunction, and double heterojunction (DH) structures. Figure 13.7 shows the schematic energy band diagrams and free carrier distribution in (a) a p-n homojunction LED and (b) a p-n heterojunction LED under forward-bias conditions. In a homojunction LED the free carriers are distributed over the diffusion length, while in a heterojunction LED the free carriers are confined to the well region with more free carriers available for radiative recombination, and hence result in more efficient luminescence. Heterojunctions are widely used in the fabrication of high-performance LEDs and LDs. Figure 13.8 shows the schematic drawing of the most commonly used DH LED, consisting of a bulk or a multiquantum well

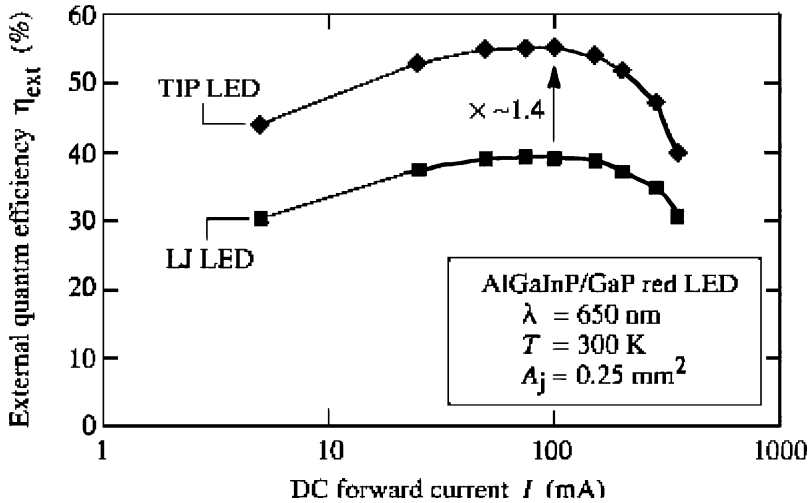


FIGURE 13.6. External quantum efficiency versus forward current for a red ($\lambda = 650$ nm) truncated-inverted-pyramid (TIP) LED and a conventional large-junction (LJ) LED in power lamp packages. The TIP LED has a 1.4-fold improvement in extraction efficiency as compared to the LJ - LED, resulting in a peak external quantum efficiency of 55% at 100 mA. After Krames et al., 1999.[3]

(MQW) active region and two confinement layers. The confinement layers are often called the cladding layers, which are formed using a wider-band-gap material than that in the active layer.

Figure 13.9 shows the schematic layer structures of two AlGaInP LEDs grown on (a) the GaAs absorbing substrate (AS) and (b) on the GaP transparent substrate (TS). In both structures the active layer is sandwiched between two carrier-confining layers, which are typically composed of $(Al_xGa_{1-x})_{0.5}In_{0.5}P$,

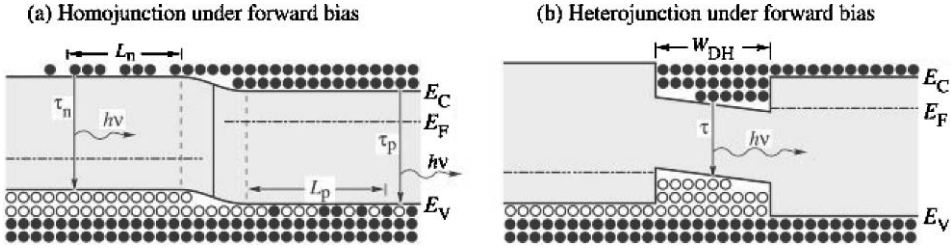


FIGURE 13.7. Schematic energy band diagrams for (a) a p-n homojunction LED and (b) a p-n heterojunction LED under forward-bias conditions, showing free-carrier distribution and radiative recombination emission. In homojunctions carriers are distributed over the diffusion length, while in heterojunctions more carriers are confined to the well region, leading to more efficient luminescence. After E. F. Schubert.[2]

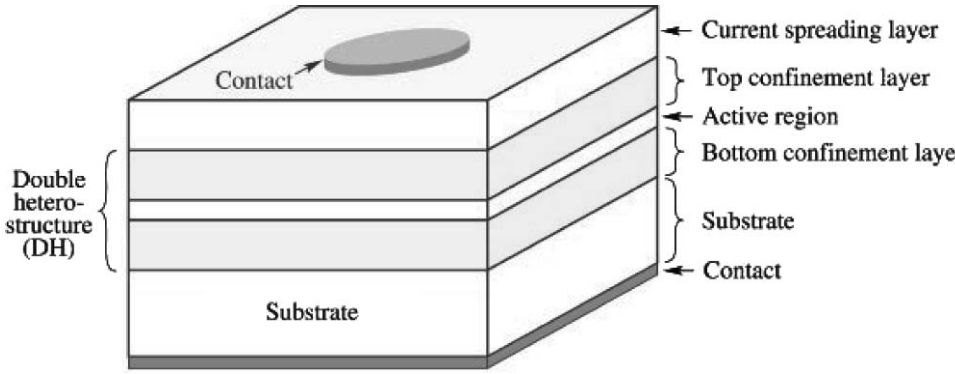


FIGURE 13.8. A schematic drawing of a double-heterostructure (DH) LED consisting of a bulk or multiquantum well (MQW) active region and two confinement layers. The confinement layers are often called the cladding layers, which are formed by wider-band-gap materials than those in the active layer. After E. F. Schubert.[2]

with $x > 0.7$. The top contact layer, usually GaP or AlGaAs, serves as both a current-spreading layer and a window layer to improve extraction of light directed toward the side of the chip. The structure shown in Figure 13.9a is an AS LED. In AS devices a distributed Bragg reflector (DBR) layer is often grown below the lower confining layer to increase on-axis light emission, and reduces the light absorption in the GaAs substrate. The structure shown in Figure 13.9b is a TS LED, in which the absorbing GaAs substrate is replaced by an optically transparent GaP

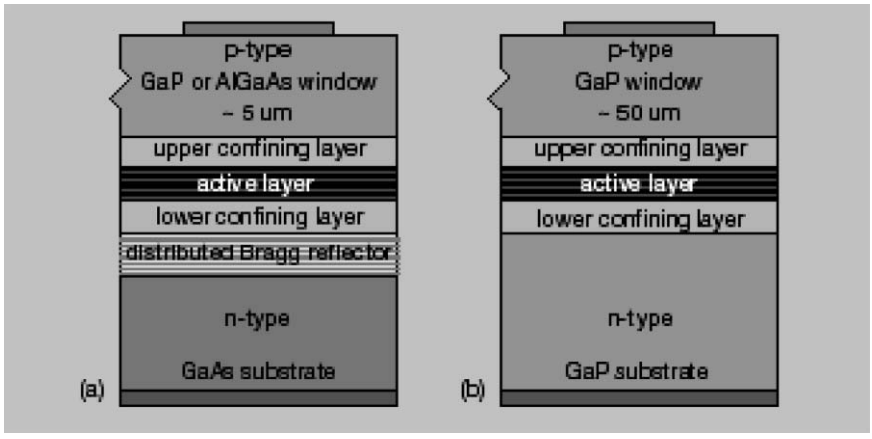


FIGURE 13.9. The schematic layer structures of two AlGaInP LEDs grown on (a) the GaAs absorbing substrate (AS) and (b) on the GaP transparent substrate (TS). In both structures the active layer is sandwiched between two carrier-confining layers, which are typically composed of $(Al_xGa_{1x})_{0.5}In_{0.5}P$, with $x > 0.7$.[7]

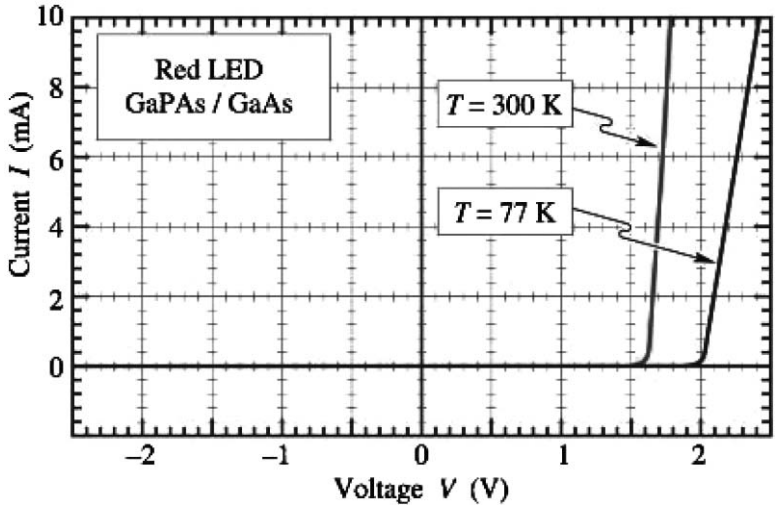


FIGURE 13.10. Current–voltage (*I–V*) characteristics of a GaAsP/GaAs LED emitting in the red part of the visible spectrum, measured at 77 K and 300 K. The threshold voltages are 2 V and 1.6 V at 77 and 300 K, respectively. After E. F. Schubert.[2]

substrate by solid-state wafer bonding. Both structures are aimed at enhancing the light-emitting efficiency in both LEDs.

The electrical characteristics of LEDs can be described by the forward current–voltage (I_F–V_F) characteristics of a p-n junction diode. Figure 13.10 shows the *I–V* characteristic of a GaAsP/GaAs LED emitting in the red part of the visible spectrum, measured at 77 and 300 K, and the threshold voltages for this LED are 2 and 1.6 V at 77 and 300 K, respectively. The forward-bias voltage required to drive an LED at a constant forward current (e.g., $I_F = 20$ mA) depends on the band gap energy of the semiconductor; typically, the larger-band-gap LED (e.g., GaN- LEDs) needs higher bias voltage. Figure 13.11 shows the forward voltage versus energy band gap for different LED materials, operating at 20 mA forward current. As can be seen in this figure, most of the LEDs in the visible spectrum can be operated in the 1.5–3.5 V range at a 20 mA forward current. The blue LED requires higher bias voltage, while the IR emitter operates at a much lower bias voltage.

Another important issue of LED operation is the light extraction scheme and packaging for different applications. Figure 13.12 shows (a) several LED lens geometries designed to increase light extraction or optical efficiency and (b) the radiation patterns of LEDs with (i) retangular, (ii) hemispheric, and (iii) parabolic geometries. Figure 13.13 shows some common LED packages: (a) LED with hemispherical epoxy dome and (b) LEDs with cylindrical and retangular epoxy packages. These packages are suitable for signal and panel displays as well as solid-state lamp applications. For fiber-optic communications the circular surface-mount LED chips are used to couple the LEDs with the optical fiber core. Both

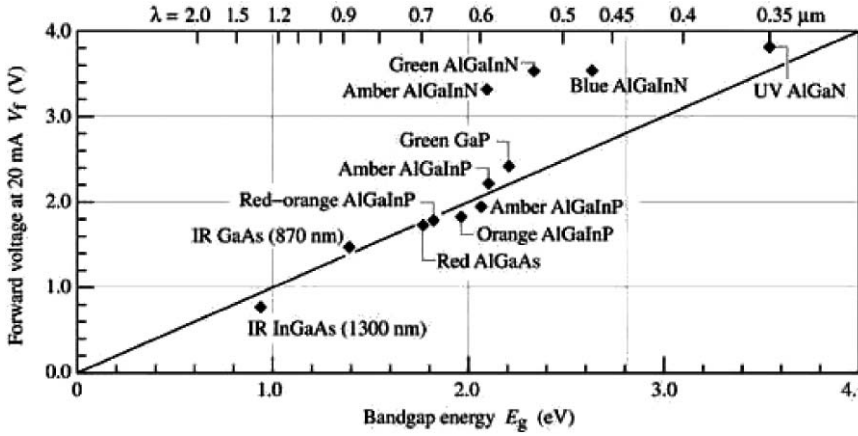


FIGURE 13.11. Typical diode forward voltage versus energy band gap for LEDs fabricated from different semiconductor materials, measured at a forward current of $I_F = 20$ mA. After Krames et al., 2000.[2]

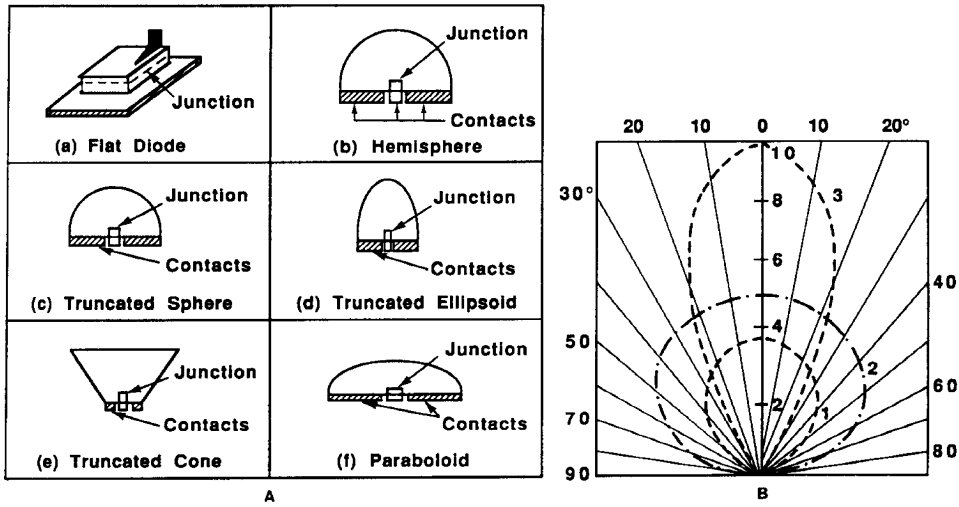


FIGURE 13.12. (a) Some LED lens geometries designed to increase light extraction or optical efficiency; (b) radiation patterns of LEDs with (1) rectangular, (2) hemispherical, and (3) parabolic geometries. After Galginaitis,[5] with permission.

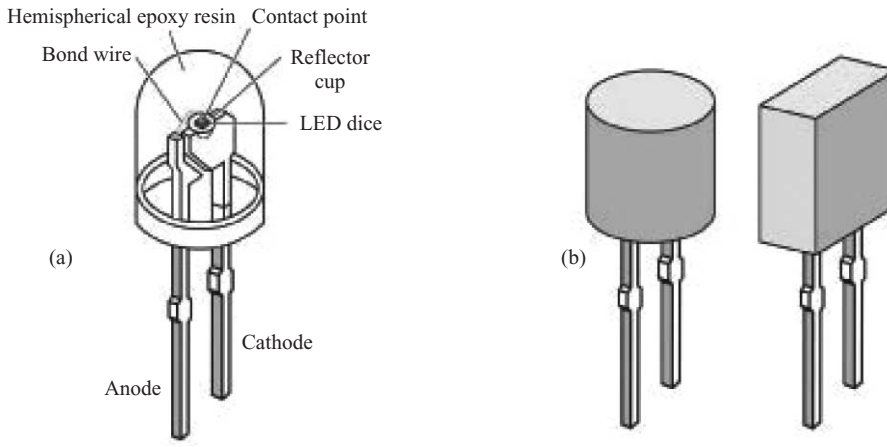


FIGURE 13.13. Typical LED packages: (a) LED with hemispherical epoxy dome and (b) LEDs with cylindrical and rectangular epoxy packages. (After E. F. Schubert.[2])

GaAs (0.85 μm emission peak) and InGaAs (1.3 μm) IR emitters are widely used for fiber-optic communications.

13.3. LED Materials and Technologies

13.3.1. Introduction

LEDs are devices designed to efficiently convert electrical energy into electromagnetic radiation, most of which is visible to the human eye. The semiconductor LEDs are most familiar as the little glowing red or green indicators on the electronic equipment and consumer electronics. Visible LEDs, introduced commercially in 1960s, offer the advantage of efficient direct monochromatic emission. However, until recently the commercial use of visible LEDs has been largely confined to indicator and display applications. The recent efficiency improvements of MOCVD-grown GaInN/GaN (white, blue, and green) and AlGaInP (red, orange, amber, yellow, and green) LEDs have enabled their use in a wide range of applications such as exterior automotive lighting, traffic signals, full-color outdoor signs, and solid-state lamps for home and office use. Further efficiency improvements and manufacturing cost reduction will enable LED-based systems to compete in the $40 billion lighting market with conventional technologies such as incandescent bulbs, fluorescent lighting, and neon and sodium vapor lamps. For LEDs that are used in optical fiber communication systems, efficient spontaneous emission originating from the excitation is favorable for reducing input power, and hence a p-n heterojunction LED structure is used for this purpose. Bulk semiconductor materials usually form the active layer in these LEDs. In LEDs that emit visible light

for display use, however, a variety of structures may be used. In the visible LEDs, the light originating in the spontaneous emission process is emitted in all directions from the light-emitting region (active layer). For that reason, several structures restricting the emitted light to a certain direction have been developed. These structures are divided into two groups: the surface-emitting and the edge-emitting types. The surface-emitting LEDs emit the light in a direction perpendicular to the p-n junction plane, while the edge-emitting LEDs emit the light in a direction parallel to the p-n junction plane.

In LED design, one needs first to characterize the physical parameters discussed above so that the performance of an LED can be optimized for a specific application. The design considerations are similar for LEDs fabricated from both direct and indirect band gap semiconductor materials. However, there are distinct differences between the direct and indirect band gap materials, which include a larger optical absorption coefficient for the direct-band-gap semiconductor in the case of light generation at the junction, and the need to introduce luminescence centers in an indirect-band-gap semiconductor to produce radiative recombination.

The most efficient LEDs using indirect-band-gap materials are the red and green GaP LEDs, while direct-band-gap materials such as GaN, $In_xGa_{1-x}N$, $(Al_xGa_{1x})_{0.5}In_{0.5}P$ $(x \leq 0.53)$, GaAs, $GaAs_{1-x}P_x$ $(x \leq 0.45)$, $Ga_{1-x}Al_xAs$ $(x \leq 0.44)$, SiC, and ZnSe have been widely used in fabricating visible LEDs. Applications of LEDs for optical display in the visible spectrum require that wavelengths of the emitted photons from these LEDs fall between 0.45 and 0.68 μm. Therefore, materials useful for this spectral range should have energy band gaps varying between 1.8 and 2.7 eV. However, for fiber-optic communication applications, IR-driven LEDs such as GaAs (0.85 μm), $In_{0.53}Ga_{0.47}As$ (1.3 μm), and InGaAsP (1.55 μm) IR emitters are the prime candidates. Some of the commercially available LEDs are discussed next.

13.3.2. AlGaInP LEDs

Hewlett Parkard (HP) and Toshiba developed the first high-brightness AlGaInP LEDs. The highest-efficiency LEDs demonstrated to date come from the quaternary AlGaInP material system, which encompasses the amber through red color spectral regime. $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ is lattice matched to GaAs for Al compositions ranging from $x = 0$ to $x = 1$, and has a direct band gap for $x \leq 0.53$. In the direct-band-gap compositional range, $(Al_xGa_{1x})_{0.5}In_{0.5}P$ emits over the red ($E_g = 1.9$ eV) to yellow-green ($E_g = 2.26$ eV) spectral range; however, the radiative efficiency drops rapidly with higher Al content as the alloy approaches the direct/indirect-band-gap crossover. Thus, commercial AlGaInP LEDs are primarily limited to red, orange, and amber emission. Figure 13.14 shows the energy band gap and the corresponding wavelength versus lattice constant of $(Al_xGa_{1-x})_yIn_{1-y}P$ at 300 K; the dashed vertical line shows $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ lattice-matched to GaAs.[6] Figure 13.15 shows the historical development stages for the high-efficiency and

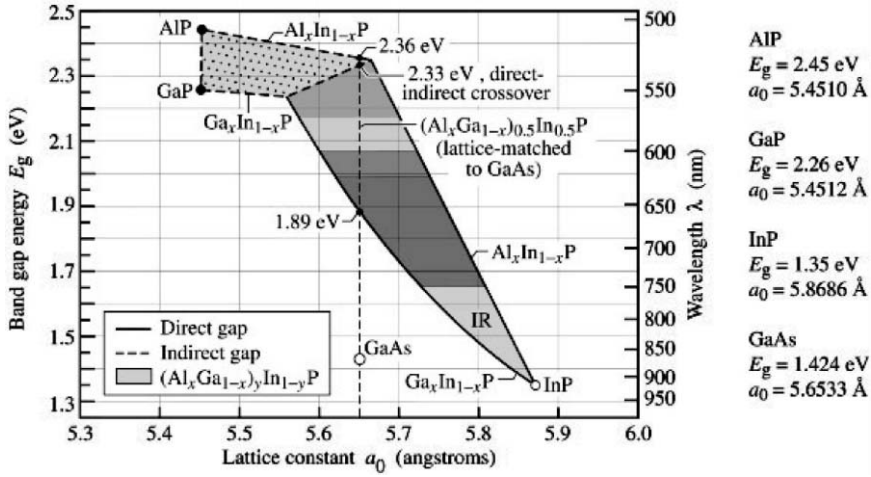


FIGURE 13.14. Energy band gap and corresponding wavelength versus lattice constant for AIP, GaP, $Ga_xIn_{1-x}P$, $Al_xIn_{1-x}P$, $(Al_xGa_{1-x})_yIn_{1-y}P$, GaAs, and InP material systems. After Chen et al., 1997.[6]

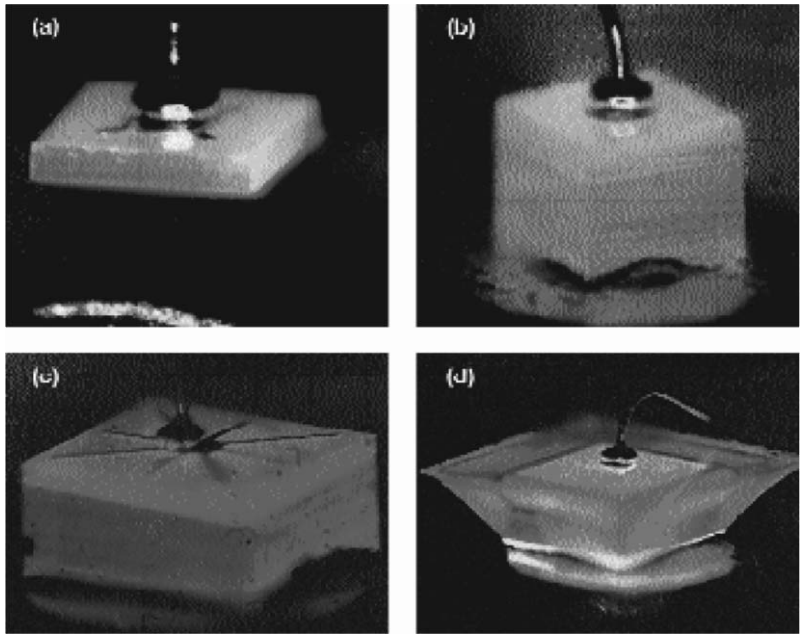


FIGURE 13.15. Historical development sequence of four generations of AlGaInP LEDs: (a) the absorbing substrate (AS) LED, (b) the transparent substrate (TS) LED with 2–3 times the AS flux, (c) the high-power LED with 5 times the TS flux, and (d) the truncated inverted pyramid (TIP) LED with 8 times the TS flux.[7]

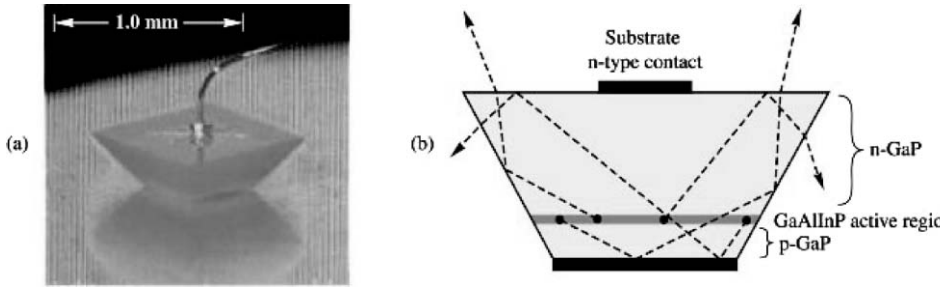


FIGURE 13.16. A truncated inverted pyramid (TIP) AlInGaP/GaP LED: (a) LED driven by an electrical injection current, and (b) schematic diagram of the LED showing the enhanced light extraction efficiency. After Krames et al., 1999.[3]

high-power LEDs using the AlInGaP material system that cover four generations of LEDs: (a) LED grown on the GaAs absorbing substrate (AS), (b) LED grown on the GaP transparent substrate (TS) with two to three times the AS flux, (c) the high-power LED with five times the TS flux, and (d) the truncated inverted pyramid (TIP) LED with eight times the TS flux. Figure 13.16 illustrates a TIP AlGaInP/GaP red LED driven by (a) an electrical injection current and (b) enhanced light extraction efficiency.

Typical illumination systems require photometric output power of several hundred lumens. Simply increasing the number of conventional LEDs is often impractical for such systems. The simplest way of increasing the flux per LED is to make the chip bigger. One example of a chip with increased die area is shown in Figure 13.15c for a high-power AlGaInP/GaP LED. The junction area of this LED chip is approximately five times that of the conventional die packaged in a 5-mm LED lamp. Therefore, driving the larger die with five times the current of the conventional die (equivalent current density) should in principle increase the flux fivefold. The highest luminous efficiency measured from a conventional 5-mm LED lamp made with TS wafers is shown in Figure 13.17 as a function of peak wavelength. The lamps have luminous efficiencies exceeding 50 lm/W at a current density of 40 A/cm^2 over the color range used for commercial AlGaInP LEDs. The highest luminous efficiency for the 5-mm lamps is 74 lm/W for lamps emitting at 615 nm. The external quantum efficiency of AlGaInP LEDs improves with increasing wavelength (lower Al composition) due to better carrier confinement, higher relative electron population of the direct minimum, reduced nonradiative impurity incorporation, and reduced absorption. For these LED lamps an external quantum efficiency of 32% at 632 nm has been achieved. However, the luminous efficiency drops due to the decreasing response of the human eye with increasing wavelength, as indicated by the CIE curve shown in Figure 13.17. The maximum DC output flux from 5-mm lamps driven at 50 mA is limited to 510 lm per LED. Super-red 5-mm AlGaInP LEDs with peak emission wavelength at 638 nm and luminous intensity of 1500 mcd are commercially available for lamp applications. This red LED has a power dissipation rating of 150 mW at 50 mA forward current

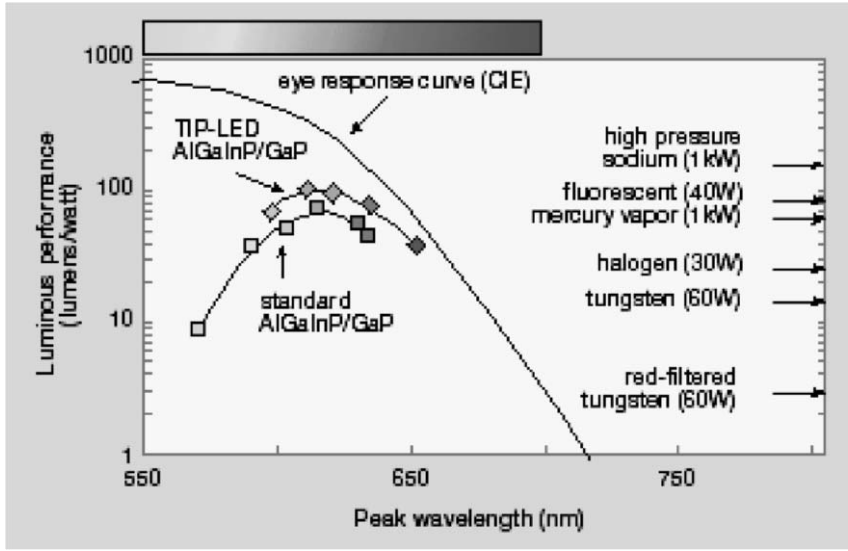


FIGURE 13.17. The measured luminescence efficiency for a standard 5-mm AlGaInP TIP LED lamp and a high-power lamps using AlGaInP TIP LED chips ($J_F = 40$ A/cm^2. The 611 nm TIP LED has a luminescence efficiency of 102 lm/W, and the 652 nm TIP LED has an external quantum efficiency of 55%. The luminescence efficiencies for all the conventional lamps and human eye response curve are also shown on the right for comparison purposes.[7]

and 2.2 V forward bias. High-brightness yellow, green, and orange color LEDs fabricated from AlGaInP/GaAs material systems are now available for a wide variety of applications.

13.3.3. GaN-Based LEDs

GaN is considered the most "environmentally friendly" III-V compound material available for the fabrication of LEDs and LDs. In comparison to toxic GaAs LEDs or even mercury-containing fluorescent lamps, GaN offers a truly safe lighting solution. The first successful development of GaN-based blue LEDs was reported by Dr. Nakamura, of Nichia Chemical Industries Ltd., Japan, in early 1990. Nichia is the world's leading manufacturer of GaN-based blue, green, and white LEDs and phosphors for lighting. Dr. Nakamura started the GaN effort in 1989 at a time when all the other optoelectronic companies were pursuing II-VI technology. His foresight has allowed Nichia to develop a large technology lead in nitride semiconductor technology. Nichia controls the market for wide-band-gap ($E_g = 3.3$ eV) GaN LED devices and is selling more than 20 million LEDs every month. The best external quantum efficienies achieved by Nichia for the blue and green LEDs were 10% and 12%, respectively, in 1998. Today, external quantum efficiency exceeding 30% has been achieved for GaN-based LEDs. The $In_xGa_{1-x}N$ LEDs ($E_g = 3.3$ eV,

for $x = 0$ and $E_g = 1.8$ eV, for $x = 1$) are capable of producing UV, blue, green, and red color light, although there is more efficiency in the blue and green spectral regimes. Super-blue GaN LEDs and InGaN LEDs with emission peaks at 463 and 470 nm and luminous intensities in the range of 2400–5500 mcd (operating at $V_F = 3.5$V and $I_F = 20$–30 mA) are commercially available for lamps and other applications. Power dissipation for these LEDs is in the range of 100–120 mW. Super-green 5 mm GaN/InGaN LEDs with peak emission wavelength at 525 nm and luminous intensity of 8,000 mcd are also commercially available for lamp applications.

The high-temperature performance of GaN amber LEDs is far superior to AlGaInP amber LEDs. The wavelength shift as a function of temperature is much smaller in GaN-based than in GaAs-based LEDs. A comparison of InGaN-based and AlGaInP-based LEDs operating at an elevated temperature of 80°C reveals that the InGaN LED light output is decreased by only 20%, whereas the AlGaInP LED light output is down by 70%. The excellent temperature performance for the GaN-based LEDs in comparison to the GaAs-based LEDs is also seen when GaN-based LEDs are compared to GaP green and AlGaAs red LEDs. The InGaN yellow LEDs are not as bright as the TS AlGaInP LEDs, but their performance rivals AS AlGaInP LEDs. Recently, Nichia has successfully developed an efficient UV (372 nm) LED. The UV LEDs are expected to find new applications in UV plastic curing, lighting, sterilization, medical, and counterfeit currency detection. Figure 13.18 shows a schematic drawing of an InGaN/GaN multiquantum well (MQW) LED structure grown on sapphire substrate. The LEDs are grown on sapphire substrates and incorporated with an InGaN/GaN MQW active region, which is sandwiched between n- and p-AlGaN/GaN superlattice and GaN cladding layers. A range of

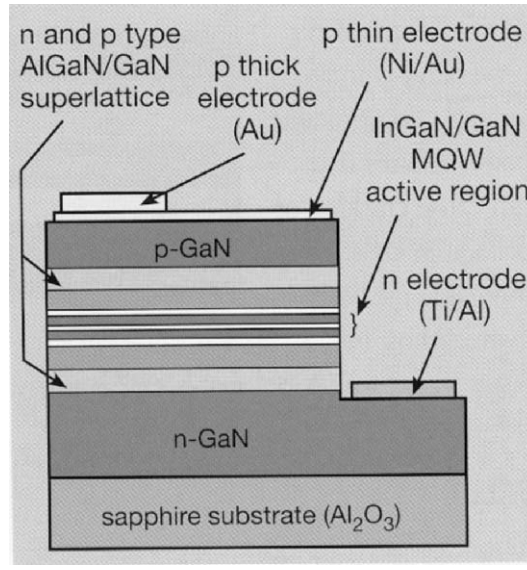


FIGURE 13.18. The layer structure of an InGaN/GaN LED grown on a sapphire substrate with an InGaN/GaN multiquantum well (MQW) active region, which is sandwiched between n- and p-AlGaN/GaN superlattice and GaN cladding layers.[8]

blue, blue/green, and yellow colors has been realized in the InGaN-based LEDs with corresponding emission peaks at 460, 500, and 560 nm, respectively. The performance of these blue and blue/green InGaN/GaN LEDs are also shown in Figure 13.18. Blue GaN LEDs and high-brightness blue and green InGaN LEDs grown on SiC substrates are now available for a wide range of commercial applications.

13.3.4. GaP-Based LEDs

Gallium phosphide (GaP), which is an indirect wide-band-gap material with an energy band gap of $E_g = 2.26$ eV at 300 K, is widely used for fabricating red and green LEDs. By doping GaP LEDs with isoelectronic impurities such as N and Zn-O, green and red light emission can be obtained from these doped GaP LEDs. For example, a red LED can be fabricated from Zn–O-doped GaP, while a green GaP LED is obtained using nitrogen (N)-doped GaP. In general, radiative recombination in an indirect band gap material such as GaP can be achieved by the excitonic radiative recombination process via luminescent impurity centers such as the N or ZnO center in the forbidden gap. The physical principles and characteristics of a red and green GaP LED are discussed next.

The zinc and oxygen (ZnO) doped GaP p-n junction diode grown by the LPE technique is an efficient red LED for commercial use. The ZnO pair impurity is an isoelectronic trap in GaP, which can replace an adjacent Ga-P pair of atoms to form a recombination center with ionization energy of 0.3 eV below the conduction band edge. Radiative recombination of electrons and holes in the ZnO centers will lead to the emission of red light ($h\nu = 1.95$ eV) from such an LED. The energy of the emitting photons from such a radiative recombination process is given by

$$h\nu = E_g - E_{DA} + \frac{q^2}{4\pi\varepsilon_0\varepsilon_s r}, \tag{13.27}$$

where $h\nu$ is the photon energy, E_{DA} is the activation energy of the ZnO center, and the last term in (13.27) is due to Coulomb potential energy of the ZnO pair separated by a distance r. The red GaP: ZnO LED produces an emission peak at 648 nm and has an emission half-width (FWHM) of 93 nm, as shown in Figure 13.19. Typical external quantum efficiencies of 2–3% have been obtained at a current level of 10 A/cm^2. Luminous performance of 1 lm/W has been achieved in ZnO-doped GaP LED. The switching speed for a typical red GaP LED is about 100 ns. Most of the commercial red LEDs are fabricated from GaAsP and AlGaInP material systems that produce higher external quantum efficiency and luminous intensity for various lamps applications. The emission peak wavelength can be controlled by varying the alloy composition x in $GaAs_xP_{1-x}$ and $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ material systems. Typical emission wavelengths for these red LEDs may vary from 626 to 660 nm.

The emission of green light ($\lambda_p = 563$ nm) from a GaP LED can be achieved in N-doped (5×10^{18} cm^{-3}) GaP LEDs grown by vapor-phase epitaxy (VPE) or liquid-phase epitaxy (LPE) techniques. The emission mechanism is due to radiative recombination of electron–hole pairs (exciton recombination) at a nitrogen impurity center on a phosphorus site. The nitrogen impurity is an isoelectronic

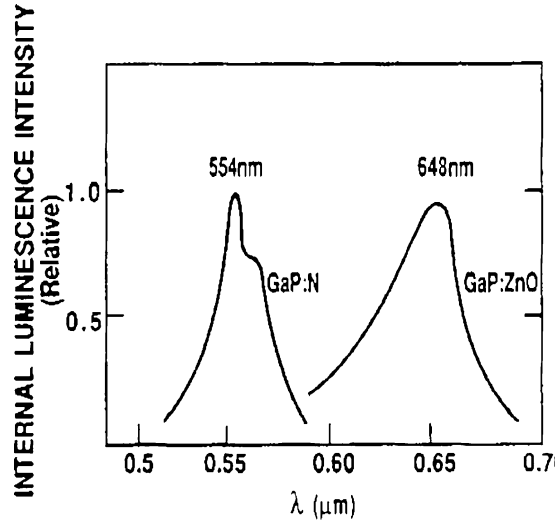


FIGURE 13.19. Normalized internal luminescence intensity spectra of a red ZnO doped GaP LED and a green N-doped GaP LED.

trap in GaP, which can replace the phosphorus atom to achieve green emission ($hv = 1.95$ eV and $\lambda = 554$ nm) in a GaP LED. A nitrogen isoelectronic trap is a highly localized potential well that can trap an electron and become charged. The resulting Coulomb field attracts a hole, which pairs with the trapped electron to form an exciton (i.e., a hydrogenlike bound electron–hole pair). The annihilation of this exciton via a radiative recombination process gives rise to a green emission with wavelength equal to 554 nm at 300 K. Because of other nonradiative recombination processes, the external quantum efficiency for a green N-doped GaP LED is usually a few percent. However, despite its low external quantum efficiency, the N-doped GaP LED provides high brightness, since green emission is near the peak of human-eye sensitivity. Figure 13.19 shows the normalized luminescence intensity spectra for a red GaP:ZnO LED and a green GaP:N LED.

Emission of yellow light can also be accomplished in a N-doped GaP LED if the nitrogen doping density is greater than 2×10^{19} cm^{-3}. Using a high nitrogen-doping density in a GaP LED will lead to a shift in the emission peak to a longer wavelength (from green to yellow light) because of the excitonic recombination at the N–N nearest-neighbor complexes. However, most of the commercial high-brightness yellow LEDs are now fabricated from AlGaInP material systems with peak emission wavelength at 589 nm.

13.3.5. GaAsP and AlGaAs LEDs

In a direct-band-gap semiconductor, the color of light emission from an LED depends on the energy band gap of the LED material. In III-V ternary compound semiconductors such as $Al_xGa_{1-x}As$, $GaAs_{1-x}P_x$, and $In_xGa_{1-x}N$, the energy band gap can be altered by varying the alloy composition x. By changing the value

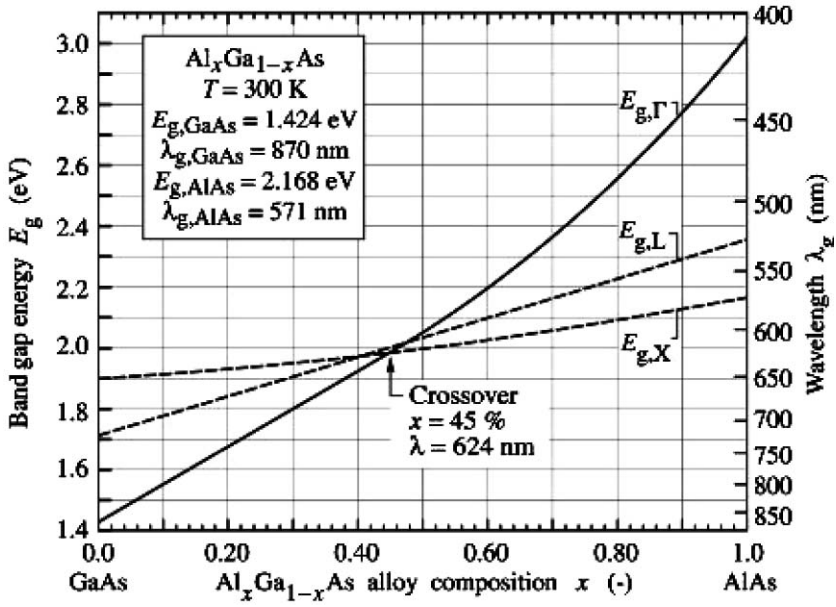


FIGURE 13.20. The band gap energy and emission wavelength of $Al_xGa_{1-x}As$ LEDs at room temperature. E_Γ is the conduction band minimum at Γ point, while E_L and E_x denote the indirect conduction band minima at the L and X points in the Brillouin zone, respectively.[2]

of x and hence the energy band gap, one can change the color of light emitted from these LEDs. The $Al_xGa_{1-x}As$ and $GaAs_{1-x}P_x$ materials are the two most commonly used ternary compound semiconductors for LED fabrications in the visible spectral range. The energy band gap for $Al_xGa_{1-x}As$ can be varied from 1.43 eV for $x = 0$ to 2.19 eV for $x = 1$. It is seen that $Al_xGa_{1-x}As$ is a direct-band-gap semiconductor for $x < 0.45$, and becomes indirect-band-gap material for $x > 0.45$. For example, if the aluminum mole fraction x is chosen equal to 0.3 (i.e., $E_g = 1.8$ eV), then red emission can be obtained from a $Ga_{0.3}Al_{0.7}As$ LED via band-to-band radiative recombination. Figure 13.20 shows the band gap energy and emission wavelength of $Al_xGa_{1-x}As$ at room temperature. Here E_Γ is the conduction band minimum at the Γ point, while E_L and E_X denote the indirect conduction band minima at the L and X points in the Brillouin zone, respectively. The band gap energy as a function of Al composition x at the Γ, L, and X points can be calculated using the formula

$$
\begin{aligned}
E_{g\Gamma} &= 1.424 + 1.247x \text{ (eV)} \qquad (0 \le x \le 0.45) \\
E_{g\Gamma} &= 1.424 + 1.247x + 1.147\,(x - 0.45)^2 \qquad (0.45 \le x \le 1.0) \\
E_{gL} &= 1.708 + 0.642x \qquad (0 \le x \le 1.0) \\
E_{gX} &= 1.900 + 0.1252x + 0.143x^2 \qquad (0 \le x \le 1.0).
\end{aligned} \tag{13.28}
$$

Since the band-to-band radiative recombination is the dominant recombination process in a direct-band-gap material, it is expected that the external quantum efficiency for a $GaAs_{1-x}P_x$ LED will decrease with increasing alloy composition and energy band gap. Nitrogen (N) is found to enhance the radiative recombination and the external quantum efficiency of a GaAsP LED. The brightness of a $GaAs_{1-x}P_x$ LED is seen to peak around $E_g = 1.9$ eV; this peak corresponds to a phosphorus mole fraction of $x = 0.4$, and the peak emission wavelength for such an LED is 660 nm, which falls in the red spectral range. Further increase in phosphorus mole fraction x (and hence the energy band gap) in a $GaAs_{1-x}P_x$ LED will shift the emission peak toward the orange color with decreasing external luminescent efficiency and brightness.

A planar technology has been employed to fabricate $GaAs_{0.6}P_{0.4}$ LED arrays for numeric and alphanumeric displays. In this technology the transparent GaP substrate is used for fabricating the $GaAs_{1-x}P_x$ LEDs (with $x \geq 0.5$) to avoid absorption of light emitted from the $GaAs_{1-x}P_x$ active layer by the GaP substrate. When GaAs substrate is used for growing the $GaAs_{1-x}P_x$ LEDs, because of the lattice mismatch between the GaAs and GaP material system it is necessary to grow a $GaAs_{1-x}P_x$ graded epilayer by gradually increasing the alloy composition x from the surface of GaAs substrate to the top of the $GaAs_{1-x}P_x$ epilayer in order to create lattice match with the GaAs substrate during the epitaxial layer growth. The junction can be formed by changing the dopants during vapor-phase deposition or by zinc diffusion into a uniformly doped structure of the graded composition. A heavily doped p^+-$GaAs_{1-x}P_x$ is usually grown on top of the p-$GaAs_{1-x}P_x$ active layer to lower the contact resistance of the LED.

13.3.6. GaAs LEDs

Gallium arsenide, a direct-band-gap material with an energy band gap of 1.43 eV at 300 K, is widely used for the fabrication of near-IR-emitting diodes with emission peak wavelength at 890 nm. The GaAs LED is the most efficient and widely used near-IR light source for a variety of applications ranging from optical communications, signal processing, fiber optic links, to optical computing. A GaAs LED can be readily fabricated using zinc diffusion in an n-type GaAs substrate to form a p-n junction. High external quantum efficiency can be obtained in a GaAs IR emitter by using a dopant density of around 10^{18} cm^{-3} in both the n- and p-type regions. If a Si-doped GaAs substrate is used for fabricating the GaAs LED, then its emission peak will shift to 1.32 eV, which is below the absorption edge of the GaAs material. As a result, the self-absorption effect in such a GaAs LED is greatly reduced and the external quantum efficiency can be greatly improved ($\eta_{ext} \approx 20\%$). Other features of a GaAs LED include high switching speed and fast recovery time (i.e., 2–10 ns), which make GaAs LED ideal for data transmission applications. Although direct-band-gap GaAsP LEDs have the fastest switching speed, the best luminescence efficiency and color coverage among the LED family of phosphor-coated GaAs LEDs have the edge in color coverage, but fall well behind in both switching speed and light conversion efficiency. In addition to

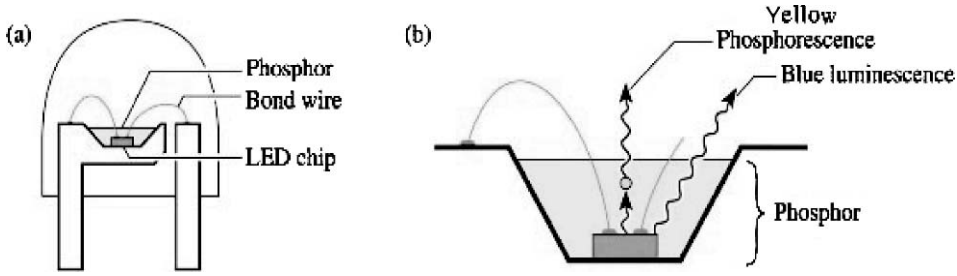


FIGURE 13.21. (a) Structure of a white LED consisting of a GaInN blue LED chip and a phosphor-containing epoxy encapsulating the semiconductor die and (b) wavelength-converting phosphorescence and blue luminescence. After Nakamura and Fasol.[9]

GaAs LEDs, InGaAsP LEDs with wavelength extended to 1.45 μm have been developed for applications in high-speed fiber-optic communications, industrial equipment, and illumination. InGaAs LEDs have also been developed for 1.3 μm fiber-optic communication and data transmission applications.

13.3.7. White LEDs

LEDs in the UV to near UV spectrum are particularly important for making white solid-state lamps using phosphors to down-convert the wavelength of the light emitter to the visible spectrum, then color mix to make a white light. High quantum efficiency is critical to making energy-efficient solid-state lamps. Figure 13.21 shows (a) the structure of a white LED consisting of a GaInN blue LED chip and a phosphor-containing epoxy encapsulating the semiconductor die and (b) wavelength converting phosphorescence and blue luminescence into white light. Figure 13.22 shows the emission spectrum of a commercial phosphor-based white GaN/InGaN LED. Super-white GaN/InGaN LED (5-mm size) lamps with luminous intensities in the range of 2,400–10,000 mcd at forward currents of 20–30 mA and forward bias of 3.4 V and power dissipation of 100–120 mW are now commercially available. This technology has the potential to revolutionize the lighting industry by enabling solid-state lamps with high efficiency and a lamp lifetime of 5–10 years. Such solid-state lamps will have efficiencies that are two to three times greater than those for incandescent bulbs. To further improve the performance of white LEDs, the peak emission wavelength of GaN/InGaN LED should shift to even shorter wavelengths below 400 nm. At these wavelengths, the light emitter better matches typical phosphor absorption bands. External quantum efficiency greater than 30% for InGaN LEDs in the UV-to-blue portion of the wavelength spectrum has been reported recently for ultrabright light-emitting and white lamp applications. White light can also be produced using a photon-recycling semiconductor LED. Figure 13.23 shows the schematic structure of a photon-recycling semiconductor (PRS) LED with one current-injected active region (1) and one optically excited active region (2). GaInN/GaN LED is used as primary source,

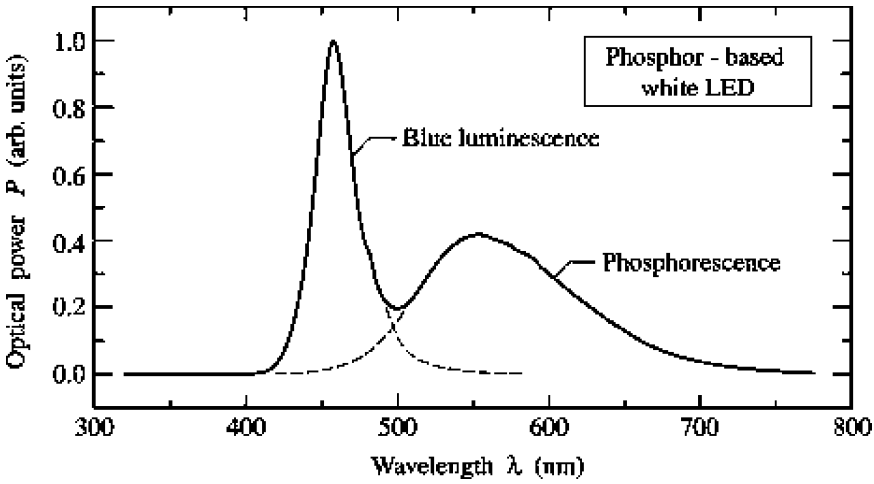


FIGURE 13.22. Emission spectrum of a commercial phosphor-based white LED manufactured by Nichia Chemical Industries Corporation.[2]

which produces the blue light, and the secondary source uses AlGaInP LED to produce the yellow light. White light is created by mixing the blue and yellow lights generated from the GaInN/GaN and AlGaInP LEDs, which are grown on opposite sides of the sapphire substrate. In addition to the two white LEDs described above, white light can also be produced by mixing three primary-color LEDs, that is, by using blue, green, and red LEDs (four-terminal device) with separate injection currents to adjust the intensity of each LED to produce pure or soft white light. Figure 13.24 shows the progress made on the luminous performance of LEDs from the mid-1960s to 2000. It is seen that a nearly three orders of magnitude increase in

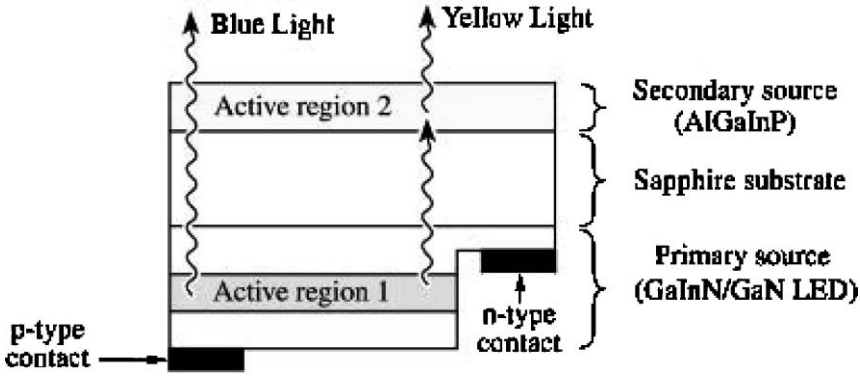


FIGURE 13.23. Schematic structure of a photon-recycling semiconductor LED with one current-injected active region (1) and one optically excited active region (2). After Gou et al., 1999.[10]

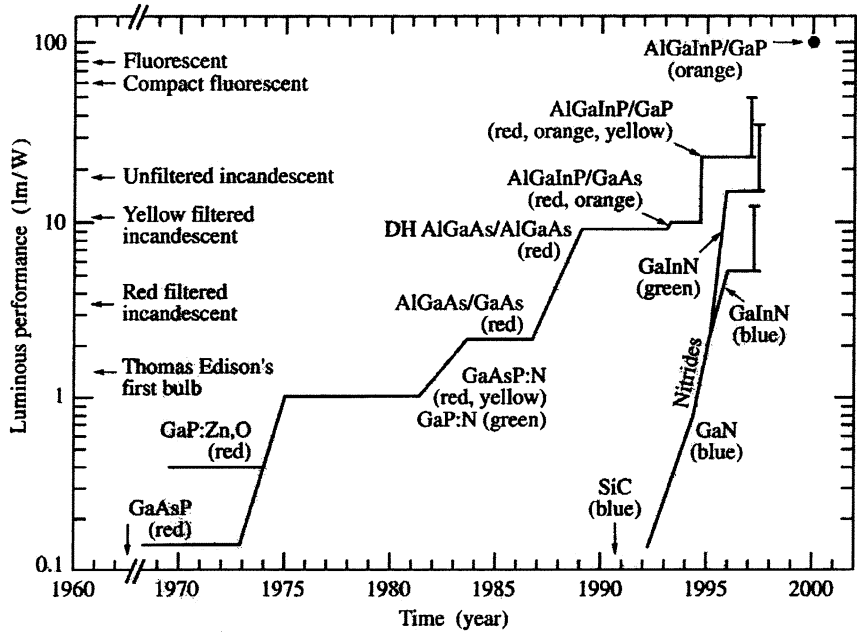


FIGURE 13.24. Luminescence performance of various LEDs developed from mid 1960 to year 2000, compared to the conventional lamps.[2]

luminous performance was achieved during this period. Table 13.2 summarizes the commercially available LEDs with emitting color, peak wavelength, and device structures. With further improvement in performance and cost reduction, LEDs are expected to play a major role in a wide range of commercial applications in lighting, display, printers, data transmission, and optical fiber communications.

High-brightness (HB) LEDs have been successfully developed in recent years and are being used in a wide variety of applications that benefit from their high visibility (even in full sunlight conditions) and full-color spectrum, including white. The spectacular growth of 2002 was led by a dramatic ramp-up in the use of HB LEDs in mobile phones, including both backlighting for full-color LCD screens and keypad backlighting. However, other applications also contributed to this vibrant market. HB LEDs are used extensively in the automotive sector, both for instrument panel lighting and for external signaling. They are the enabling components for full-color outdoor video screens used in sports stadiums, outdoor advertising, and rock concerts. Moreover, they have been widely adopted in red, green, and yellow traffic signals, as well as in highway signs and moving message panels. Illumination applications are the latest to benefit from the high efficiency and long lifetimes of HB LEDs. Based on continuing positive trends in this dynamic industry, the market for HB-LEDs is forecast to grow to $4.7 billion by 2007. HB LEDs will continue to penetrate the outdoor sign, automotive, traffic, signal, and display backlighting markets, as well as to capture an increasing variety of illumination

TABLE 13.2. Some commercial LEDs, their emitting colors (wavelengths), and device structures.

Material	Color	Wavelength (nm)	Structure
InGaN+phosphor	White	460–800	p-n junction
GaN	UV	370–390	p-n junction
GaN/SiC	Blue	430	p-n junction
InGaN	Blue	450, 473	p-n junction
ZnSe	Blue	460	p-n junction
SiC	Blue	490	p-n junction
InGaN	Turquoise	495–505	p-n junction
InGaAlP/GaAs	Pure green	562	heterojunction
GaP		555	p-n junction
InGaN	Green	525	p-n junction
InGaAlP/GaAs		574	heterojunction
GaP		567–585	
GaAsP	Yellow-green	555–575	p-n junction
AlInGaP	Yellow		
InGaAsP/GaAs		585–595	p-n junction
		590	heterojunction
GaAsP/GaP		585	heterojunction
GaP/GaP		570	
AlGaAs	Orange	605–620	p-n junction
InGaAlP/GaAs		612–620	heterojunction
GaAlP/GaP		610	heterojunction
GaP	Red	700	p-n junction
AlGaAs		660	p-n junction
InGaAlP/GaAs		623–644	heterojunction
GaAsP/GaP		635	heterojunction
GaAs	Near IR	840	p-n junction
InP	Near IR	900	p-n junction
InGaAs	IR	1.3 μm	p-n junction,

applications. From modest beginnings in the mid-1990s, the non-Japan region of Asia, including Taiwan, South Korea, and China, has become the world's largest volume producer of HB LEDs. Using advanced device manufacturing techniques based on metal-organic chemical vapor deposition (MOCVD), 23 companies in the region produced the equivalent of 13.4 billion red–orange–yellow (InGaAlP-based) LED chips and 3.4 billion blue and green (GaN-based) LED chips in 2003, representing 80% and 40% of the world totals, respectively. The HB LED market grew by 51% in 2002 to total sales of $1.84 billion.

13.3.8. RC LEDs

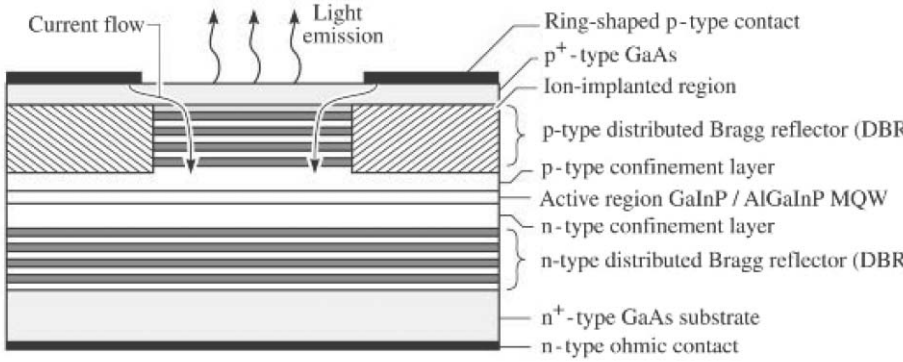
A new type of photonic device is now available in large-scale production: resonant cavity (RC) LEDs. The RC LED emitting red light at $\lambda_p = 650$ nm has been successfully developed for plastic optical fiber (POF) communication use. The RC LED is superior to its predecessors in luminous intensity, light purity, and modulation capabilities. With properties somewhere between a standard LED and

a LD, an RC LED produces intense light from quantum wells. Compared with the conventional LEDs, the RC variety produces a brighter, more directional beam with higher spectral purity and modulation speed, making it suitable for use in POF short-haul communications. It is less temperature sensitive and has a longer lifetime than competing laser light sources. RC LEDs are a key technology for the next generation of data communication via high-speed polymethyl methacrylate (PMMA) POF, which is now entering the information and entertainment or "infotainment" market in automotive and consumer applications.

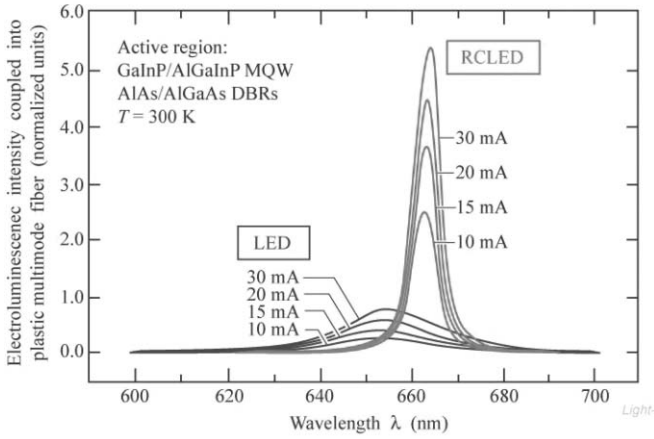
A typical RC LED consists of two distributed Bragg reflectors (DBRs) made of semiconductor materials (e.g., AlGaAs/GaAs DBR mirrors), which form a resonant cavity. An active layer is located between the DBR mirrors and includes quantum wells (e.g., GaInP/GaAs QWs) a few nanometers thick for light generation. If the thickness of the active layer between the DBR mirrors of the RC LED is chosen as an integer multiple of the half-wavelength of the emitting light, then the condition for vertical resonance is fulfilled. However, there are also resonances at off-axis angles, and the more half-wavelengths that fit between the mirrors the more off-axis resonances will be observed. The off-axis resonances are outside the extraction cone, which is defined by the angle of total internal reflection. Light in these resonances is absorbed, not emitted. Since the same amount of light is emitted in every resonance, the cavity order has to be as low as possible. The penetration depth of the light in the DBR mirrors contributes to the effective resonator length. Even if the active layer is only about the size of a wavelength, the arrangement still amounts to a multiple of half-wavelengths. External quantum efficiency as high as 12% has been reported for the red RC LED with 650 nm light emission. Commercial red RC LED has been grown on GaAs substrates using the MOCVD technique. These red RC LEDs feature GaInP QW active layer and AlGaAs/GaAs DBR mirrors and are characterized by low turn-on voltage ($V_F = 1.8$ V at $I_F = 30$ mA) and high optical output power ($P_0 > 0.5$ mW at $I_F = 30$ mA). Typical spectral width is 3 nm at a peak wavelength of 650 nm.[11,12]

RC LEDs are being designed for emerging high-speed POF applications in homes, offices, and automobiles. It transmits data at 650 nm over short-distance plastic fiber connections to link appliances such as PCs, storage devices, digital cameras, and set-top boxes. Data rates of up to 500 Mbit/s over 50 m of graded-indexed POF are supported, or 250 Mbit/s over 50 m of step-index POF. Figure 13.25a shows a schematic illustration of an RC LED formed using InGaP/AlGaInP MQW active layer and two DBRs emitting at 650 nm for POF communications, and Figure 13.25b shows spectra of light coupled into a POF from a GaInP/AlGaInP MQW RC LED and a conventional GaInP/AlGaInP LED at different driving currents. Note the narrow spectrum and higher coupled power of the RC LED.

The POF can carry more data over longer distances than Category 5 copper cable, costs less, and is easier to install compared to glass optical fiber. A number of other devices such as LEDs, edge-emitting lasers, and VCSELs also operate at 650 nm. For POF applications, conventional LEDs offer only poor coupling efficiency and operate at slower modulation speeds, while RC LEDs have a high coupling efficiency and feature a small active area, enabling higher data rate



(a)



(b)

FIGURE 13.25. (a) Schematic illustration of an RC LED emitting at 650 nm. The active region is an InGaP/AlGaInP MQW structure. Two distributed Bragg reflectors (DBRs) from the optical cavity. (b) Spectra of light coupled into a POF from a GaInP/AlGaInP MQW RC LED and a conventional GaInP/AlGaInP LED at different driving currents. Note the narrow spectrum and higher coupled power of RC LED. After Streubel et al, 1998; Whitaker, 1999.[11,12]

operation. RC LEDs are much cheaper than edge-emitting lasers and are less temperature sensitive and considerably more reliable than currently available red VCSELs. RC LEDs exhibit narrow emission line width comparable to that of LDs. This narrower emission spectrum minimizes chromatic dispersion and attenuation and allows transmission over longer lengths of POF. Unlike glass fiber, there is also no need for special alignment equipment to join the plastic fiber and emitter, and hence greatly reduces the cost of installation.

13.4. Semiconductor LDs

13.4.1. Introduction

The first coherent light source became available when Maiman introduced a pulsed solid-state ruby laser in 1960. Since then a wide variety of lasers including gas, solid state, semiconductor, and dye lasers have been developed. Today, semiconductor laser diodes (LDs) offer coherent light sources with wavelengths extending from the extremely short wavelength (i.e., X-ray) to long-wavelength infrared ($>10\mu m$) spectral regime. Semiconductor lasers were first reported by IBM, General Electric, and MIT's Lincoln Laboratory in 1962. These p-n junction LDs diodes were fabricated from $Al_xGa_{1-x}As$ alloys on the GaAs substrates using the liquid phase epitaxy (LPE) technique, and emitted coherent light in the 600–900 nm wavelength range for optical communication use.

The advances in III-V semiconductor materials processing using MBE, MOCVD, and chemical beam epitaxy (CBE) have allowed the creation of various semiconductor lasers that can operate in wavelengths from the UV, visible, and near-IR, to the mid-infrared spectral range. These advanced growth techniques allow greater control in the thickness of epitaxial layers and types of materials used, resulting in a wide variety of heterostructures including quantum well (QW), multiple quantum well (MQW), strained MQW, and quantum cascade lasers.

A semiconductor LD differs from conventional solid-state lasers or gas lasers in several aspects: (i) an LD is extremely small, (ii) it exhibits high-power conversion efficiency, (iii) it can be pumped directly by an electric current, and (iv) the intensity of output light can be easily modulated by varying the forward current of the LD. Furthermore, an LD is usually operating at a much lower power level than that of a solid-state laser or gas laser. Therefore, an LD can be used as a portable and easily controlled coherent radiation source. LDs play an important role in a wide range of applications such as coherent light sources, optical communications and data transmission, optical computing, optical displays, CD, DVD, laser printers, and optoelectronic integrated circuits (OEICs). In this section, the basic device physics and structures, operation principles, and electrical and optical characteristics of an LD are described.

13.4.2. Population Inversion

An LD can be formed by a heavily doped p^+-n^+ junction structure using a direct-band-gap semiconductor. The dopant density in both regions of the diode is usually greater than 10^{19} cm^{-3}. When a large forward-bias voltage is applied to the LD, a state of population inversion (i.e., the conduction band states are filled with electrons and the valence band states are empty) occurs near a narrow region of the p-n junction. Under this condition, radiative recombination takes place between electrons and holes in a narrow population inversion region near the junction, and lasing action is followed if a resonant cavity is provided and the oscillation

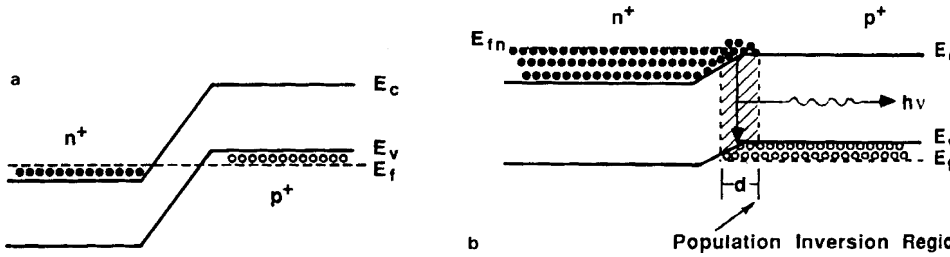


FIGURE 13.26. Energy band diagrams for a heavily doped p-n junction laser diode (a) in thermal equilibrium ($V_a = 0$) and (b) under large forward bias voltage ($V_a \gg 0$).

condition is satisfied. It should be noted that the radiative recombination is not confined merely to the conduction-valence band transition. In fact, transitions from impurity band states have also been used in many LDs.

Figure 13.26 shows the energy band diagrams of an LD (a) in thermal equilibrium and (b) under large forward-bias and population inversion conditions. Since the total rate of radiative recombination is directly proportional to the product of electron and hole densities available in the conduction and valence bands, the radiative recombination is most intense in a narrow region near the metallurgical junction of the LD, as shown in Figure 13.26b.

Population inversion in a p-n junction LD may be obtained by injection (or pumping) of minority carriers under a sufficiently large forward-bias voltage. To understand how the population inversion condition is accomplished, consider a GaAs LD. The applied forward-bias voltage V_a is chosen such that

$$V_a > \frac{h\nu}{q}, \tag{13.29}$$

where $h\nu = E_1 - E_2$ is the emitted photon energy, E_1 is the electron energy in the conduction band, and E_2 is the hole energy in the valence band. As shown in Figure 13.26b, the distribution functions for electrons in the conduction band and the valence band under a forward-bias condition are given, respectively, by

$$f_c(E_1) = \frac{1}{1 + \exp(E_1 - E_{fn})/k_B T}, \tag{13.30}$$

$$f_v(E_2) = \frac{1}{1 + \exp(E_2 - E_{fp})k_B T}, \tag{13.31}$$

where $E_2 = E_1 - h\nu$, E_{fn} and E_{fp} are the quasi-Fermi levels for electrons and holes, respectively. Now consider the rate of stimulated emission at a frequency ν due to transition from energy state E_1 in the conduction band to energy state E_2 in the valence band. The rate of stimulated emission is proportional to the product of the density of occupied states in the conduction band, $g_c(E) f_c(E)$, and the density of unoccupied states in the valence band, $g_v(E)[1 - f_v(E)]$. Therefore, the total stimulated emission rate is obtained by integrating over all the energy states in the

population inversion region. This can be expressed by

$$W_{\text{emission}} = \int P_{\text{cv}} g_{\text{c}} g_{\text{v}} f_{\text{c}} (1 - f_{\text{v}})\, \mathrm{d}E, \tag{13.32}$$

$$W_{\text{absorption}} = \int P_{\text{vc}} g_{\text{c}} g_{\text{v}} (1 - f_{\text{c}}) f_{\text{v}}\, \mathrm{d}E. \tag{13.33}$$

In (13.32) and (13.33) it is assumed that the rate of transition probability from the valence band to the conduction band and its inverse are equal (i.e., $P_{\text{cv}} = P_{\text{vc}}$), where P_{cv} denotes the rate of transition probability from the conduction band to the valence band and P_{vc} is the rate of transition probability from the valence band to the conduction band. The condition for lasing action to occur is that W_{emission} must be greater than $W_{\text{absorption}}$. Solving (13.30) through (13.33) one obtains

$$f_{\text{c}}(1 - f_{\text{v}}) > f_{\text{v}}(1 - f_{\text{c}}), \quad \text{or} \quad f_{\text{c}} > f_{\text{v}}. \tag{13.34}$$

Equation (13.34) implies that more states are occupied in the conduction band than in the valence band, which is the condition for population inversion. Solving (13.30), (13.31), and (13.34) yields

$$E_{\text{fn}} - E_{\text{fp}} > (E_1 - E_2) = h\nu. \tag{13.35}$$

It is noted that (13.35) is identical to (13.29) since $(E_{\text{fn}} - E_{\text{fp}})$ is equal to the applied voltage, qV_{a}. Thus, the condition for population inversion in an LD is given by either (13.29) or Eq. (13.35).

13.4.3. Oscillation Conditions

Two conditions must be met to achieve sustained oscillation in an LD. First, a resonant cavity must be provided so that photons generated via radiative recombination can make several passages within the cavity to be further amplified in the active medium before leaving the cavity. A typical resonant cavity consists of two parallel reflecting surfaces perpendicular to the junction plane of an LD. This type of resonant cavity is known as a Fabry–Perot cavity or an interferometer. The air–semiconductor interface boundary may serve adequately as a reflecting surface if the refractive index of the semiconductor is large enough (e.g., for GaAs, $n = 3.46$). The second requirement is that the overall amplification constant per round trip through the cavity must be positive. If R is the reflection coefficient at the two reflecting surface boundaries and L is the distance between the two reflecting boundaries of the cavity, then for each trip between the boundaries the radiation power density is reduced by a factor of $2R$ at the interface boundaries. If the gain through stimulated emission is designated by $g(\nu)$, and the loss in the laser medium due to free-carrier absorption and defect-center scattering is denoted by l, then the condition for sustained oscillation can be expressed as

$$\Gamma g(\nu) = l + \left(\frac{1}{L}\right) \ln(1/R), \tag{13.36}$$

where Γ is the carrier confinement factor, L is the length of the laser cavity, $g(\nu) = \alpha_d(f_c - f_v)$ is the gain factor due to stimulated emission, and α_d is the absorption coefficient for direct transition.

A physical interpretation of (13.36) is given as follows: In laser operation, the condition for a complete population inversion is that $f_c = 1$ and $f_v = 0$. In general, there exist three distinct regimes in an LD under different forward-bias conditions. When the nonequilibrium condition is established by current injection (i.e., $V_a > 0$), spontaneous emission occurs. As the injection current increases, a condition for population inversion (i.e., for $f_c > f_v$) is reached near the physical junction (see Figure 13.26b), which marks the beginning of stimulated emission, and the gain factor $g(\upsilon)$ becomes positive. Upon further increase of the injection current, the difference between f_c and f_v widens and hence $g(\upsilon)$ increases. When the threshold condition is met, the LD undergoes an oscillating mode of operation, resulting in the emission of coherent light from the LD. As the operating temperature is lowered, the difference between f_c and f_v will increase for the same amount of injection current. Therefore, the threshold current will decrease with lowering the diode temperature, as is evident from the experimental results of GaAs and other LDs. Figure 13.27a shows the experimental curves of emission intensity versus diode current for a GaAs LD operating at 4.2 and 77 K, and Figure 13.27b illustrates the emission intensity versus wavelength at 77 and 4.2 K for conditions below and above the threshold current of the same GaAs LD.

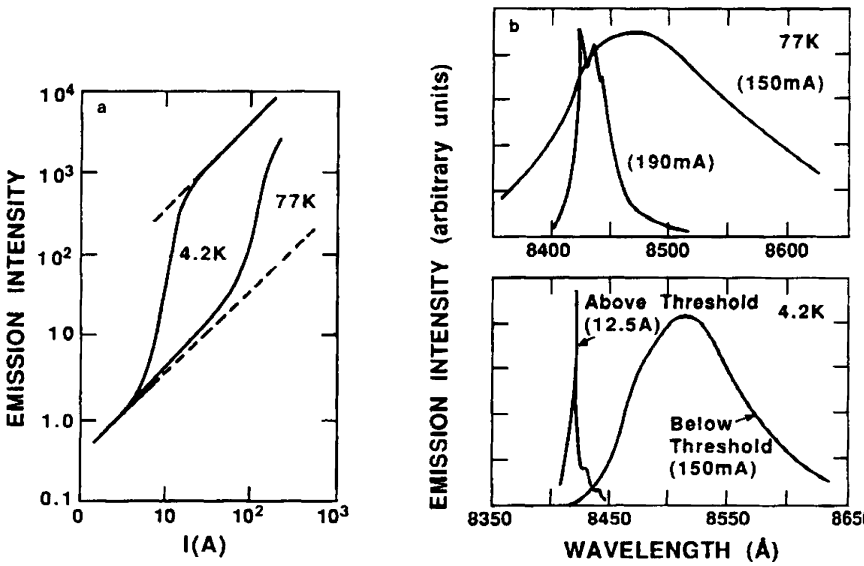


FIGURE 13.27. (a) Relative emission intensity versus diode current for a GaAs laser diode at 4.2 K and 77 K and (b) emission peak intensity versus wavelength before and after reaching threshold oscillation conditions. After Kressel and Butler,[13] by permission.

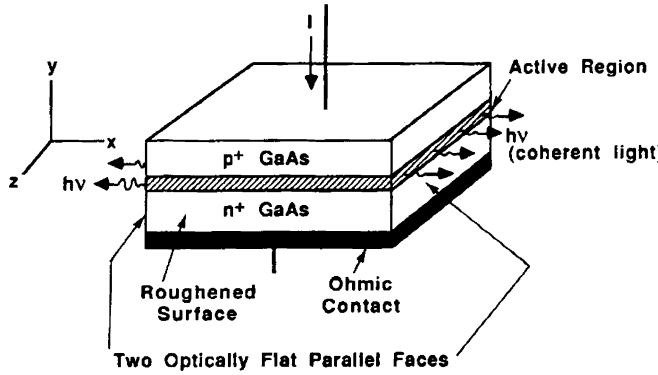


FIGURE 13.28. Structure of a GaAs p-n junction laser diode with a Fabry–Perot cavity. The optically flat parallel faces serve as the optical cavity (with length equal to L) along the x-direction, while the other two faces are roughened to prevent positive optical feedback along the z-direction.

The results reveal that a significant improvement in LD performance such as lowering the threshold current and reducing the emission line width can be obtained at 4.2 K.

13.4.4. Threshold Current Density

Figure 13.28 shows the device structure of a GaAs LD, which is used for analyzing the lasing threshold current density. The LD is constructed using a Fabry–Perot cavity with a pair of parallel planes (left and right sides) cleaved or polished to serve as end mirrors. The two remaining sides (front and back) of the LD are roughened to reduce emission of photons from both faces.

To facilitate analysis, it is assumed that the junction area is equal to A, the thickness of population inversion region is d, and the distance between the two end mirrors is equal to L. For simplicity, it is further assumed that the reflection coefficients R_1 and R_2 at the two parallel plane mirrors are equal. At $T = 0$ K, the population inversion condition requires that in the population inversion region, N_2 be equal to n (where n is the electron density in the conduction band) and N_1 be zero (where N_1 is the electron density in the valence band). Thus, the total number of conduction electrons in the population inversion region is equal to N_2Ad. If the conduction electron making a downward transition from the conduction band to the valence band has a lifetime of τ_2 and with each such decay requiring the injection of one electron into the junction region, then the total current flow in the LD needed to maintain the conduction band population density of N_2 is given by

$$I = \frac{qN_2Ad}{\tau_2} \quad \text{or} \quad J = \frac{qnd}{\tau_2}. \tag{13.37}$$

The inverted population density $N_2(= n/A)$ is related to the lasing frequency ν_0, linewidth $\Delta\nu_0$, and cavity decay time constant τ_c by

$$N_2 = \left(\frac{4\pi^2}{3}\right)\left(\frac{\Delta\nu_0}{\nu_0}\right)\left(\frac{\tau_r}{\tau_c}\right)\left(\frac{n_0}{\lambda}\right)^3, \tag{13.38}$$

where τ_r is the total lifetime due to radiative and nonradiative recombination processes. The cavity decay time τ_c is given by

$$\tau_c = \frac{(n_0/2cL)}{(2/L + \ln l/R)}, \tag{13.39}$$

where c is the speed of light, n_0 is the index of refraction for the laser material, and l is the loss associated with free-carrier absorption and scattering events in the cavity media. Now solving (13.37)–(13.39), one obtains the threshold current density J_{th}, which reads

$$J_{th} = \frac{I_{th}}{A} = (4\pi^2/3)(2qcdL/\eta)(\Delta\nu_0/\nu_0)\left(n_0^2/\lambda_0^3\right)[2lL + \ln(R^{-1})], \tag{13.40}$$

where $\eta = \tau_2/\tau_r$ is the quantum efficiency of LD, which measures the ratio of radiative lifetime to total decay lifetime by both the radiative and nonradiative recombination processes.

As an example, calculations of the threshold current density J_{th} for a p-n junction LD using (13.40) are as follows. The physical parameters for a GaAs LD operating at very low temperatures are quantum yield, $\eta = 1$ (i.e., $\tau_r = \tau_2$); index of refraction for GaAs, $n_0 = 3.46$; linewidth, $\Delta\lambda = 20$ nm; laser wavelength, $\lambda = 840$ nm; cavity length, $L = 0.3$ mm; junction depth, $d = 10^{-4}$ cm; reflection coefficient, $R = 0.32$; junction area, $A = 3x10^{-4}$ cm^2; and losses, $l = 0$. Substituting the values of these physical parameters into (13.40), one obtains a threshold current density $J_{th} = 120$ A/cm^2 or $I_{th} = 36$ mA. Experimental data for a GaAs LD at 4.2 K agree well with this calculation. It is noted, however, that for $T > 60$ K the threshold current density J_{th} for lasing is found to increase with T^3 due to the increase of absorption in the bulk GaAs near the junction region. This in turn increases the density of electrons in the valence band states. The measured threshold current at room temperature is about two orders of magnitude higher than the predicted value given above.

13.5. Laser Diode Materials and Technologies

13.5.1. GaAs-Based LDs

The structure of a typical GaAs LD with a Fabry–Perot resonant cavity is illustrated in Figure 13.28. To obtain the population inversion condition in the active region of the p-n junction LD, both p and n regions are heavily doped (i.e., the doping densities in both regions are greater than 10^{19} cm^{-3}). A pair of parallel planes are cleaved and polished perpendicular to the junction to act as an optical resonance

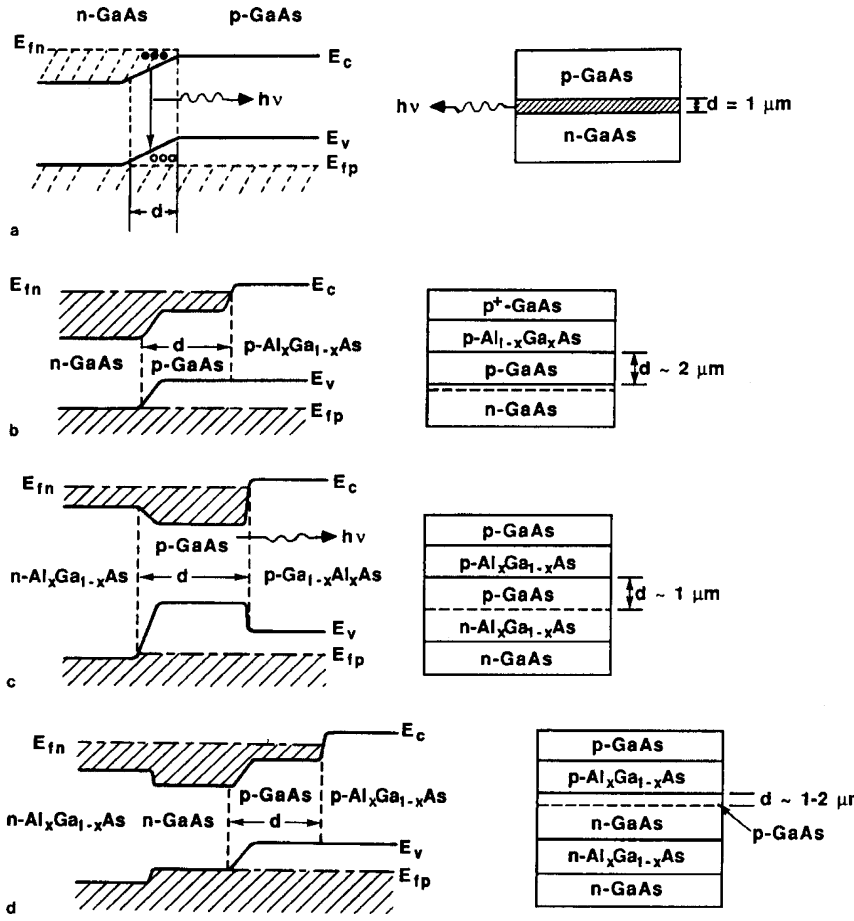


FIGURE 13.29. Multilayer heterostructures for carrier confinement in an injection laser diode (LD): (a) homojunction GaAs LD, (b) single heterostructure (SH) LD, (c) double heterostructure (DH) LD, and (d) large optical cavity (LOC) DH LD.

cavity, while the two remaining sides of the LD are roughened to eliminate possible lasing in directions other than the two main parallel mirror planes.

Several types of conventional GaAs LD structures have been reported in the literature. These include (a) epitaxial homostructure, (b) single heterostructure (SH), (c) double heterostructure (DH), and (d) large optical cavity (LOC) DH LD structure. The schematic diagrams for these structures and their corresponding energy band diagrams are shown in Figures 13.29a through 13.29d, respectively. The GaAs p-n homojunction LD structure becomes obsolete due to its high lasing threshold current density and poor confinement of both charge carriers and photons in the junction region. These shortcomings can be partially corrected in the SH LD structure, in which an AlGaAs layer with a significantly different refractive

index from that of GaAs is incorporated into the GaAs LD structure. Furthermore, the inherent energy band gap difference in both materials confines the carriers to one side (i.e., the p-GaAs side) of the junction. Significant improvement can be made if the injected carriers and emitted light are kept near the active region of the p-n junction. These techniques are known as carrier and optical confinements, and they can be achieved using the DH LD shown in Figure 13.29c. The basic device structure consists of a thin p-GaAs active layer sandwiched between the p-AlGaAs and n-AlGaAs cladding layers. Electrons injected into p-GaAs under forward-bias conditions are prevented from reaching the p-AlGaAs layer by the conduction band discontinuity. For example, the potential barrier (ΔE_c) for the $Al_{0.3}Ga_{0.7}As$ layer is equal to 0.28 eV, which povides excellent carrier confinement. The optical confinement of the DH LD is due to the difference in the refractive index of GaAs and AlGaAs. Although the refractive index of GaAs is only 5% larger than that of AlGaAs, the optical confinement is excellent for the GaAs/AlGaAs DH LD. Population inversion in a GaAs/AlGaAs DH LD can be readily reached, and radiative recombination is limited to the p-GaAs active layer. Figure 13.30 shows

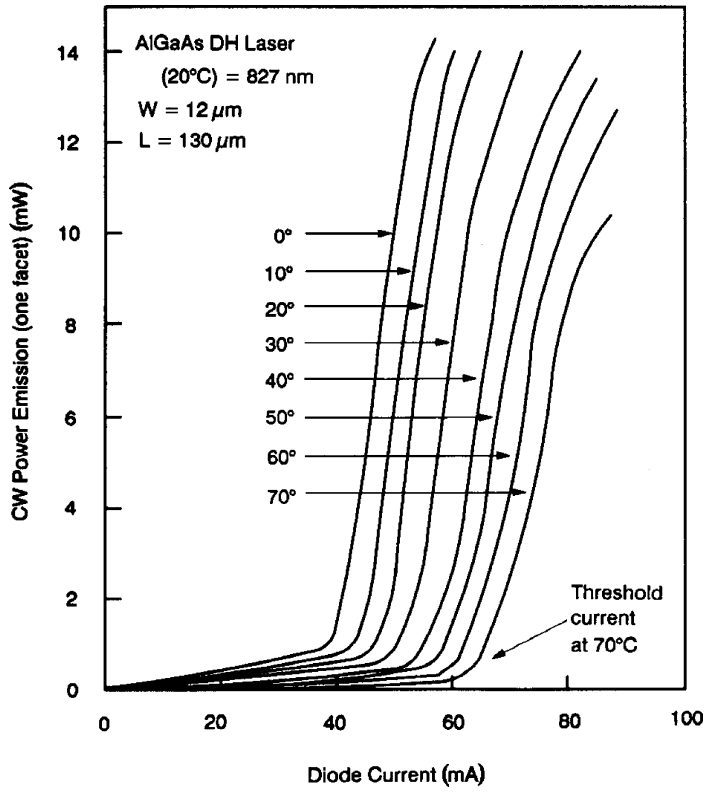


FIGURE 13.30. The CW output power versus forward current for a GaAs/AlGaAs DH LD measured at different diode temperatures. After Kressel,[14] by permission.

the CW output power as a function of current and heat-sink temperature for an oxide-defined GaAs/AlGaAs DH LD. As can be seen in this figure, the threshold current of an LD decreases with decreasing operating temperature. A differential efficiency (i.e., dP_0/dI_F) for an LD can be obtained from the slope of the output power (P_0) versus forward current (I_F) plot shown in Figure 13.30. A steeper slope represents a higher differential efficiency for the LD (i.e., a small increase in I_F will lead a large increase in LD output power above the threshold current). Figure 13.31 shows the typical optical characteristics of the CW laser spectrum for a GaAs/AlGaAs DH LD with a cavity length of 250 μm. As illustrated in this figure,

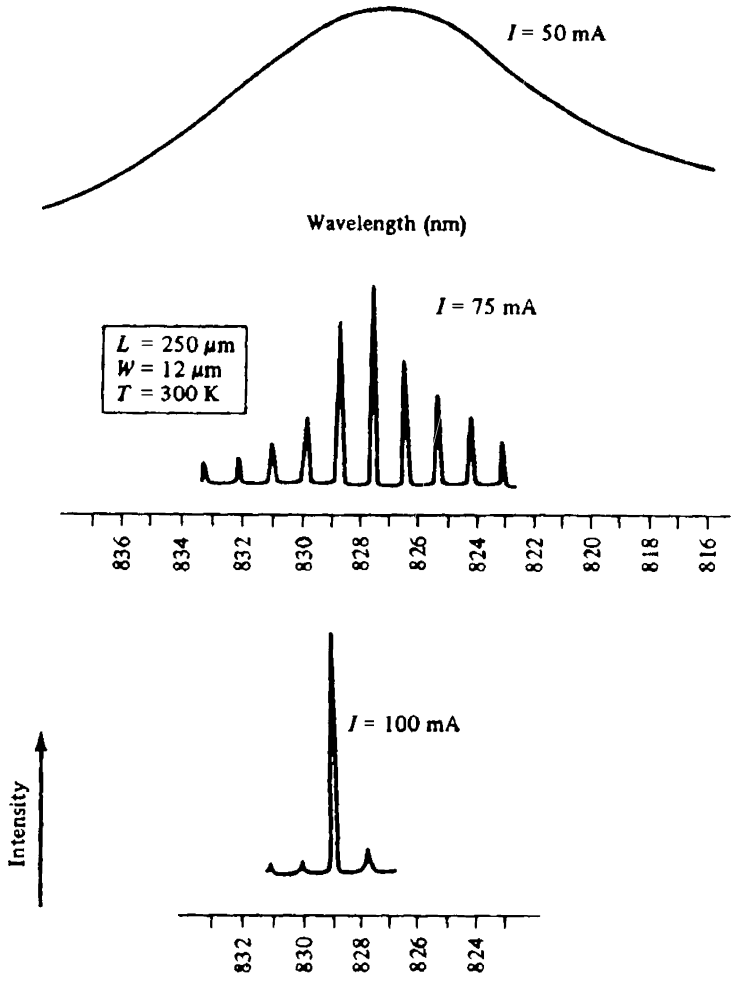


FIGURE 13.31. Lasing spectra of an oxide-defined GaAs/AlGaAs DH laser with cavity length 250 μm and different driving currents: $I_F = 50$, 75, and 100 mA. After Kressel,[14] by permission.

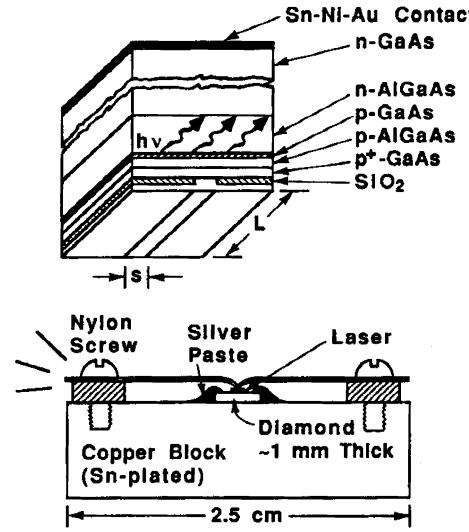


FIGURE 13.32. A GaAs/AlGaAs DH laser diode with a stripe geometry contact. The laser diode is mounted with the stripe down on a metallized diamond heat sink having five times the thermal conductivity of copper. Light emits from the p-GaAs active layer. After D'Asaro,[15] by permission.

this LD behaves like an LED at a forward current of 50 mA (or smaller) with a broad emission line width (a). As forward current increases to 75 mA, lasing action occurs with multimode emission spectra (b), and at 100 mA forward-current single-mode lasing was observed (c). The peak lasing wavelength is defined by the maximum spectral intensity in either mode.

The DH LD has by far shown the best performance in the LD operation among the bulk p-n junction LDs. Some DH LDs mounted on the diamond-II heat sinks can operate continuously at room temperature and above. A comparison of the LD structures shown in Figure 13.29 reveals that the LOC LD has the advantage of reducing the diffraction of the laser beam at the face of the p-n junction active region. This is achieved by allowing the laser beam to emerge from an opening that is much larger than the 1.2 μm openings of the other types of LDs shown in Figure 13.29.

Figure 13.32 shows a GaAs-AlGaAs DH LD with a stripe geometry contact. The substrate material is (111) or (100) oriented GaAs with a dopant density of around 1 to 4×10^{18} cm^{-3}. The first layer grown on the GaAs substrate is n-$Al_xGa_{1-x}As$, 2–5-μm thick, typically Sn-doped, with x varying between 0.2 and 0.4. The second layer is the p-GaAs active region (0.4–2-μm thick) doped with Si, and usually contains a small amount of Al, either deliberately provided or carried over from the first layer. The third layer is p-$Al_xGa_{1-x}As$ with a dopant density of around 3.8×10^{18} cm^{-3} and thickness of 1–2-μm; the composition x for this layer varies between 0.2 and 0.4. The fourth is a p^+-GaAs layer with dopant density of 3.5×10^{18} cm^{-3}. The main function for this layer is to provide better ohmic contact. The third and fourth layers are kept quite thin, and the fourth layer (i.e., the p^+ GaAs layer) is used as the main heat sink for the LD. A type-II diamond heat sink is used to improve the thermal performance of the LD and possesses thermal conductivity up to five times better than that of copper. The stripe geometry contact

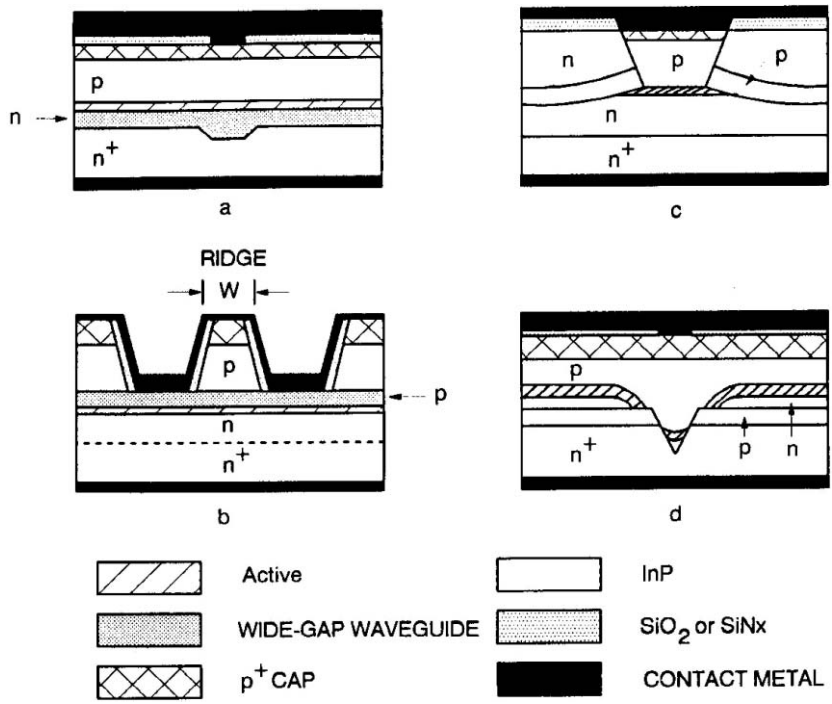


FIGURE 13.33. Some important stripe-geometry structures for a GaInAsP/InP laser diode: (a) inverted rib, (b) ridge waveguide, (c) etched-mesa buried heterostructure, and (d) channeled substrate buried structure. After Bowers and Pollack,[16] by permission.

provides a convenient way of attaining a small device area with some lateral heat flow, which allows a high continuous wavelength (CW) operating temperature. The optimum stripe width is usually between 10 and 15 μm for a typical GaAs LD. Figure 13.33 shows the cross-sectional views of several stripe-geometry InGaAsP infrared (IR) LD structures grown on InP substrates for 1.3–1.55 μm wavelength. Figure 13.33a is an inverted-rib laser structure. The wide-band-gap (e.g., AlGaAs) waveguide layer underneath the active layer (GaAs) has a riblike structure inside the optical cavity. The change in thickness produces a larger effective index of refraction in the rib region than on both sides of the rib stripe, which results in waveguiding along the rib region. In this structure, the active layer is a planar structure and the injection of current through the active layer is limited to the narrow stripe of the rib region. The ridge-waveguide structure shown in Figure 13.33b employs the same principle as the rib structure for waveguiding, which occurs underneath the ridge region. Figure 13.33c shows the etched-mesa buried heterostructure LD, which is quite different from those shown in Figures 13.33a and b. In this structure the active layer (InGaAsP) is first etched into a narrow stripe, and InP is then regrown over the active stripe. Since the InP layer has a lower refractive index and larger energy band gap than the InGaAsP active stripe, both optical and carrier confinements can be achieved using such a structure. It

is seen that current injection for structure (c) is confined by the SiO_2 on the top of the device and the reverse-bias p-n junction on both sides of the active stripe, which results in more efficient use of injected carriers and hence lowers the lasing threshold. Figure 13.32d illustrates another InGaAsP LD structure. In this laser structure a V-groove is first etched into the substrate, and then a crescent stripe of InGaAsP active layer and InP cladding layers are grown on top of the V-groove in the wide-band-gap InP by a liquid-phase epitaxy (LPE) technique to achieve optical and carrier confinements in such an LD structure.

13.5.2. Quantum-Well and Quantum-Dot Lasers

In the DH LDs discussed above, the typical thickness of the LD active layer is about 0.1 μm. If the thickness of the active layer is reduced to less than 200 Å, quantum size effects will occur. A new class of QW lasers based on such effects has been widely investigated in recent years. These new QW lasers display characteristics that are quite different from those of conventional DH LDs. In a QW laser structure, the confinement of carriers in one dimension causes quantization in the allowed energy levels along the direction perpendicular to the QW plane. The density-of-states function changes from a square-root dependence on energy in a DH LD to a steplike dependence in a QW laser. If carriers are confined in two or three dimensions (i.e., a quantum wire or quantum dot), the peak density of states of the quantum-wire or quantum-dot lasers becomes even larger at each discrete level. Such modifications in the density-of-states function of the QW lasers can greatly reduce the threshold temperature dependence, lower the threshold current, and narrow the laser line width. Of the two most throughly developed QW lasers, AlGaAs/GaAs QW lasers are found to have superior performance characteristics to those of InGaAsP QW lasers and conventional DH lasers. As an example, a very efficient AlGaAs/GaAs QW laser can be formed using a graded-index waveguide, separate-confinement heterostructure (GRIN SCH) with SQW or MQWs in the active layer, as shown in Figure 13.34, which illustrates that both optical and carrier confinement efficiencies can be greatly improved in a GRIN SCH AlGaAs/GaAs QW laser structure. The optical confinement efficiency is improved by using a parabolic refractive index profile, which can focus more optical energy to the active quantum well. The improvement in carrier collection efficiency may be attributed to the fact that change in the density of states in the graded layers is reduced. It should be noted that the GRIN-SCH laser has yielded the lowest threshold current ever reported for any semiconductor laser. The GRIN-SCH-SQW lasers and MQW ridge-waveguide lasers using GaAs/AlGaAs ($\lambda_p = 0.82$ μm), InGaAs/GaAs (0.98 μm), and InGaAsP/InP (1.3–1.55μm) material systems have been developed for applications in fiber-optic communications and fiber-optic networking in recent years. These MQW lasers show superior performance with lower threshold current, narrower and sharper emission peak, and higher differential efficiency when compared to the conventional bulk semiconductor LDs.

Quantum-dot (QD) lasers have received great attention in recent years. The main advantages of quantum-dot (QD) based lasers are low threshold current,

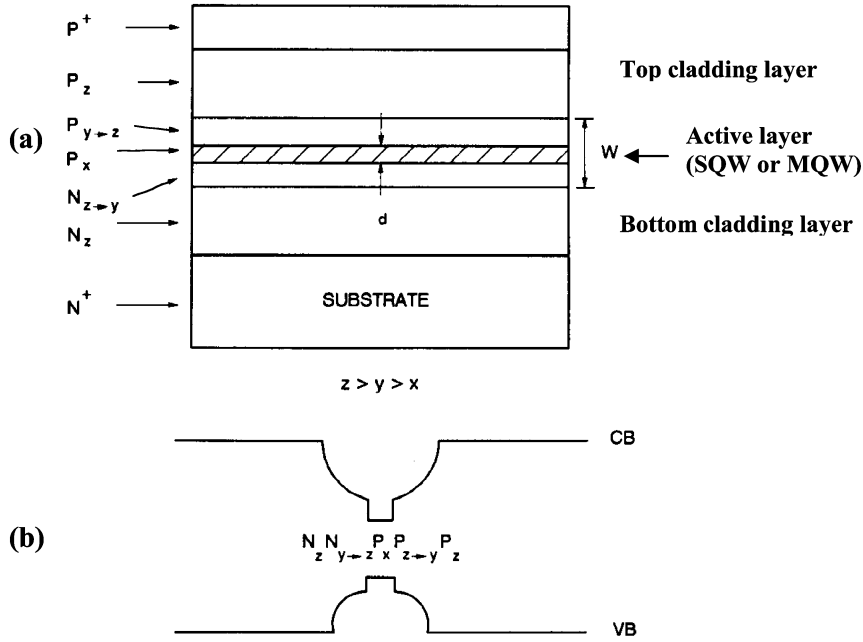


FIGURE 13.34. (a) Schematic diagram of a graded-index waveguide, separate-confinement heterostructure (GRIN-SCH) single-quantum-well (SQW) laser; (b) energy band diagram of a GRIN-SCH-SQW laser. After Tsang et al.,[17] by permission.

temperature insensitivity, excellent carrier and optical confinements, reduced linewidth, smaller size, and wavelength tunability. QD lasers fabricated from a wide variety of III-V compound semiconductors have been reported in recent years.[18–20] The growth of QDs can be achieved using the direct method or self-organized technique. In the direct method, a combination of high-resolution lithography and etching is used to form the quantum dots. In the direct method, the QW structure is formed first to provide confinement only in one dimension. By etching the QW structures to form pillars, one can provide confinement in the other two directions. The major disadvantage of this technique is the generation of nonradiative defects in the QDs, which makes this technique unsuitable for LD applications. The main requirement during the growth of QDs for semiconductor LDs is the minimization in fluctuation of dot size and position. The self-organized method using the Stranski–Krastanow (SK) growth mode can be used to grow QD LDs. The advantages of this technique include (i) complete maskless process, (ii) high uniformity in the size, location, and composition of the QDs, and (iii) that large number of QDs can be fabricated in a single step.

Some of the key features that make QD LDs popular include the following: (i) The QD structure can be easily integrated into arrays; (ii) low-threshold current density enables high-density arrays; (iii) QD-based VCSELs are readily possible with improved performance over conventional VCSELs; (iv) vertically stacked

QDs allow one to realize room-temperature lasing at a relatively low threshold current density of 90 A/cm^2; (v) QD lasers with emission wavelengths up to 1.8 μm have been fabricated using InAs/(In,Ga,Al)As QDs; (vi) CW output power of 3.5–4.0 W has been demonstrated. Another improvement predicted in QD lasers is the simultaneous reduction of threshold current density to less than 5 A/cm^2 and complete temperature insensitivity.

13.5.3. Other Semiconductor LDs

In the previous section, only GaAs-based LDs and QW lasers were discussed. The reasons that GaAs is chosen for LD fabrication are that it is a direct-band-gap material, and high-purity GaAs epilayers with relatively low defect density can be grown routinely using the MBE and MOCVD techniques. This is important from the standpoint of fabricating a reliable LD. Besides GaAs, various LDs have been fabricated using direct-band-gap compound semiconductors such as InGaN, InGaAs, InAsP, InGaAsP, ZnSe, $PbSn_xTe_{1-x}$, and $CdHg_xTe_{1-x}$, with emitting wavelengths varying from 0.4 μm to greater than 30 μm, depending on the selected alloy compositions of these materials. For example, the energy band gap for $In_{1-x}Ga_xN$ alloy may vary from 3.3 to about 1.8 eV at 300 K as x changes from 1 to 0; this corresponds to a shift of the emission peak wavelength from 0.376 μm (UV) to 0.688 μm (red).

A GaAs LD that emits coherent infrared radiation at 0.84 μm is an important IR light source for many applications including data transmission, signal processing, optical links, and optical communications. In the visible spectrum where excellent detectors are available, visible coherent light sources can be obtained from LDs fabricated on large-band-gap ternary and quaternary compound semiconductor materials such as $In_{1-x}Ga_xN$, $Al_xGa_{1-x}As$, $Ga_xIn_{1-x}As_yP_{1-y}$, and $Al_xGa_{1-x}As_yP_{1-y}$. These materials can be grown on sapphire, InP, or GaAs substrates. LDs fabricated from these materials can extend the useful wavelengths of coherent radiation from 0.4 to about 3 μm.

Although a change of energy band gap can be obtained from a list of compound semiconductors for LD applications, the choice is limited by two requirements, namely, that the material must be a direct-band-gap semiconductor and capable of forming a p-n junction on such a material system. III-V compound semiconductors such as GaN, GaAs, GaSb, InP, InSb, and II-VI compounds such as ZnS, ZnSe, and the ternary and quaternary compounds of these materials have been widely investigated for laser fabrication. III-V compound semiconductors can be grown on both n-type and p-type materials, and hence p-n junction LDs can be readily fabricated from these material systems using the MBE or MOCVD technique. Furthermore, since the band gap energies of the III-V ternary compounds such as $In_xGa_{1-x}N$, $Al_xGa_{1-x}As$, $GaAs_xP_{1-x}$, $In_xGa_{1-x}As$, and $In_xGa_{1-x}P$ can be varied by changing their alloy compositions, LDs fabricated from these materials can cover the emission wavelengths from UV to IR spectral regimes. For example, the energy band gap of the AlAs material is 2.19 eV, while the energy band gap for GaAs is 1.43 eV at 300 K. Thus, the mixed compounds of $Ga_xAl_{1-x}As$ can

produce coherent light emissions from the yellow and red to the near-infrared (IR) spectral regime. UV, blue, and blue/green lasers have been successfully developed in recent years using wide-band-gap GaN ($E_g = 3.3$ eV) based ternary compounds such as $Ga_xAl_{1-x}N$ and $Ga_xIn_{1-x}N$ material systems as well as SiC material.

As for the II-VI compound semiconductors, fabrication of LDs from these material systems is not as simple and straightforward as it is for the III-V semiconductor lasers. The self-compensation problem in II-VI semiconductors prevents possible fabrication of p-n junction LDs from several of these materials. For example, CdS can be produced only in n-type, while materials such as ZnS and ZnTe are available only in p-type. Recent progress in material growth for some of the II-VI compound semiconductors has made it possible to fabricate LDs from some of the II-VI material systems. Although it is possible to fabricate a p-n heterojunction laser structure from the II-VI materials, lattice mismatch at the interface of two different II-VI semiconductors creates further complications for efficient laser operation. In the absence of an adequate heterostructure, LDs made from II-VI materials must utilize an optical or electron-beam pumping technique. Recently, successful fabrication of a ZnSe blue LED operating at 77 and 300 K has been reported. Successful conversion of n-type ZnSe into p-type ZnSe has been achieved using nitrogen implantation followed by thermal annealing with an IR lamp, while n-type conduction can be obtained in Ga-doped ZnSe. Nitrogen doping in ZnSe enables one to produce p-type ZnSe. As a result, blue/green ZnSe p-n junction LDs have been successfully developed for commercial applications. Most ZnSe LDs reported are grown on GaAs substrates using the MOCVD or MBE technique. However, the lifetime for ZnSe LDs is still shorter than that of GaN LDs, and further improvement in material quality is needed for improving the performance of ZnSe LDs.

It should be noted that the number of semiconductor materials exhibiting laser action has been continually growing in recent years. Most of the semiconductor materials listed in Table 13.2 for LEDs have also been used in the fabrication of LDs.

13.5.4. Recent Advances in Semiconductor LDs

The worldwide market for LDs in noncommunications applications exceeded $2 billion in 2004. Driven particularly by the demand for DVD players and DVD-ROM drives, the market for LDs with wavelengths less than 1 μm grew from $966 million in 1999 to more than $2.2 billion in 2004. LDs are an enabling technology for a wide variety of consumer, computer, business, and industrial products including the familiar audio CD players and computer CD-ROM drives, DVD players and DVD-ROM drives, laser printers, laser pointers, barcode scanners, industrial material processing systems, and computer-to-plate digital printing presses.

The first significant volume application of LDs was in audio CD players, beginning in 1981. Historically, most applications of LDs have required near-IR-wavelength devices (780–850 nm); however, visible red (630–680 nm) LDs became available in 1988. Initially, visible laser diodes (VLDs) were used primarily

for barcode scanner and laser pointer applications. More recently, these LDs have been used in higher-density optical storage systems, such as DVD-ROM drives.

Another major area of applications for LDs is in the telecom laser market. Infrared LDs are a key component in local and long-distance telecom networks, as well as in cable television distribution systems. The LDs for telecommunication and fiber-optic links are fabricated from GaAs (850–980 nm), InGaAs (1.3 μm), and InGaAsP (1.55 μm) material systems for fiber-optic links. The InGaAs and InGaAsP LDs are used mainly for long-haul fiber-optic communications and data transmission, while GaAs LDs are used mainly for short-distance local area networks (LAN). The LDs are used in these systems as transmitters into the optical fiber, and also as high-power "pump" sources to drive the optical amplifiers used along long-distance routes. The performance requirements on LDs for these applications are considered exceptionally demanding, and especially for the LDs used in transcontinental cables. Demand for these products is so strong that the worldwide market for LDs for telecom applications is forecast to grow from $1.95 billion in 1999 to over $5 billion in 2004, for a compound growth rate of over 22% per year. Several semiconductor lasers developed recently for fiber-optic communication and DVD applications are described next.

13.5.4.1. Visible Laser Diodes (VLDs)

LDs continue to find new product applications as the lasing wavelength is pushed shorter into the visible spectrum. The latest generation of visible laser diodes (VLDs) operate at or near 635 nm; this wavelength, equivalent to a helium neon gas laser, is highly visible to the human eye. VLDs in the range from 635 to 685 nm are replacing the traditional HeNe laser in many commercial products for good reasons: lower cost, compact size, and superior long-term reliability. Another intrinsic benefit is that LDs are generally more suitable for battery-operated devices and other low-voltage applications. The key technology for the VLDs is based on AlGaAs and AlGaInP LDs grown on GaAs substrates by the MOCVD technique. Commercial applications for these VLDs include laser pointer, line marker, leveler, bar-code scanner, DVD, DVD-R/RW, laser printer, CD, CD-ROM, CD-R/RW, and optical communications. The emission wavelengths can vary from 635 to 850 nm depending on the alloy compositions used in the fabrication of these LDs. The AlGaInP LDs with low threshold current (20 mA) and short wavelength (635 nm) are achieved by using a strained InGaP/AlGaInP MQW active layer. A reduction in the peak emission wavelength is achieved by increasing the band gap of the QW active region. The 635-nm LD is eight times brighter than a 670-nm LD. Typical output power for such an LD is 3–5 mW CW, which is suitable for battery-powered laser pointers due to its low operating current and voltage.

For shorter-wavelength (400–470 nm) operation, blue LDs based on GaN/InGaN material systems have been developed in recent years. Nichia Corporation has demonstrated a long-lived (10,000 hours) GaN-based blue LD, and Sony, Fujitsu, Toshiba, and many other companies have aggressive research programs to develop these devices for high-density optical storage applications. There are now more

than 30 LD manufacturers around the world. The major suppliers of low-power LDs (i.e., $P_0 < 100$ mW) are Japanese companies such as Sony, Rohm, Sharp, and Matsushita. North American companies such as SDL and Opto Power have taken a leading role in the high-power LD ($P_0 > 10$ W) market. European producers, such as OSRAM Opto Semiconductor and DILAS, are gradually increasing their market share. In Asia, Samsung and several Taiwanese companies are also producing LDs. There are many applications using LDs operating in the 780–850 nm infrared spectral range, since some machine vision systems and sensors are optimized for near-infrared light sources, and GaAs LDs operating at 850 nm are widely used in short-distance LAN and optical links. High-performance InGaAsP/InP strained MQW lasers grown on InP substrates have been developed for 1.3 and 1.55 μm fiber-optic communications applications.

13.5.4.2. Vertical Cavity Surface Emitting Lasers

The vertical cavity surface emitting laser (VCSEL) is a semiconductor LD that emits light in a cylindrical beam vertically from the surface of a fabricated wafer. It offers significant advantages when compared to the edge-emitting lasers currently used in the majority of fiber-optic communications devices. VCSELs can be fabricated efficiently on a 3-inch-diameter wafer. Even more important, the ability to manufacture these lasers using standard microelectronic fabrication methods allows integration of VCSELs on board with other components without requiring prepackaging. As an enabling technology, VCSELs allow superior new systems and products to be created at a lower cost. VCSELs provide a higher coupling efficiency with an optical fiber due to the emission of a circular laser beam and the ease of the manufacture of an array, and enable error detection and characteristics measurement in a wafer state. As a result, they are emerging as a promising light source in optical communications and optical interconnections. In particular, index-guided VCSELs, in which an aperture through which current flows is confined by selective oxidation, have a very low threshold current and power consumption, high efficiency, and excellent linearity of current to light output. Therefore, the index-guided VCSELs can be applied to transceiver modules for local area communications and for optical interconnection between computers and digital displays; and some of these are under commercial development. The VCSELs can be fabricated from a list of material systems including GaN, GaAs, InP, GaSb, ZnSe, and their compounds.

VCSELs have a number of important advantages that have catapulted them to the distinctive position of being the technology of choice for a wide range of data communications products. With a low threshold current of between 1 and 6 mA, VCSELs offer very efficient power conversion. They can deliver transmission speeds between 1 and 10 Gbit/s, yet have a modulation swing of only 5–10 mA, which keeps power consumption low. The latest generation of VCSELs does not require hermetic packaging, yet typical mean lifetimes for a well-manufactured VCSEL device range from 10 to 100 years. At the same time, the circular, low-divergence output beams provided by VCSELs eliminate the need for corrective

optics in most applications. The GaAs-based VCSELs with peak emission wavelength at 850 nm are used as optical switch interconnects for very-short-reach (VSR), rack-to-rack or switch-to-aggregation device interconnects, short-reach storage-area networks (SAN), and LAN links for enterprise networks (typical distance is less than 500 m). They are solidly entrenched in the network for SAN and LAN backbones. The dominant laser communication for these optical interconnects is transceivers based on VCSELs. For longer-reach links (10 km) in optical metro/access interconnects, the InP-based VCSELs with peak emission wavelength at 1310 nm are used in these applications. The InGaAsP VCSEL arrays grown on InP substrates with emission wavelength at 1550 nm are used primarily for long-haul, core fiber-optic interconnects with distances greater than 100 km.

Currently, optical networks use fixed-wavelength lasers emitting at particular wavelengths to achieve data multiplexing and higher bandwidth. Advances in GaAs-based VCSELs have led to commercial applications in data communications at 850 nm, while InP-based VCSELs operating at 1.3 and 1.55 μm are used as sources for long-haul telecommunications.

The VCSEL devices are characterized by a short optical cavity length on the order of $n\lambda$, where λ is the laser operating wavelength and n is the refractive index of the cavity material. The cavity is formed between two highly reflecting DBR mirrors, which consist of alternate quarter-wavelength ($^1/_4\ \lambda$) low and high refractive index layers. The gain region generally consists of quantum wells positioned at the antinodes of the cavity resonance. The VCSEL's geometry is such that the optical cavity is defined normal to the wafer surface, and hence the laser is surface emitting. The surface normal emission means that subsequent process steps are used to define a circularly symmetric optical aperture. The resulting transverse mode thus produces low-divergence circular-beam profiles that are ideal for coupling into optical fibers. The very short cavity length also ensures support of only one longitudinal mode.

Figure 13.35 shows a schematic drawing of an oxide-confined top-emitting AlGaAs/GaAs-based VCSEL for ultra-high-speed 850-nm optical interconnects. The layer structure is grown by molecular beam epitaxy (MBE) on GaAs substrate. The active region is composed of three 8-nm-thick GaAs/AlGaAs QWs embedded in carrier confinement layers and bottom and top DBRs. Devices with small mesa diameters are desirable in order to reduce the low-pass filter effect of series resistance and parasitic capacitances on the modulation bandwidth. As shown in Figure 13.35, upper mesas of 20 μm diameter are formed using chemically assisted ion beam etching. After selective wet oxidation of the 30-nm-thick AlAs layer current aperture, the mesas are passivated and a second etching step gives access to the n-doped GaAs substrate on which a large-area n-contact is evaporated. The surface is planarized with polyimide and a coplanar contact layout is put on top that allows wire bonding to a transmission line or alternatively flexible testing with a microwave probe tip. Measured electrical 3-dB bandwidths easily exceed 12 GHz for small-signal modulation. For apertures below 4 μm, the VCSEL emits in a single fundamental mode closely resembling a Gaussian field profile. Highest reported output powers are in the 5 mW range. Both types of VCSELs can

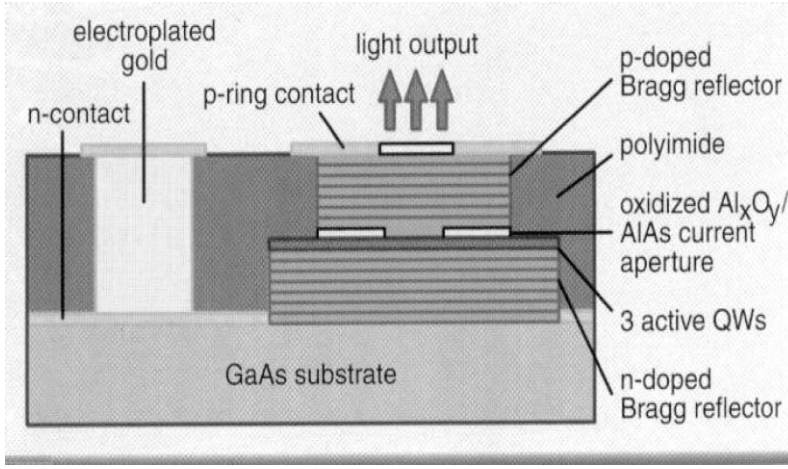


FIGURE 13.35. Schematic drawing of an oxide-confined top-emitting AlGaAs/GaAs-based VCSEL for ultra-high-speed 850-nm optical interconnects. The active region is composed of three 8-nm-thick GaAs/AlGaAs QWs embedded in carrier confinement layers and bottom and top DBRs. After mederer.[21]

be designed for 10 Gbit/s operation. For larger diameter, lasing occurs in several transverse modes. Both types of VCSELs can be designed for 10 Gbit/s operation.

VCSELs operating at telecom wavelengths of 1,310 and 1,550 nm are significantly more cost effective and easier to build than standard edge-emitting lasers used in current high-speed communications. A number of approaches have been developed to produce 1.3 and 1.55 μm VCSELs. One approach is using wafer fusion to combine the 1300 nm active region with the GaAs-based DBR mirrors and an integrated optical pump. Another approach is to use an InGaAsN active layer lattice-matched to GaAs-based DBRs. Cielo and Sandia demonstrated the first electrically pumped 1.3 μm VCSEL. The monolithic device exhibited single-mode and continuous-wave operation at 1,294 nm with an output power of 60 mW. The device is grown by a single MBE growth process, and does not involve the use of wafer fusion techniques. It is a monolithic GaAs-based structure with GaAs/AlGaAs DBR mirrors and an active region containing two QWs made from the quaternary $In_{0.35}Ga_{0.65}As_{0.983}N_{0.017}$ alloy, which is lattice-matched to GaAs at low nitrogen composition. The significant cost reduction provided by 1.3 μm VCSELs has made increased bandwidth more accessible and cost effective for the telecommunications and Internet infrastructure. Furthermore, since 1.3 μm light can be transmitted through silicon, the additional flexibility this offers for integrating the 1.3 μm VCSELs with silicon-based microsystems will have significant implications for security systems applications. The development of 1310 nm VCSEL arrays and 10 Gbit/s transceivers is causing VCSELs to expand into ultrahigh-bandwidth enterprise switching and SAN applications, as well as metro/access applications.

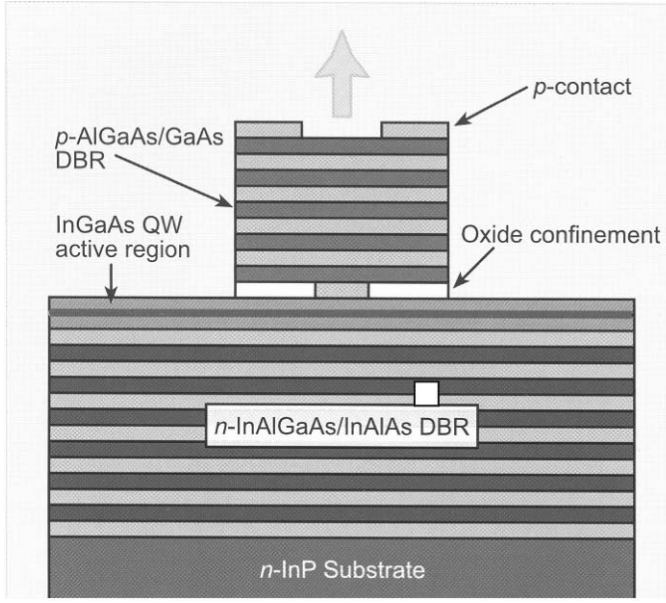


FIGURE 13.36. The schematic layer structure of a single-epitaxy 1.6 μm InP-based VCSEL developed by Bandwidth. This VCSEL is grown on InP substrate, and is composed of a lattice-matched n-InGaAlAs/InAlAs bottom DBR mirror and an InGaAs QW active region. After Whitaker.[22]

The InP-based VCSEL is capable of emitting light in the 1550 nm transmission window for fiber-optic links. Figure 13.36 shows the schematic layer structure of a single-epitaxy 1.6 μm InP-based VCSEL developed by Bandwidth9. This VCSEL device is grown on InP substrate, which is composed of a lattice-matched n-InGaAlAs/InAlAs bottom DBR mirror and an InGaAs QW active region. The key feature of the device is the inclusion of a metamorphic p-GaAs/AlGaAs top DBR mirror, which provides high reflectivity and allows direct current injection. The device emitted a single transverse mode with fixed polarization, and demonstrated CW output power of 0.45 mW at 25°C, as well as error-free transmission at 2.5 Gb/s through 50 km of single-mode fiber. The maximum output power of 0.45 mW was achieved for a device with a 9 μm aperture and a threshold voltage of 1.7 V. The minimum threshold current is 0.87 mA for a 32-μm device.

Today,1,310 and 1,550 nm VCSELs with low threshold current (10 mA) and modulation capability up to 2.5 and 1.25 Gb/s are commercially available for applications in telecommunications, fiber-optic links, cable TV, fiber channel, and ATM transceiver modules and systems. Typical high-speed 1310 nm LD devices deliver a CW output power of 5 mW at operating current of 22 mA and operating voltage of 1.15 V with a slope efficiency of 0.40 mW/mA. The 1,550 nm LD device delivers a CW output power of 5 mW at 33 mA and 1.15 V with a slope efficiency of 0.26 mW/mA.

13.5.4.3. Tunable Lasers

The high data capacity provided by dense-wavelength division multiplexing (DWDM) optical fiber links is the reason that Internet exists in its current form. The DWDM technology has been instrumental in allowing network operators to send multiple signals down a single fiber, enabling a huge growth in network resources for no extra fiber deployment. With the exponential growth of traffic, it is clear that optical network capacity installation must continue apace by scaling dramatically. Currently, optical networks employ fixed lasers emitting at particular wavelengths to achieve data multiplexing and higher bandwidth. Therefore, key to satisfying the requirements of future optical networks will be the ability to supply bandwidth on demand. These can be nicely met by tunable lasers: light sources that can be tuned over a range of wavelengths.

There are several approaches to tunable lasers on the market, including DFB laser arrays, DBR lasers, tunable VCSELs, and ECLs. Each of these laser technologies has strong points and weaknesses to be overcome. The ECL device is capable of very high spectral purity, high output power, and wide tuning range. These features make the emerging ultra-long-haul applications a natural target for this technology. Since wavelength tuning in an ECL is achieved by physically moving the mirror and grating in concert, there are some major challenges to overcome in qualifying the product and proving telecom-level reliability. Furthermore, the ECL is fairly complex to manufacture, and hence is unlikely to achieve the cost points necessary to address the metropolitan or regular long-haul applications.

Tunable VCSELs are used with a MEMS mirror to provide the wavelength tuning. There are two main types of tunable VCSELs: electrically pumped and optically pumped. The electrically pumped architecture offers wide tunability at relatively low power levels. Using direct modulation, this approach can be well suited to metro-access solutions where low power and short-dispersion limited reach can be accommodated. This technology has the potential for low-cost manufacturing; the laser's chips can be tested on wafer, so only known good dies are assembled and coupling efficiency can be quite high due to the circular nature of the beam. The power capabilities largely preclude this technology from addressing long-haul applications.

The optically pumped variant of the tunable VCSEL offers slightly greater power and is targeted more toward long-haul applications. To achieve the power necessary for long-haul applications, it is necessary to use an additional semiconductor optical amplifier (SOA) chip to boost the output power; this adds cost and complexity to the manufacturing process. It is also possible to modulate the pump laser to address the metro-access market, but the multichip assembly approach does not lend itself well to addressing the cost requirements for this space. Both of these VCSEL approaches rely on small mechanical movement to provide the tuning function. The same reliability and qualification challenges outlined for the ECLs also apply here.

Figure 13.37 shows a cross-sectional view of a grating coupled sample reflector (GCSR) tunable laser developed by Attitun. This GCSR laser employs four separate regions fabricated from a monolithic InP structure with an InGaAsP quantum-well

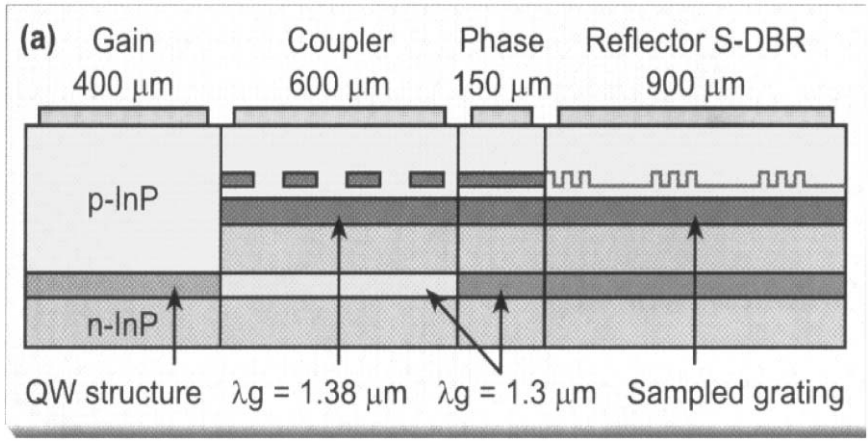


FIGURE 13.37. A grating coupled sample reflector (GCSR) tunable laser that uses four separately controlled regions fabricated in the monolithic InP structure with an InGaAsP QW quaternary active region. The design allows tuning across a 100-nm range. After Plasto.[23]

(QW) active region, each section determines the optical power of the device, while the second vertical coupler region acts as a coarse tuning of up to 100 nm. This enables the choice of one of the ten or more wavelengths available from the peaks reflected by the sampled grating Bragg reflector (S-DBR). Once the wavelength range is selected, the grating region tunes the current across a 4-nm band in the same way as that of a DBR, and a phase region gives the fine control at the GHz level. The control of power and wavelength of this device is managed at the module level and the power output can be at 10 dBm or higher. Tunable lasers will be extensively used for bandwidth provisioning and for fast optical switching in the new generation of optical networking.

Problems

13.1. If band-to-band radiative recombination is responsible for the emission of photons, what color of light may be expected from an LED made from the following materials and explain why: GaAs, GaN, $Ga_{0.3}Al_{0.7}As$, $GaAs_{0.6}P_{0.4}$, $In_{0.53}Ga_{0.47}As$, $In_{0.5}Al_{0.5}P$, and $In_{0.5}Ga_{0.5}P$? If the band gap energy for GaAs is $E_g = 1.42$ eV; $E_g = 3.5$ eV for GaN, 2.26 eV for GaP, 2.45 eV for AlP, 2.19 eV for AlAs, and 1.35 eV for InP, use linear extrapolation to find the band gap energy variation with alloy composition x (i.e., $Al_xGa_{1-x}As$) for the ternary compounds listed above and find their corresponding emission wavelengths.

13.2. (a) Draw a cross-sectional view of the layer structure of an AlGaInP DH LED consisting of p-GaP window-layer, p-AlInP cladding layer ($N_A = 5 \times 10^{17}$ cm^{-3}), p-AlGaInP active layer ($N_A = 5 \times 10^{16}$ cm^{-3}),

n-AlInP cladding layer, and AlAs/AlGaAs DRB grown on n-GaAs absorbing substrate; this DH LED emits light at 565 nm (yellow). (b) Explain how the light emission, carrier, and optical confinements are achieved in this DH LED and (c) if yellow light is emitted from this LED, what would happen if the GaAs substrate were replaced by GaP substrate.

13.3. A ZnO-doped GaP LED is used as a red LED. The Zn impurity is an acceptor level with $E_A = E_v + 0.04$ eV, an oxygen (O) impurity is a deep-donor level with $E_D = E_c - 0.43$ eV, and the band gap energy for GaP is $E_g = 2.26$ eV. What is the peak emission wavelength for this LED if the light emission is due to the radiative recombination via ZnO impurity centers.

13.4. The refractive indices of GaN and GaAs LEDs are given by 2.5 and 3.4, respectively. (a) Calculate the critical angle of total internal reflection for GaAs, and GaN, using (13.25). (b) If the fraction of light power that escapes from the semiconductor is given by

$$\frac{P_{\text{escape}}}{P_{\text{source}}} = \frac{n_{\text{air}}^2}{4n_s^2}, \tag{1}$$

where P_{source} is the light power emitted from the semiconductor and n_{air} and n_s are refractive indices of air and semiconductor, respectively, calculate the fraction of light power that can escape from a planar GaAs and GaN LED structure. (Answer: $\phi_c = 17.1^\circ$ for GaAs, 23.6° for GaN; fraction of light escape: 2.21% for GaAs, and 4.18% for GaN.)

13.5. Using (13.23) through (13.26) and the latest LED growth technologies, design an AlGaInP red LED that could yield an external quantum efficiency of 50% or higher.

13.6. High-power white LEDs are in great demand for lighting use in homes and offices. Describe two or three approaches that could be used to produce high-efficiency white LED solid-state lamps.

13.7. What is the optical power density (per unit volume) generated in a typical GaAs injection LD at threshold for $T = 4$, 77, and 300 K? If 1% of this power density is absorbed and converted into heat in the junction volume, what would be the rate of temperature rise in this GaAs LD?

13.8. Draw the energy band diagram of a degenerate GaAs p^+-n^+ junction LD with a sufficiently large forward bias to cause population inversion in a narrow region of the junction, and explain the operation principle of this LD.

13.9. Design a single-mode double-heterojunction LD with carrier and optical confinements using InP-based lattice-matched materials for light emission at wavelength 1.55 μm.

13.10. Using (13.40), calculate the threshold current density J_{th} for a GaAs p-n junction LD. The physical parameters for this LD operating at very low temperatures (e.g., at 4.2 K) are quantum yield, $\eta = 1$ (i.e., $\tau_r = \tau_2$); index of refraction for GaAs, $n_0 = 3.46$; linewidth, $\Delta\lambda = 20$ nm; laser wavelength, $\lambda = 840$ nm; cavity length, $L = 0.3$ mm; junction depth, $d = 10^{-4}$

cm; reflection coefficient, $R = 0.32$; junction area, $A = 3 \times 10^{-4}$ cm^2; and losses, $l = 0$. (Answer: $J_{\text{th}} = 120$ A/cm^2 or $I_{\text{th}} = 36$ mA at 4.2K.)

13.11. The power output (P_0) of an LD as a function of injection current density can be described by

$$P_0 = A(J - J_{\text{th}}) \left(\frac{\eta_i h\nu}{q} \right) \frac{(1/2L)\ln(R)}{[l + (1/2L)\ln(R)]}, \tag{1}$$

where J_{th} is the threshold current density, η is the internal quantum efficiency, $R = R_1 = R_2$ is the reflectivity at two laser facets, l is the loss in the laser medium due to free carrier absorption and scattering by defect centers, and $h\nu$ is the emission photon energy. Suppose the temperature dependence of threshold current density is given by

$$J_{\text{th}} = J_{\text{th0}} \exp(T/T_0), \tag{2}$$

where J_{th0} is the threshold current density at $T = 0$ K and T_0 is the characteristic temperature of LD material.

Using (1) and (2), plot the power output (P_0) of an AlGaAs/GaAs DH laser versus diode current for $T = 0$, 10, 20, 30, 40, 50, 60, and 70 °C (see Figure 13.30). Given: $T_0 = 160$ K for an AlGaAs/GaAs LD.

(Note that (1) can also be used to find the power conversion efficiency of an LD, which is defined by $\eta_p = P_0/V_a AJ$.)

References

1. H. F. Ivey, *IEEE J. Quantum Electron.* **QE-2**, 713 (1966).
2. E. F. Schubert, *Light Emitting Diodes*, Cambridge University Press, Cambridge, UK (2003).
3. M. R. Krames, M. Ochiai-Holcomb, G. E. Hofler, C. Carter-Coman, E. I. Chen, I.-H. Tan, P. Grillot, N. F. Gardner, H. C. Chui, J.-W. Huang, S. A. Stockman, F. A. Kish, and M. G. Craford, *Appl. Phys. Lett.* **75**, 2365 (1999).
4. W. N. Carr, *IEEE Trans. Electron Dev.* **ED-12**, 531 (1965).
5. S. V. Galginaitis, *J. Appl. Phys.* **36**, 460 (1965).
6. C. H. Chen, S. A. Stockman, M. J. Peanasky, and C. P. Kuo, in *"High Brightness Light Emitting Diodes,"* edited by G. B. Stringfellow and M. G. Cradford, *Semiconductors and Semimetals*, Vol. **48**, Academic Press, San Diago (1997).
7. M. Holcomb, P. Grillot, G. Hfler, M. Krames and S. Stockman, *Compound Semiconductor* Magazine, April issue (2001).
8. M. di Forte-Poisson, *Compound Semiconductor*, **Vol. 7** (3), pp. 70 (2001).
9. S. Nakamura and G. Fosol, "*The blue laser diode,"* Springer, Berlin (1999).
10. X. Gou, J. W. Graff, and E. F. Schubert, *IEDM Technical Digest*, **IEDM-99**, 600 (1999).
11. K. Streubel and R. Stevens, *Electronics Lett.* **34**, 1862 (1998).
12. T. Whitaker, *Compound Semiconductors* **5**(4), 32 May (1999).
13. H. Kressel and J. K. Butler, *Semiconductor Lasers and Heterojunction LEDs*, Academic Press, New York (1977).
14. H. Kressel, in: *Fundamentals of Optical Fiber Communications* (M. K. Barnoski, ed.), 2nd ed., Chap. 4, Academic Press, New York (1981).
15. L. A. D'Asaro, *J. Luminescence*. **7**, 310 (1973).

16. J. E. Bowers and M. A. Pollack, in: *Optical Fiber Telecommunications II* (S. E. Miller and I. P. Kaminow, eds.), Academic Press, New York (1988).
17. W. T. Tsang, R. A. Logan, and J. P. Van der Ziel, *Appl. Phys. Lett.*, **34**, 644 (1979).
18. J. Y. Tsao, *Solid State Lighting*, IEEE Circuits & Devices Magazine, May/June, pp. 28–37 (2004).
19. N. Nikolai, et al., *Quantum Dot Heterostructure Lasers,* IEEE J. of Selected Topics in Quantum Electronics, **vol. 6** (3), pp. 439–451, May/June (2000).
20. A. Yasuhiko, *Progress in GaN based Quantum Dots for Optoelectronic Applications*, IEEE J. of Selected Topics in Quantum Electronics, **vol. 8** (4), pp. 823–833, July/August (2002).
21. F. Mederer, R. Michalzik, and K.J. Ebeling, *Compound Semiconductor*, **6**(6), pp. 60–64 (2000).
22. T. Whitaker, *Compound Semiconductor*, 6(5), pp. 65 (2000).
23. R. Plastow, *Compound Semiconductor*, **6**(6), pp. 58 (2000).

Bibliography

A. A. Bergh and P. J. Dean, *Light-Emitting Diodes*, Clarendon Press, Oxford (1976).
W. F. Brinkman, T. L. Koch, D. V. Lang, and D. P. Witt, *Bell Labs Tech. J.* **5**, 150–167 (2000).
H. C. Casey Jr. and M. B. Panish, *Heterojunction Lasers*, Academic Press, New York (1978).
L. A. Coldren and S. W. Corzine, *Diode Lasers and Photonic Integrated Circuits*, Wiley, New York (1995).
K. Gillessen and W. Shairer, *Light Emitting Diodes*, Prentice-Hall, New York (1987).
S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*, Prentice Hall, New Jersey (2001).
H. Kressel, in: *Fundamentals of Optical Fiber Communications* (M. K. Barnoski, ed.), 2nd ed., Academic Press, New York (1981), Ch. 4.
H. Kroemer, *Proc. IEEE* **51**, 1782 (1963).
C. H. Lee, *Picosecond Optoelectronic Devices*, Academic Press, New York (1984).
B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York (1991).
B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*, 5th ed., Prentice Hall, New York (2001).
S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1981).
G. H. B. Thompson, *Physics of Semiconductor Laser Devices*, Wiley, New York (1980).
V. M. Ustinov, A. E. Zhokov, A. Yu, Egorov, and Nikolai A. Maleev, *Quantum Dot Lasers*, 1st ed., Oxford University Press, London, UK (2003).

14
Bipolar Junction Transistors

14.1. Introduction

The invention of germanium alloy bipolar junction transistors (BJTs) by Bardeen, Brattain, and Shockley in 1948 has revolutionized the electronics industry. The BJT device is considered one of the most important electronic components used in modern integrated circuit (IC) chips for computers, communications and power systems, and in many other digital and analog electronic circuit applications. The subsequent developments of silicon BJTs, metal-oxide-semiconductor field-effect transistors (MOSFETs), and ICs based on BJTs and MOSFET have changed the landscape of the entire electronics industry. As a result, silicon BJTs and FETs have replaced bulky vacuum tubes for various electronic circuits, computers, microwave, and power systems applications. Furthermore, advances in silicon-processing technologies such as the development of optical and electron-beam (E-beam) lithographies, new metallization and etching techniques, as well as ion-implantation enable the fabrication of high-performance silicon BJTs with submicron geometries for very large scale integrated circuit (VLSIC) applications. Recent development of new Si/Si-Ge heterojunction bipolar transistors (HBTs) grown by molecular beam epitaxy (MBE) and metal-organic-chemical vapor deposition (MOCVD) techniques on silicon substrates offer even higher speed performance for next-generation supercomputer applications.

Conventional n^+-p-n or p^+-n-p BJTs may be fabricated using either alloying, thermal-diffusion, or ion-implantation techniques. Various semiconductor-processing technologies such as epitaxy, planar, beam-lead, optical and E-beam lithographies, oxidation, passivations, and dry etching (i.e., reactive ion etching (RIE)) have been developed to facilitate fabrication of silicon ICs and III-V semiconductor optoelectronic devices. Recent advances in processing technologies of III-V compound semiconductor materials and devices have made it possible to develop new high-speed and high-frequency devices using GaAs- and InP-based III-V compound semiconductors. For example, high-speed HBTs have been developed using AlGaAs/GaAs and InAlAs/InGaAs material systems. In general, semiconductor devices fabricated from GaAs/AlGaAs, InGaAs/InP, and other III-V compound semiconductors (e.g., AlGaN/GaN) can be operated at a much higher

frequency and speed than silicon devices because of the inherent high electron mobility of III-V semiconductors. This will be discussed further in Chapter 16.

A BJT device may be operated as an amplifier or as an electronic switch, depending on its bias condition. Unlike a unipolar field-effect transistor (FET), which is a majority carrier device, the BJT is a bipolar device since its current conduction is due to the diffusion of the minority carriers (i.e., electrons in the p region and holes in the n-region) across the p-n junctions. Therefore, the p-n junction theories described in Chapter 11 can be used to derive the minority and majority carrier distributions, current conduction, and the static and dynamic characteristics of a BJT device.

If one adds another p-n junction to the n-p-n BJT structure, a four-layer p-n-p-n switching device can be formed. A p-n-p-n structure is a bistable device whose operation depends on internal feedback mechanisms that can produce high- and low-impedance stable states under bias conditions. This enables the p-n-p-n device to operate as a switching device. The p-n-p-n devices are available for a wide range of voltage and current ratings. The low-power p-n-p-n devices are designed mainly for use in switching and logic circuitry, while high-power p-n-p-n devices find wide applications in AC switching, DC choppers, phase-control devices, and power inverters.

In this chapter, the basic principles that govern the operation of a BJT and a four-layer p-n-p-n device are discussed. A general description of the basic BJT structure and modes of operation is presented in Section 14.2. The distribution of excess carrier densities, current components, and current–voltage (*I–V*) characteristics for a BJT under bias conditions are discussed in Section 14.3. The current gain, base transport factor, and emitter injection efficiency of a BJT are examined in Section 14.4. In Section 14.5, we present the Ebers–Moll and Gummel–Poon models, which provide a powerful means for elucidating the physical insights of the transistor action under different operation modes and biasing conditions. The frequency response and switching properties of a BJT as well as the effect of heavy doping on current gain and limitations due to base resistance and junction breakdown are also discussed. Section 14.6 describes basic device theory and performance characteristics of a Si BJT switching device. Finally, the device structure, operation principles, and performance characteristics of a p-n-p-n four-layer switching device are presented in Section 14.7.

14.2. Basic Device Structures and Modes of Operation

The physical makeup of a typical n^+-p-n (or p^+-n-p) BJT consists of three distinct regions, namely, a heavily doped n^+ (or p^+) emitter region, a thin (0.5 μm or less) p (or n) base region, and a lightly doped n (or p) collector region. Figures 14.1a and b show a vertical n^+-p-n and a vertical p^+-n-p BJT, respectively. Figure 14.1c is a schematic representation of an n^+-p-n transistor, and Figure 14.1d shows the circuit symbols of an n^+-p-n and a p^+-n-p transistor. In general, a BJT may be operated in three different configurations: the common-emitter mode, the common-base

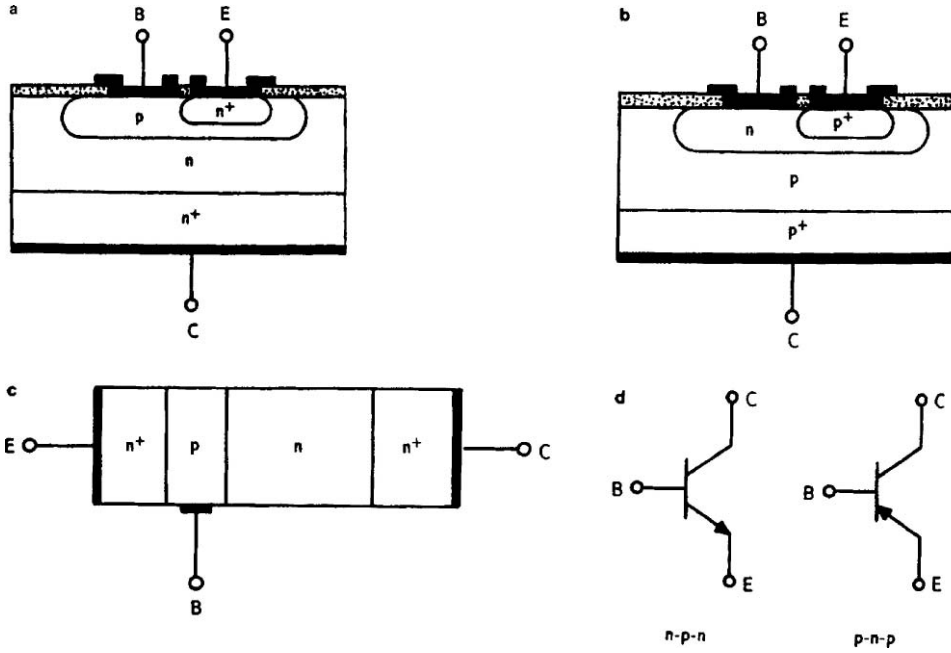


FIGURE 14.1. (a) Cross-sectional view of an n^+-p-n bipolar junction transistor (BJT), (b) a p^+-n-p BJT, (c) schematic diagram of an n^+-p-n BJT, and (d) circuit symbols for an n-p-n and a p-n-p transistor.

mode, and the common-collector mode, as shown in Figure 14.2. In normal active-mode operation, the emitter–base junction is forward-biased and the collector–base junction is reverse-biased. Under this condition, the transistor is operated as an amplifier. In the saturation-mode operation, both the emitter–base and collector–base junctions are forward-biased, while in the cutoff-mode

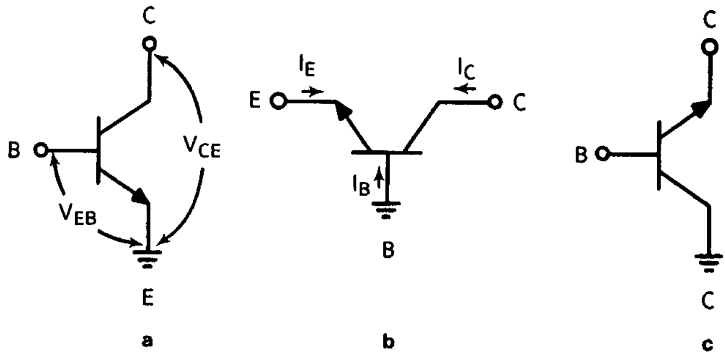


FIGURE 14.2. Three different configurations of an n^+-p-n BJT: (a) common-base, (b) common-emitter, and (c) common-collector connections.

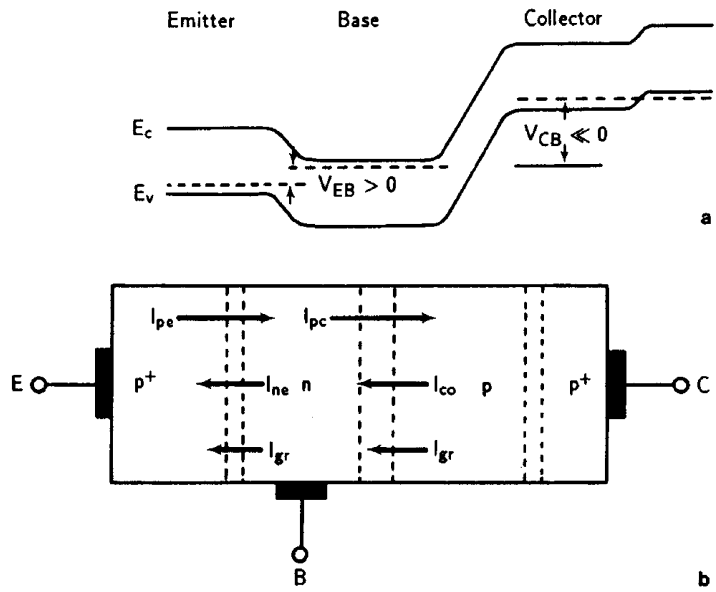


FIGURE 14.3. (a) Energy-band diagram of a p^+-n-p transistor operating in the normal active mode as an amplifier ($V_{EB} > 0$ and $V_{CB} \ll 0$) and (b) current components for a p^+-n-p transistor amplifier.

operation both the emitter–base and collector–base junctions are reverse-biased. Figures 14.3a and b show the energy band diagram and current components of a p^+-n-p BJT operating as an amplifier under normal active mode conditions, respectively.

14.3. Current–Voltage Characteristics

The current–voltage (I–V) characteristics for a BJT under bias conditions can be derived from the p-n junction theories described in Chapter 11. To analyze the dc characteristics of a BJT, it is assumed that the I–V characteristics in the emitter–base (E-B) junction and the collector–base (C-B) junction follow the ideal diode equation. Under this assumption, effects due to surface recombination-generation, series resistance, and high-level injection may be neglected. Figure 14.4 shows the spatial distributions of excess carrier densities in the emitter, base, and collector regions of the BJT. The excess electron and hole densities at the edge of the E-B junction of a p^+-n-p transistor can be expressed as

$$p'_{nb}(0) = p_{nb}(0) - p_{0b} = p_{0b}\left(e^{qV_{BE}/k_BT} - 1\right) \quad \text{at } x = 0 \tag{14.1}$$

and

$$n'_{pe}(-x_E) = n_{pe}(-x_E) - n_{0e} = n_{0e}\left(e^{qV_{BE}/k_BT} - 1\right) \quad \text{at } x = -x_E, \tag{14.2}$$

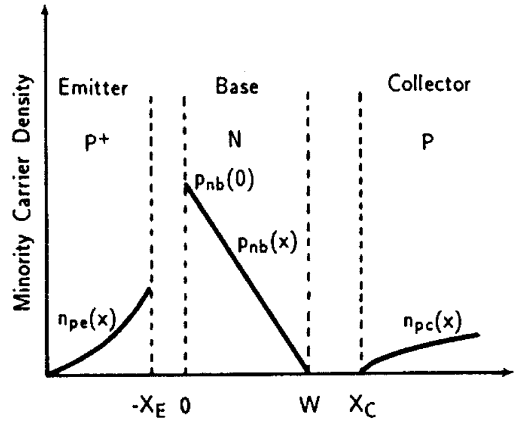


FIGURE 14.4. Minority carrier density distributions for a p^+-n-p transistor with E-B junction forward-biased and C-B junction reverse-biased (assuming $W_E \gg L_{pe}$ and $W_B \ll L_{pb}$).

where p_{0b} and n_{0e} denote the equilibrium hole and electron densities at the edge of the E-B junction, respectively. Similarly, the excess hole and electron densities at the edge of the base–collector junction are given, respectively, by

$$p'_{nb}(W_b) = p_{nb}(W_b) - p_{0b} = p_{0b}\left(e^{qV_{CB}/k_BT} - 1\right) \quad \text{at } x = W_b \tag{14.3}$$

and

$$n'_{pc}(x_C) = n_{pc}(x_C) - n_{0e} = n_{0e}\left(e^{qV_{CB}/k_BT} - 1\right) \quad \text{at } x = x_C, \tag{14.4}$$

where n_{0c} is the equilibrium density of electrons in the collector region. Since the potential drop occurs mainly across the depletion region, the continuity equation for holes in the quasineutral n-base region can be written as

$$D_{pb}\frac{\partial^2 p'_{nb}}{\partial x^2} - \frac{p'_{nb}}{\tau_b} = 0. \tag{14.5}$$

The general solution of (14.5) for the excess hole density in the quasineutral n-base region (i.e., $0 < x < W_b$) is given by

$$p'_{nb}(x) = C_1 e^{-x/L_{pb}} + C_2 e^{x/L_{pb}}, \tag{14.6}$$

where C_1 and C_2 are constants to be determined. Substituting the boundary conditions given by (14.1) and (14.3) into (14.60), we obtain

$$p'_{nb}(x) = p_{n0}\left(e^{qV_{EB}/k_BT} - 1\right)\frac{\sinh\left((W_b - x)/L_{pb}\right)}{\sin\left(W_b/L_{pb}\right)} + p_{n0}\left(e^{qV_{CB}/k_BT} - 1\right)\frac{\sinh\left(x/L_{pb}\right)}{\sinh\left(W_b/L_{pb}\right)}, \tag{14.7}$$

where $L_{pb} = (D_{pb}\tau_b)^{1/2}$ is the hole diffusion length in the n-base region. Equation (14.7) is important, because it relates the minority hole density in the base region to the base width W_b. For example, if W_b approaches infinity (i.e., $W_b/L_{pb} \gg 1$),

then (14.7) reduces to the case of a p-n junction diode, and the transistor action is halted. If the E-B junction is forward-biased and the C-B junction is reverse-biased, then the second term in (14.7) is negligible, and the hole density profile in the base region becomes

$$p'_{\mathrm{nb}}(x) \approx p_{\mathrm{n0}}(\mathrm{e}^{qV_{\mathrm{EB}}/k_{\mathrm{B}}T} - 1)\frac{\sinh((W_{\mathrm{b}} - x)/L_{\mathrm{pb}})}{\sinh(W_{\mathrm{b}}/L_{\mathrm{pb}})}. \tag{14.8}$$

Equation (14.8) can be further simplified for most practical transistors since the base width is much smaller than the hole diffusion length L_{pb}, and hence for $V_{\mathrm{EB}} \gg k_{\mathrm{B}}T/q$, (14.8) reduces to

$$p'_{\mathrm{nb}} \approx p_{\mathrm{n0}}\,\mathrm{e}^{qV_{\mathrm{EB}}/k_{\mathrm{B}}T}\left(1 - \frac{x}{W_{\mathrm{b}}}\right). \tag{14.9}$$

This shows that the minority hole density in the base region decreases linearly with distance x from the edge of the E-B junction to the edge of the base–collector (C-B) junction, as shown in Figure 14.4. Deviation from this linear dependence with distance x can be attributed to the recombination loss occurring in the base region. The hole current in the base region may be derived from the diffusion equation for the excess hole density, which is given by

$$I_{\mathrm{pb}} = -qAD_{\mathrm{pb}}\frac{\mathrm{d}p'_{\mathrm{nb}}(x)}{\mathrm{d}x}, \tag{14.10}$$

where $p'_{\mathrm{nb}}(x)$ is given by (14.7). Thus, the injected hole current entering the base can be evaluated at $x = 0$ using (14.10), and the hole current flowing out of the base is evaluated at $x = W_{\mathrm{b}}$ using (14.10). This yields

$$\begin{aligned} I_{\mathrm{pb}}(0) &= \frac{qD_{\mathrm{pb}}p_{\mathrm{n0}}}{L_{\mathrm{pb}}}\coth\left(\frac{W_{\mathrm{b}}}{L_{\mathrm{pb}}}\right)(\mathrm{e}^{qV_{\mathrm{EB}}/k_{\mathrm{B}}T} - 1) \\ &\quad - \frac{qD_{\mathrm{pb}}p_{\mathrm{n0}}}{L_{\mathrm{pb}}\sinh\left(W_{\mathrm{b}}/L_{\mathrm{pb}}\right)}(\mathrm{e}^{qV_{\mathrm{CB}}/k_{\mathrm{B}}T} - 1), \end{aligned} \tag{14.11}$$

$$\begin{aligned} I_{\mathrm{pb}}(W_{\mathrm{b}}) &= \frac{qD_{\mathrm{pb}}p_{\mathrm{n0}}}{L_{\mathrm{pb}}\sinh\left(W_{\mathrm{b}}/L_{\mathrm{pb}}\right)}(\mathrm{e}^{qV_{\mathrm{EB}}/k_{\mathrm{B}}T} - 1) \\ &\quad - \frac{qD_{\mathrm{pb}}p_{\mathrm{n0}}}{L_{\mathrm{pb}}}\coth\left(W_{\mathrm{b}}/L_{\mathrm{pb}}\right)(\mathrm{e}^{qV_{\mathrm{CB}}/k_{\mathrm{B}}T} - 1). \end{aligned} \tag{14.12}$$

It is seen that the hole current in the base region is, in general, a function of the applied bias voltages at both the E-B and C-B junctions. The polarity of the bias voltages at both junctions can be changed depending on the operation modes of the BJTs.

Similarly, the excess electron densities in the p^+-emitter and p-collector regions of the BJTs can be determined by solving the continuity equations for the excess

electron densities in both regions; they are given respectively by

$$n'_{\mathrm{pe}}(x) = n'_{\mathrm{pe}}(-x_{\mathrm{E}})\, \mathrm{e}^{(x+x_{\mathrm{E}})/L_{\mathrm{ne}}} \quad \text{for } x < -x_{\mathrm{E}}, \tag{14.13}$$

$$n'_{\mathrm{pc}}(x) = n'_{\mathrm{pc}}(x_{\mathrm{C}})\, \mathrm{e}^{-(x-x_{\mathrm{C}})/L_{\mathrm{nc}}} \quad \text{for } x > x_{\mathrm{C}}, \tag{14.14}$$

where $n'_{\mathrm{pe}}(-x_{\mathrm{E}})$ and $n'_{\mathrm{pc}}(x_{\mathrm{C}})$ are the excess electron densities at the edges of the E-B and C-B junctions defined by (14.2) and (14.4), respectively. Equations (14.13) and (14.14) show that the excess electron densities decrease exponentially with distance from the depletion edge of both the E-B and C-B junctions, as shown in Figure 14.4. The electron current in the emitter and collector regions can be derived from the diffusion equation given by

$$I_{\mathrm{n}} = qAD_{\mathrm{n}}\frac{\mathrm{d}n'_{\mathrm{p}}}{\mathrm{d}x}. \tag{14.15}$$

The total emitter current, which consists of the electron injection current from the emitter to the base and the hole-injection current from the base to the emitter, is given by

$$\begin{aligned} I_{\mathrm{E}} &= A'\left(I_{\mathrm{pE}} + I_{\mathrm{nE}}\right) \\ &= -A'qD_{\mathrm{pb}}\frac{\mathrm{d}p'_{\mathrm{nb}}}{\mathrm{d}x}\bigg|_{x=0} + A'qD_{\mathrm{ne}}\frac{\mathrm{d}n'_{\mathrm{pe}}}{\mathrm{d}x}\bigg|_{x=-x_{\mathrm{E}}} \\ &= I_{\mathrm{BO}}\coth\left(\frac{W_{\mathrm{b}}}{L_{\mathrm{pb}}}\right)\left[\left(\mathrm{e}^{qV_{\mathrm{EB}}/k_{\mathrm{B}}T}-1\right) - \frac{1}{\cosh\left(\frac{W_{\mathrm{b}}}{L_{\mathrm{pb}}}\right)}\left(\mathrm{e}^{qV_{\mathrm{CB}}/k_{\mathrm{B}}T}-1\right)\right] \\ &\quad + I_{\mathrm{EO}}\left(\mathrm{e}^{qV_{\mathrm{EB}}/k_{\mathrm{B}}T}-1\right), \end{aligned} \tag{14.16}$$

where

$$I_{\mathrm{BO}} = A'qn_{\mathrm{i}}^2\left(\frac{D_{\mathrm{pb}}}{N_{\mathrm{db}}L_{\mathrm{pb}}}\right), \tag{14.17}$$

$$I_{\mathrm{EO}} = A'qn_{\mathrm{i}}^2\left(\frac{D_{\mathrm{ne}}}{N_{\mathrm{ae}}L_{\mathrm{ne}}}\right), \tag{14.18}$$

are the saturation currents in the base and emitter regions of the BJTs, respectively. Similarly, the collector current can be expressed by

$$\begin{aligned} I_{\mathrm{C}} &= A\left(I_{\mathrm{pC}} + I_{\mathrm{nC}}\right) \\ &= -AqD_{\mathrm{pb}}\frac{\mathrm{d}p'_{\mathrm{nb}}}{\mathrm{d}x}\bigg|_{x=W_{\mathrm{B}}} + AqD_{\mathrm{nc}}\frac{\mathrm{d}n'_{\mathrm{pc}}}{\mathrm{d}x}\bigg|_{x=x_{\mathrm{C}}} \\ &= \frac{I_{\mathrm{BO}}}{\sinh\left(W_{\mathrm{b}}/L_{\mathrm{pb}}\right)}\left[\left(\mathrm{e}^{qV_{\mathrm{EB}}/k_{\mathrm{B}}T}-1\right) - \coth\left(\frac{W_{\mathrm{b}}}{L_{\mathrm{pb}}}\right)\left(\mathrm{e}^{qV_{\mathrm{CB}}/k_{\mathrm{B}}T}-1\right)\right] \\ &\quad + I_{\mathrm{CO}}\left(\mathrm{e}^{qV_{\mathrm{CB}}/k_{\mathrm{B}}T}-1\right), \end{aligned} \tag{14.19}$$

where

$$I_{\mathrm{CO}} = Aqn_{\mathrm{i}}^2\left(\frac{D_{\mathrm{nc}}}{N_{\mathrm{ac}}L_{\mathrm{nc}}}\right) \tag{14.20}$$

is the collector saturation current, A is the C-B junction area, and A' is the E-B junction area. If the direction of current flow into the emitter, base, and collector junctions shown in Figure 14.3b is defined as positive, then the base current I_B is related to the emitter and collector currents by

$$I_B = -I_E - I_C, \tag{14.21}$$

where I_E and I_C are the emitter and collector currents given by (14.16) and (14.19), respectively. Since the emitter current is nearly equal to the collector current, the base current I_B is usually very small. It is noted that (14.21) does not include the recombination current in the base region. As shown by (14.9), for a uniformly doped base with negligible base recombination (i.e., $W_b \ll L_{pb}$) the injected excess hole density is a linear function of x in the base region. In general, the recombination current in the base region can be expressed by

$$I_r = \left(\frac{qA'}{\tau_p}\right)\int_0^{W_b} p'_{nb}(x)\,dx \approx \left(\frac{qA'n_iW}{2\tau_0}\right)e^{qV_{EB}/2k_BT}. \tag{14.22}$$

If the recombination current component in the base is not negligible, then (14.22) should be added to (14.21) to obtain the total base current.

It should be noted that (14.16) through (14.22) for the current components derived above are valid only for a BJT with a uniformly doped base. However, for a practical BJT in which the E-B and B-C junctions are formed by double diffusion or ion implantation, the base impurity doping profile is no longer uniform and a built-in electric field exists in the base region. This is illustrated in Figure 14.5 for

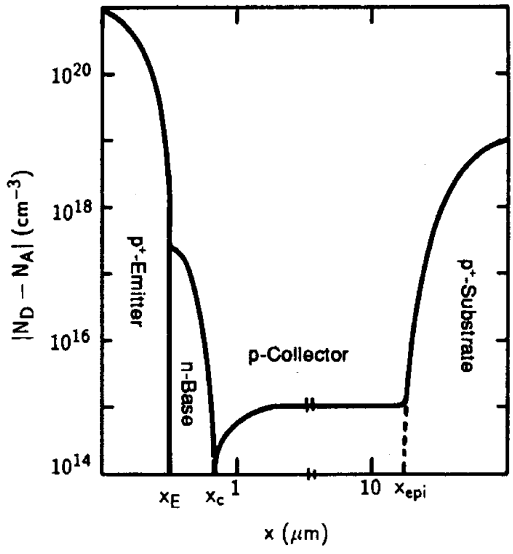


FIGURE 14.5. Impurity density profile for a double-diffused planar epitaxial structure of a p$^+$-n-p-p$^+$ transistor.

a double-diffused planar p^+-n-p-p^+ BJT. The built-in electric field $\mathcal{E}$ due to the nonuniform doping profile in the base region of a BJT can be expressed by

$$\mathcal{E} = \frac{k_B T}{q N_d(x)} \frac{dN_d(x)}{dx}. \tag{14.23}$$

The polarity of the built-in electric field given by (14.23) is such that it assists the transport of injected holes in the base region. Thus, the hole current density in this case is given by

$$J'_{pb} = q\mu_p p_{nb}\mathcal{E} - qD_{pb}\frac{dp_{nb}}{dx} = qD_{pb}\left[\left(\frac{p_{nb}}{N_{db}}\right)\frac{dN_{db}(x)}{dx} + \frac{dp_{nb}}{dx}\right]. \tag{14.24}$$

Equation (14.24) is obtained by substituting $\mathcal{E}$ given by (14.23) for the electric field and using μ_p from the Einstein relation (i.e., $\mu_p = (k_B T/q)^{-1} D_p$) in the first term of (14.24). Multiplying both sides of (14.24) by N_{db} and integrating the equation yields

$$p'_{nb}(x) = \frac{J'_{pb}}{qD_{pb}N_{db}} \int_x^{W_b} N_{db}\, dx. \tag{14.25}$$

It is noted that (14.25) is obtained using the boundary condition $p'_{nb}(W_b) = 0$ and $x = W_b$ and by assuming that J'_{pb} is constant (i.e., the recombination current is negligible in the base region). From (14.25), the hole density at $x = 0$ is given by

$$p'_{nb}(0) = \frac{n_i^2}{N_{db}(0)} e^{qV_{BE}/k_B T} = \frac{J'_{pb}}{qD_{pb}N_{db}(0)} \int_0^{W_b} N_{db}(x)\, dx. \tag{14.26}$$

Thus, the hole current in the base region can be derived from (14.26), which yields

$$I'_{pb} = J'_{pb}A' = \frac{\left(qA'D_{pb}n_i^2\right) e^{qV_{BE}/k_B T}}{\int_0^{W_b} N_{db}(x)\, dx}. \tag{14.27}$$

The integral in the denominator of (14.27) represents the total number of impurity atoms in the base and is known as the Gummel number. For silicon BJTs the Gummel number may vary between 10^{12} and 10^{13} cm^{-2}. Therefore, a larger electron current flow can be realized with a smaller Gummel number, which corresponds to a narrower base width. Figure 14.6 shows the base and collector currents versus the E-B junction bias voltage for a silicon BJT.[1] Four regions are observed in this plot: (i) the low-V_{BE}, nonideal region in which the base current is dominated by the recombination current and I_B varies with $e^{qV_{EB}/2k_B T}$; (ii) the ideal region

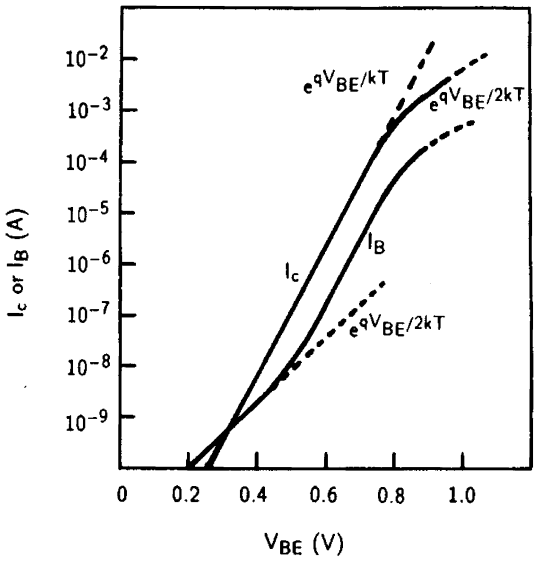


FIGURE 14.6. Base and collector currents as a function of the emitter-base bias voltage for a silicon BJT under forward-bias conditions. After Jespers,[1] by permission.

in which both the base and collector currents are dominated by the diffusion current (i.e., I_B and $I_C \approx e^{qV_{EB}/k_BT}$; (iii) the moderate-injection region in which a significant voltage drop occurs across the base resistance (i.e., r_bI_B drop); (iv) the high-injection region in which I_C and I_B vary with $e^{qV_{EB}/2k_BT}$. In general, the recombination current generated in the base region can be decreased by reducing the processing-related defects in this region, while the high-injection and base-resistance effects can be minimized by modifying the base doping profile and the transistor structure.

The output *I–V* (i.e., I_C vs. V_{CE}) characteristics for a silicon p^+-n-p BJT with a common-emitter configuration is shown in Figure 14.7.[2] Also shown in this figure are the Early voltage V_A, the collector saturation current I'_{CO} (also known as I_{CEO}),

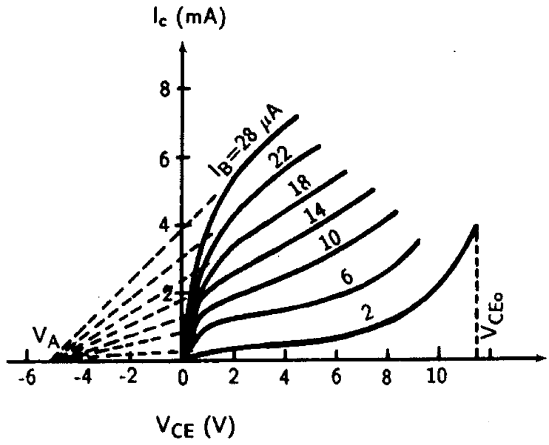


FIGURE 14.7. I_c versus V_{CE} for a p^+-n-p silicon BJT with common-emitter configuration. Also shown in the figure are the Early voltage V_A, saturation current I'_{CO}, and breakdown voltage V_{CEO} for the common-emitter configuration. After Gummel and Poon,[2] by permission.

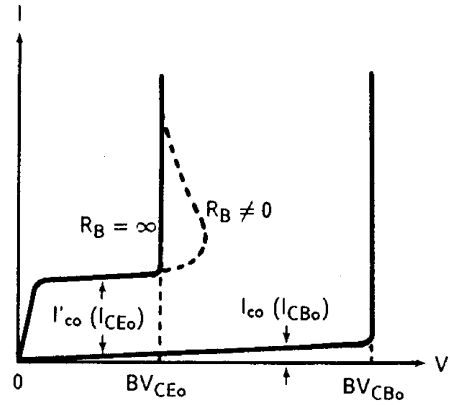


FIGURE 14.8. Breakdown voltage and saturation current for p^+-n-p silicon BJTs with common-base and common-emitter configurations.

and the breakdown voltage V_{CEO}. The breakdown voltages and saturation currents for p^+-n-p silicon BJTs with common-base and common-emitter configurations are shown in Figure 14.8. It is seen that the saturation current I_{CO} (also known as I_{CBO}) for the common-base configuration is substantially smaller than I_{CEO} for the common-emitter configuration (i.e., by a factor of β_0). In the common-base configuration, the current gain α_0 is close to unity and I_C is nearly independent of V_{CB}. From the current equations derived earlier, we see that both the emitter current I_E and collector current I_C are functions of the applied voltage across the E-B and C-B junctions. In the common-emitter configuration the current gain β_0 can be quite large and I_C usually increases with increasing V_{CE}. The collector saturation current I'_{CO} for the common emitter configuration (i.e., base opened and $I_B = 0$) is related to the saturation current I_{CBO} for the common-base configuration by

$$I'_{CO} = I_{CEO} = \frac{I_{CO}}{(1 - \alpha_0)} = \beta_0 I_{CBO}. \tag{14.28}$$

The base width W_b will decrease with increasing V_{CE}, which in turn will cause an increase in the current gain. The continued increase of I_C with increasing V_{CE} is attributed to a large increase of β_0 with V_{CE} in the common-emitter mode of operation. This phenomenon is called the *Early effect*, which is a direct result of the base-width modulation by the C-B junction bias voltage variation. It is noted that in a BJT, a change of C-B junction bias voltage will result in a change of space-charge layer width at the C-B junction and consequently will modify the width of the quasineutral base region. This variation will result in several effects that further complicate the performance of a BJT as a linear amplifier. The voltage V_A in which the extrapolated I_C versus V_{CE} curves (see Figure 14.7) meet the negative V_{CE} axis is called the *Early voltage*, which is given by

$$V_A = \frac{qN_B W_b^2}{\varepsilon_0 \varepsilon_s}. \tag{14.29}$$

Equation (14.29) is valid for a BJT whose base width W_b is much larger than its depletion layer width in the base. To reduce the influence of the C-B voltage on the

collector current, the value of V_A must be increased. From (14.29), it is seen that this can be accomplished by increasing the base doping density, which, in turn, will reduce the C-B depletion layer width and hence the *Early effect*. This reduces the movement of the C-B boundary to the base region of a BJT.

The breakdown voltage BV_{CEO} for the common-emitter configuration can be related to the breakdown voltage BV_{CBO} for the common-base configuration by

$$BV_{CEO} = BV_{CBO}(1 - \alpha_0)^{1/m}, \tag{14.30}$$

where m is an integer. Since α_0 is very close to unity for most BJTs, BV_{CBO} is usually much larger than BV_{CEO}. It is noted that BV_{CBO} under open-base conditions can be related to the multiplication factor M by

$$M = \frac{1}{1 - (V/BV_{CBO})^m}, \tag{14.31}$$

where V is the applied bias voltage. When the base is opened, the emitter and collector currents are equal (i.e., $I_E = I_C = I$). Both I_{CO} and $\alpha_0 I_E$ are multiplied by M as they flow across the C-B junction. From (14.30) it is seen that for $\alpha_0 \approx 1$, BV_{CBO} becomes much larger than BV_{CEO}. This is clearly illustrated in Figure 14.8, which shows the breakdown voltage BV_{CBO} and saturation current I_{CO} for the common-base configuration, and the corresponding BV_{CEO} and I'_{CO} for the common-emitter configuration.

The current–voltage equations derived in this section for a BJT will be used in the Ebers–Moll model for large-signal and transient analysis, which will be discussed in detail in Section 14.5.

14.4. Current Gain, Base Transport Factor, and Emitter Injection Efficiency

When a BJT is biased in the normal active mode, it operates as an amplifier, and hence a current gain results. For a p^+-n-p transistor, the emitter current consists of two components: a hole current I_{pE}, which is due to hole injection from the p^+-emitter into the n-base region, and an electron current I_{nE}, which is due to electron injection from the n-base into the p-emitter region. The collector current also consists of two components, namely, a hole current I_{pC} injecting from the n-base into the p-collector region, and an electron current I_{nC} injecting from the p-collector into the n-base region. Expressions for the emitter- and collector-current components are given by (14.16) and (14.19), respectively. For a common-base BJT amplifier, the key parameters affecting its performance include the emitter injection efficiency, the base transport factor, and the current gain. If the base recombination current component is included, then the emitter injection efficiency

γ can be expressed by

$$\gamma = \frac{I_{\mathrm{pE}}}{I_{\mathrm{E}}} = \frac{I_{\mathrm{pE}}}{(I_{\mathrm{nE}} + I_{\mathrm{pE}} + I_{\mathrm{r}})}. \tag{14.32}$$

And the base transport factor β_{T} is given by

$$\beta_{\mathrm{T}} = \frac{I_{\mathrm{pC}}}{I_{\mathrm{pE}}}. \tag{14.33}$$

The common-base current gain α_0 is defined by

$$\alpha_0 = h_{\mathrm{FB}} = \frac{\mathrm{d}I_{\mathrm{C}}}{\mathrm{d}I_{\mathrm{E}}} = \frac{-(I_{\mathrm{C}} - I_{\mathrm{CO}})}{I_{\mathrm{E}}} = \frac{I_{\mathrm{pC}}}{(I_{\mathrm{nE}} + I_{\mathrm{pE}} + I_{\mathrm{r}})} = \gamma\ \beta_{\mathrm{T}}, \tag{14.34}$$

which shows that for a common-base BJT the current gain is equal to the product of emitter injection efficiency and base transport factor. Since I_{pC} is smaller than I_{pE}, the common-base current gain α_0 is always smaller than unity. However, for a well-designed BJT, this gain factor can be very close to unity (e.g., $\alpha_0 = 0.9999$), and from (14.34) one obtains

$$I_{\mathrm{C}} = -\alpha_0 I_{\mathrm{E}} + I_{\mathrm{CO}}, \tag{14.35}$$

which relates the collector current to the emitter current with the base as a common terminal. It is noted that I_{CO} is the collector reverse saturation current for the C-B configuration.

To obtain current amplification in a BJT, the transistor is usually operating in the common-emitter configuration. In this configuration, the emitter terminal is used as a common ground, the B-E terminal is used as an input port, and the C-E terminal serves as an output port. The common-emitter current gain β_0 (or h_{FE}) is defined by

$$\beta_0 = h_{\mathrm{FE}} = \frac{\mathrm{d}I_{\mathrm{C}}}{\mathrm{d}I_{\mathrm{B}}} = \frac{\alpha_0}{(1 - \alpha_0)}, \tag{14.36}$$

which is obtained by solving (14.21) and (14.35). Since the value of α_0 for a well-designed BJT is very close to unity, β_0 for the common-emitter operation is usually much larger than unity (e.g., if $\alpha_0 = 0.99$, then $\beta_0 = 99$).

For a p^+-n-p transistor operating under normal active-mode conditions (i.e., $V_{\mathrm{BE}} > 0$ and $V_{\mathrm{CB}} \ll 0$), the emitter injection efficiency γ can be derived using (14.16), and one obtains

$$\gamma \approx \frac{I_{\mathrm{pE}}}{(I_{\mathrm{pE}} + I_{\mathrm{nE}})} = \frac{1}{1 + (N_{\mathrm{db}} D_{\mathrm{ne}} L_{\mathrm{pb}} / N_{\mathrm{ae}} D_{\mathrm{pb}} L_{\mathrm{ne}}) \tanh(W_b / L_{\mathrm{pb}})}. \tag{14.37}$$

Equation (14.37) neglects the base recombination current. Similarly, the base transport factor can be obtained by solving (14.16) and (14.19), which yields

$$\beta_{\mathrm{T}} = \frac{I_{\mathrm{pC}}}{I_{\mathrm{pE}}} = \frac{1}{\cosh(W_{\mathrm{b}} / L_{\mathrm{pb}})} \approx 1 - \frac{W_{\mathrm{b}}^2}{2L_{\mathrm{pb}}^2}. \tag{14.38}$$

In the above derivation it is assumed that the base width is much smaller than the hole-diffusion length in the base region. An interesting physical insight of β_T can be obtained if (14.38) is expressed in terms of the transit time τ_B and the minority carrier lifetime τ_p in the base region. It can be shown that in order to have a base transport factor close to unity, the base transit time must become so short (i.e., $\tau_B \ll \tau_p$) that the injected holes have little chance to recombine with electrons in the base region. For practical silicon BJTs, β_T is very close to unity, and hence the current gain β_0 can be obtained by solving (14.36) and (14.37), which yields

$$\begin{aligned}\beta_0 = h_{FE} &= \frac{\alpha_0}{1-\alpha_0} = \frac{\gamma\beta_T}{1-\gamma\,\beta_T} \approx \frac{\gamma}{(1-\gamma)} \\ &= \left(\frac{N_{ae}D_{pb}L_{ne}}{N_{db}D_{ne}L_{pb}}\right)\coth\left(\frac{W_b}{L_{pb}}\right) \approx \frac{N_{ae}}{Q_B},\end{aligned} \tag{14.39}$$

where Q_B is the Gummel number defined by the denominator of (14.27). Thus, for a given emitter doping density N_{ae}, the static common-emitter current gain β_0 is inversely proportional to the base charge density Q_B. Figure 14.9 illustrates the current gain h_{FE} versus collector current I_C for the silicon BJT shown in Figure 14.6. As shown in Figure 14.9, β_0 is small when I_C is small. This can be attributed to bulk and surface recombination losses that occurred in the base region at small-bias voltages. The recombination current in the base region may be larger than the diffusion current component at low current level. Thus, by reducing the bulk trap density and surface recombination loss, the current gain can be increased substantially at low collector current. As the collector current continues to increase, h_{FE} will also increase and eventually reach a saturation value. At still higher collector current, the minority carrier density injected into the base approaches the majority carrier density, and the injected carriers effectively increase the base doping density, which, in turn, will cause the emitter injection efficiency to decrease. This is the so-called high-injection condition. At high-injection

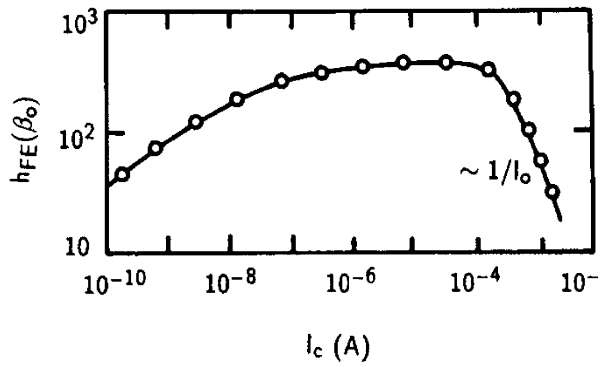


FIGURE 14.9. Common-emitter current gain versus collector current for the silicon BJTs shown in Figure 14.6.

levels, the current gain varies inversely with collector current (i.e., $h_{FE} \sim e^{-qV_{EB}/2k_BT} \sim 1/I_c$), as shown in Figure 14.9.

Equation (14.39) shows that h_{FE} is directly proportional to the emitter doping density. Therefore, in order to increase h_{FE} it is necessary to increase the doping density in the emitter region. However, two adverse effects associated with heavy doping in the emitter region may result, namely, band gap narrowing and Auger recombination. Both of these effects can severely affect the current gain of the BJTs. The effect of band gap narrowing on the current gain can be evaluated by examining the band gap narrowing effect on the effective intrinsic carrier density in the heavily doped emitter region. The square of the effective intrinsic carrier density under heavy doping conditions is given by

$$n_{ie}^2 = N_c N_v \exp\left[-\left(E_g - \Delta E_g\right)/k_B T\right] = n_i^2 \, e^{\Delta E_g/k_B T}, \tag{14.40}$$

where N_c and N_v are the effective densities of the conduction and valence band states, respectively, while n_i is the intrinsic carrier density for the nondegenerate case. The quantity ΔE_g is the band gap shrinkage due to the heavy doping effect in the emitter region. The minority carrier densities in the base and emitter regions are given respectively by

$$p_{nb} = \frac{n_i^2}{N_{db}} \tag{14.41}$$

and

$$n_{pe} = \frac{n_{ie}^2}{N_{ae}} = \frac{n_i^2}{N_{ae}} e^{\Delta E_g/k_B T}. \tag{14.42}$$

Therefore, the effect of band gap narrowing on the current gain can be estimated qualitatively from (14.39) to (14.42), which yields

$$h_{FE} \approx \frac{p_{nb}}{n_{pe}} \approx e^{-\Delta E_g/k_B T}. \tag{14.43}$$

Equation (14.43) shows that h_{FE} will decrease exponentially with increasing band gap narrowing ΔE_g.

Another heavy doping effect that can greatly degrade the transistor performance is associated with the reduction of minority carrier lifetime with increasing doping density in the emitter region of the BJTs. As the doping density in the emitter region increases, Auger recombination becomes the dominant recombination process for the minority carriers. In this case, the minority carrier lifetime in the emitter region of a BJT is controlled by the Auger recombination process instead of the Shockley–Read–Hall (SRH) process. As a result, the minority carrier lifetime in the emitter region will decrease with the square of the majority carrier density. This, in turn, will reduce the emitter minority carrier diffusion length and degrade the emitter injection efficiency. Figure 14.10 shows the effects of band gap narrowing and Auger recombination on the current gain of a silicon power transistor.[3] The results clearly show that in order to accurately predict the measured current gain data,

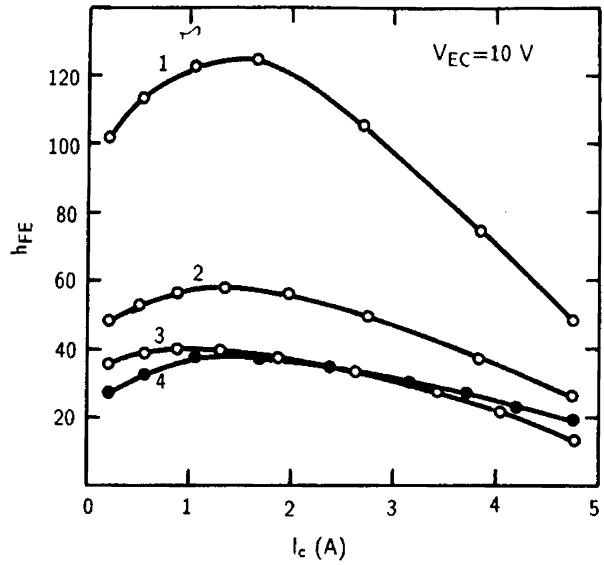


FIGURE 14.10. Calculated and measured common-emitter current gain h_{FE} versus collector current for a silicon power transistor by considering (1) the SRH process only, (2) SRH and bandgap narrowing, (3) measured values, and (4) Auger, SRH, and bandgap narrowing. After McGrath and Navon,[3] by permission, © IEEE–1977.

the effects of band gap narrowing and Auger recombination in the heavily doped emitter region should be taken into account. The relative importance of these effects on the SRH recombination process depends on the emitter junction depth and the dopant density as well as the injection level.

The base-spreading resistance is another important parameter that will affect the performance of a BJT at very high frequencies. In general, it will increase the B-E junction voltage drop at high base current for power and switching transistors. A close examination of the cross-section of the BJT shown in Figure 14.1 reveals that the base current must flow some distance from the base terminal to the bulk base region between the E-B and C-B junctions. Since the base region is very thin (i.e., ≤ 0.5 μm) and not highly doped, a parasitic resistance known as base spreading resistance $r_{b'b}$ exists in this region. This spreading resistance must be included in the BJT device modeling to account for its adverse effect on transistor performance at high frequencies and at high injection level.

14.5. Modeling of a Bipolar Junction Transistor

In this section, we present the Ebers–Moll model and the Gummel–Poon model for BJTs. In order to predict the performance of a BJT for large signals or transient behavior under any biasing conditions, it is necessary to develop a simple and accurate device model so that its electrical output characteristics can be

correlated to the physical parameters of the transistor. This is particularly important for IC designs, since an accurate device model is needed in the design of any integrated circuit. The first BJT device model for large-signal circuit simulation was introduced by Ebers and Moll in 1954, and later modified by Gummel and Poon to account for various physical effects that were not included in the Ebers–Moll model. The Ebers–Moll model is the simplest device model for the BJTs that can be used to predict carrier injection and extraction phenomena in a BJT.

The BJT device model developed by Gummel and Poon is based on the integral charge equation that relates terminal electrical characteristics to the charges in the base region. By taking into account many physical effects in the device modeling parameters, the Gummel–Poon model predicts the transistor behavior more accurately than does the Ebers–Moll model. To implement the Gummel–Poon model for computer circuit simulation, many physical parameters must first be determined. It can be shown that a simplified version of the Gummel–Poon model can be reduced to the basic Ebers–Moll model.

Figure 14.11a shows the equivalent circuit of the simplest Ebers–Moll model for a BJT.[4] This large-signal transistor model consists of two diodes connected back to back, and each diode is connected in parallel with a current source. The current sources are driven by the diode currents, which are assumed to have ideal diode characteristics. Using the results derived in Section 14.2 for p^+-n-p BJTs, the terminal current equations for the Ebers–Moll model can be written as

$$I_E = I_F - \alpha_R I_R, \tag{14.44}$$

$$I_C = I_R - \alpha_F I_F, \tag{14.45}$$

$$I_B = -(I_E + I_C) = -(1 - \alpha_F) I_F - (1 - \alpha_R) I_R, \tag{14.46}$$

where

$$I_F = I_{ES} \left(e^{q V_{BE}/k_B T} - 1 \right), \tag{14.47}$$

$$I_R = I_{CS} \left(e^{q V_{BC}/k_B T} - 1 \right). \tag{14.48}$$

Note that I_F is the forward current flowing through the E-B junction and I_R is the reverse current flowing through the C-B junction; α_F and α_R denote the forward and reverse C-B current gains, respectively, while I_{ES} and I_{CS} denote the emitter and collector saturation currents, respectively. Expressions for α_F, α_R, I_{ES}, and I_{CS} can be derived from (14.16) through (14.20) for a p^+-n-p BJT and are given by

$$I_{ES} = \frac{q A' D_{pb} n_i^2}{N_{db} L_{pb}} \coth\left(\frac{W_b}{L_{pb}}\right) + \frac{q A' D_{ne} n_i^2}{N_{ae} L_{ne}}, \tag{14.49}$$

$$I_{CS} = \frac{q A D_{pb} n_i^2}{N_{db} L_{pb}} \coth\left(\frac{W_b}{L_{pb}}\right) + \frac{q A D_{nc} n_i^2}{N_{ac} L_{nc}}, \tag{14.50}$$

$$\alpha_F = \frac{1}{I_{ES}} \frac{q A' D_{pb} n_i^2}{N_{db} L_{pb}} \frac{1}{\sinh(W_b/L_{pb})}, \tag{14.51}$$

$$\alpha_R = \frac{1}{I_{CS}} \frac{q A D_{pb} n_i^2}{N_{db} L_{pb}} \frac{1}{\sinh(W_b/L_{pb})}. \tag{14.52}$$

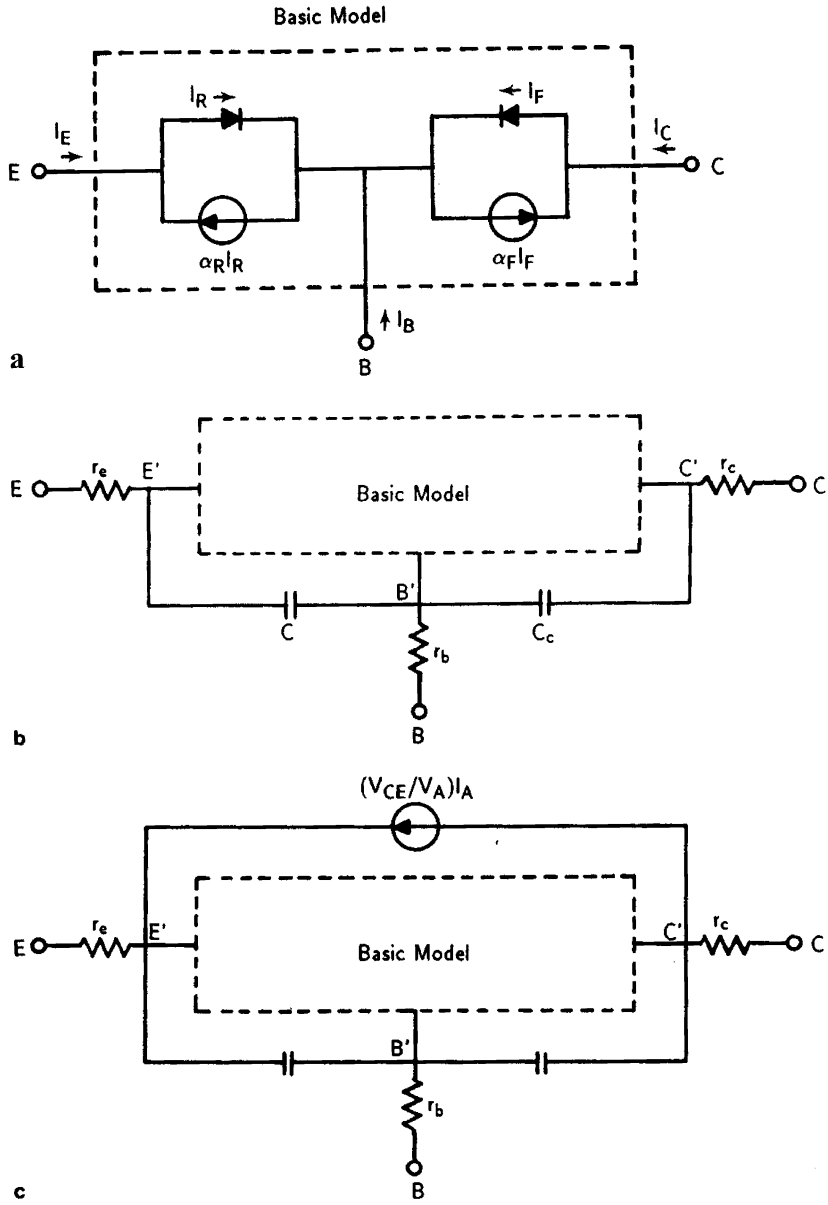


FIGURE 14.11. Equivalent circuit diagrams of an n-p-n transistor based on the Ebers–Moll model: (a) basic model, (b) modified model including series resistance and depletion capacitances, and (c) additional current source for the Early effect. After Ebers and Moll,[4] by permission, © IEEE–1961.

Equations (14.44) and (14.45) relate currents I_E and I_C to the terminal voltages V_{EB} and V_{CB}, and the four transistor parameters I_{ES}, I_{CS}, α_F, and α_R. The current equations for the emitter and collector junctions given above enable the general expressions for the emitter and collector currents to be rewritten as

$$I_E = a_{11}\left(e^{qV_{EB}/k_BT} - 1\right) + a_{12}\left(e^{qV_{CB}/k_BT} - 1\right), \tag{14.53}$$

$$I_C = a_{21}\left(e^{qV_{EB}/k_BT} - 1\right) + a_{22}\left(e^{qV_{CB}/k_BT} - 1\right), \tag{14.54}$$

where

$$a_{11} = -I_{ES}, \qquad a_{12} = \alpha_R I_{CS}, \qquad a_{21} = \alpha_F I_{ES}, \qquad a_{22} = -I_{CS}. \tag{14.55}$$

Based on the reciprocity property of a two-port device, one obtains $a_{12} = a_{21}$, and hence $\alpha_R I_{CS} = \alpha_F I_{ES}$. Therefore, only three unknowns are involved in the basic Ebers–Moll model shown in Figure 14.11a. The accuracy of this basic model can be improved by adding the emitter- and collector-series resistances (r_e and r_c) and the emitter- and collector-depletion capacitances (C_e and C_c) to the equivalent circuit shown in Figure 14.11a, and the result is shown in Figure 14.11b. In this case, the diode is controlled by the internal junction voltages $V_{E'B'}$ and $V_{C'B'}$ but not by the external voltages. If one adds the Early effect (i.e., the base-width modulation) to the model, then an extra current source must be included between the internal emitter and the collector terminals, as shown in Figure 14.11c. A comparison of Figures 14.11a and b shows that in order to improve the model accuracy from Figures 14.11a and b, the unknown physical parameters must increase threefold. This makes the model too complicated to handle and more difficult to solve. Furthermore, the model shown in Figure 14.11b can be improved by adding a diode to the base lead to account for the two-dimensional current crowding effect along the E-B junction. Therefore, it is evident that the basic Ebers–Moll model can provide a first-order solution for relating the device physical parameters to the large-signal dc and transient characteristics of a BJT. The accuracy and complexity of this model depend on the number of physical effects being considered in the model. This can be best illustrated by using the Gummel–Poon model, in which more than 20 physical parameters are incorporated in the equivalent circuit.

The Ebers–Moll model described above may also be applied to n^+-p-n BJTs provided that the polarities defined for I_E, I_C, I_B, V_{EB}, and V_{CB} are reversed. On the basis of the Ebers–Moll model given by (14.53) and (14.54) one notices that there are three regions of operation for the common-base or the common-emitter configuration. As shown in Figure 14.7, the three regions of operation for a common-emitter configuration are (i) the cutoff region with $V_{EB} < 0$ and $V_{CB} \ll 0$, (ii) the active region with $V_{EB} > 0$ and $V_{CB} \ll 0$, and (iii) the saturation region with $V_{EB} > 0$ and $V_{CB} \gg 0$. In region (i) both diodes are reverse-biased and only leakage currents flow through the transistor. This region corresponds to the "off" state in the switching transistor operation. In region (ii), the transistor operates as an amplifier. In this region of operation (i.e., normal active mode), a change in the base current due to a small change in input voltage V_{EB} across the E-B junction at the input terminal will result in a large change in the collector current, and

hence a voltage drop across the load resistance in the collector output terminal with consequent voltage and power amplification. In region (iii), both junctions are forward-biased, and V_{CE} is nearly equal to 0 but with a large collector current. This region corresponds to the "on" state in the switching transistor operation.

The Gummel–Poon model is widely used in modeling the BJTs for various IC designs.[2] It is based on the integral charge model that relates the terminal electrical characteristics to the base charge. This device model is very accurate since it takes many physical effects into consideration. For example, over two dozen physical parameters are needed to cover a wide range of transistor operation. In the Gummel–Poon model, the current that flows from the emitter to the collector terminals with unit current gain is given by

$$I_{CC} = \frac{(qn_i A)^2}{Q_b} \left(e^{qV_{EB}/k_B T} - e^{qV_{CB}/k_B T} \right), \tag{14.56}$$

where

$$Q_b = qA \int_0^{W_b} p_b(x)\, dx \tag{14.57}$$

is the base charge and A is the junction area. The Gummel–Poon model is based on the control of base charge given by (14.57), which links junction voltages, collector current, and base charge. The base charges consist of five components, and are given by

$$\begin{aligned} Q_b &= Q_{b0} + Q_{je} + Q_{jc} + Q_{de} + Q_{dc} \\ &= Q_{b0} + Q_{je} + Q_{jc} + \tau_F I_F + \tau_R I_R, \end{aligned} \tag{14.58}$$

where Q_{b0} is the zero-bias charge in the base region, Q_{je} and Q_{jc} are charges associated with the emitter and collector junction depletion capacitances, respectively, and $Q_{de}(= \tau_F I_F)$ and $Q_{dc}(= \tau_R I_R)$ represent minority carrier charges associated with the emitter and collector diffusion capacitances, respectively. As the injection level increases, the diffusion capacitance also increases, which results in high-injection gain degradation. The current flow from the emitter region to the collector region may be written as

$$I_{CC} = I_F - I_R, \tag{14.59}$$

where

$$I_F = \frac{I_s Q_{b0}}{Q_b} \left(e^{qV_{BE}/k_B T} - 1 \right), \tag{14.60}$$

$$I_R = \frac{I_s Q_{b0}}{Q_b} \left(e^{qV_{BC}/k_B T} - 1 \right). \tag{14.61}$$

It is interesting to note that (14.60) and (14.61) resemble (14.47) and (14.48) given by the Ebers–Moll model.

The base current I_B, which is related to the base charges and base recombination current, can be expressed by

$$I_B = \frac{dQ_b}{dt} + I_{rB}. \tag{14.62}$$

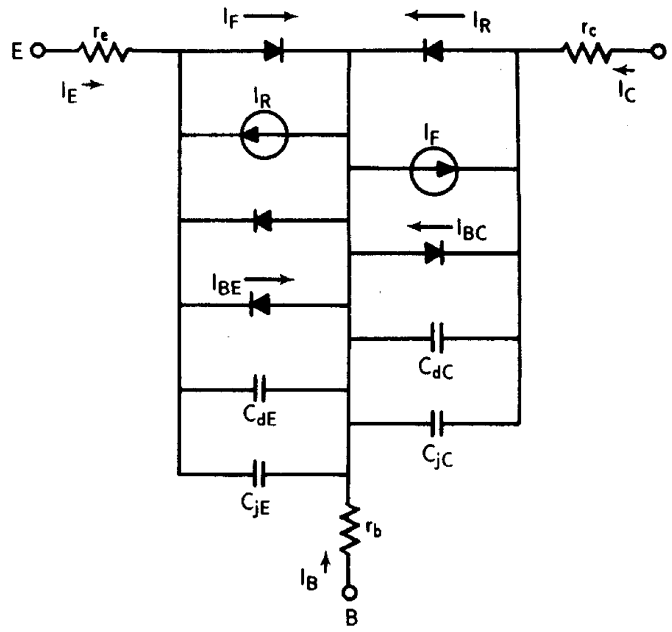


FIGURE 14.12. Equivalent circuit diagram of a p-n-p BJT based on the Gummel–Poon model: C_{jE} and C_{jC} denote the emitter and collector depletion capacitances, C_{dE} and C_{dC} are the emitter and collector diffusion capacitances, while r_e, r_h, and r_c are the emitter, base, and collector resistances, respectively. After Gummel and Poon,[2] by permission.

Here I_{rB} denotes the base recombination current, which consists of two components given by

$$I_{rB} = I_{EB} + I_{CB}, \tag{14.63}$$

where

$$I_{EB} = I_1 \left(e^{qV_{EB}/k_BT} - 1\right) + I_2 \left(e^{qV_{EB}/n_e k_B T} - 1\right), \tag{14.64}$$

$$I_{CB} = I_3 \left(e^{qV_{CB}/n_c k_B T} - 1\right). \tag{14.65}$$

Here I_{EB} is the emitter part of the base current and I_{CB} is the collector part of the base current, n_e and n_c denote the diode ideality factors for the E-B and C-B junctions, respectively. Values of n_e and n_c may vary between 1 and 2, depending on whether the diffusion or the recombination current is the dominant component in the base region. Thus, referring to Figure 14.12, the total emitter and collector currents can be written separately as

$$I_E = I_{CC} + I_{EB} + \tau_F \frac{dI_F}{dt} + C_{jE} \frac{dV_{EB}}{dt}, \tag{14.66}$$

$$I_C = I_{CC} - I_{CB} - \tau_R \frac{dI_R}{dt} + C_{jC} \frac{dV_{CB}}{dt}. \tag{14.67}$$

Figure 14.12 presents the equivalent circuit diagram for the Gummel–Poon model, which includes the junction depletion and diffusion capacitances of the

E-B and B-C junctions as well as series resistances r_e, r_h, and r_c.[(2)] Since Q_b is voltage dependent, the effect of high injection in the base (i.e., $\tau_F I_F \ll Q_{b0}$) is included. The Early effect is also included in the model by incorporating the voltage dependence of the collector charge Q_{jC} ($= C_{jC} V_{CB}$). The emitter part of the base current I_{EB} is represented by two diodes connected in parallel, one ideal and one with a diode ideality factor greater than 1 (i.e., to account for the bulk or surface recombination current), which makes the current gain bias-dependent at low current levels. Other effects, such as current-induced base push-out (i.e., the *Kirk* effect), can be incorporated into the model by adding a multiplication factor B to the $\tau_F I_F$ term given by (14.58). Therefore, the Gummel–Poon model is indeed a very accurate device model for predicting large-signal dc or transient behavior in a BJT. It allows one to predict the device terminal characteristics with good physical insight over a wide range of transistor operation. For a complete description of this device model, refer to the original paper published by Gummel and Poon.[2]

14.6. Switching and Frequency Response

As pointed out earlier, depending on the biasing conditions and modes of operation, a BJT can be operated either as an amplifier or as a switching device. In general, the BJTs are operated in the active mode only in linear or analog circuits. However, in digital circuits all four modes of operation may be involved. In this section, we discuss the switching properties and frequency response of a BJT.

When a BJT is operating as a switching device, the transistor has to change its bias condition from the low-current, high-voltage state (off) to the high-current, low-voltage (on) state within a very short period of time (e.g., tens of nanoseconds or shorter). Figure 14.13 shows the operation regimes and switching modes of a BJT.[5] The switching behavior of a BJT is seen to be a large-signal transient phenomenon. Since the switching speed is a key parameter in the operation of a switching transistor, one must include the junction depletion and diffusion capacitances, as well as the base-spreading resistance in the Ebers–Moll model shown in Figure 14.11. The junction depletion capacitance is important under reverse-bias conditions, while the diffusion capacitance becomes dominant under forward-bias conditions. The diffusion capacitance is related to the excess carrier stored charge in the transistor. In the active mode this charge is stored in the base, but in the saturation mode a large part of this charge is stored in the collector region. Nonlinear computer programs like SPICE are available for computer-aided simulation of digital bipolar transistors for all regions (i.e., cutoff, active, and saturation) of operation.

A switching transistor can be operated in several different modes. The saturation mode and current mode are the two most commonly used modes of operation for switching applications. Figure 14.13 shows these two basic modes of operation and their corresponding load lines. If the transistor is used as a current switch in digital circuits, it is always operated in the common-emitter configuration. In this

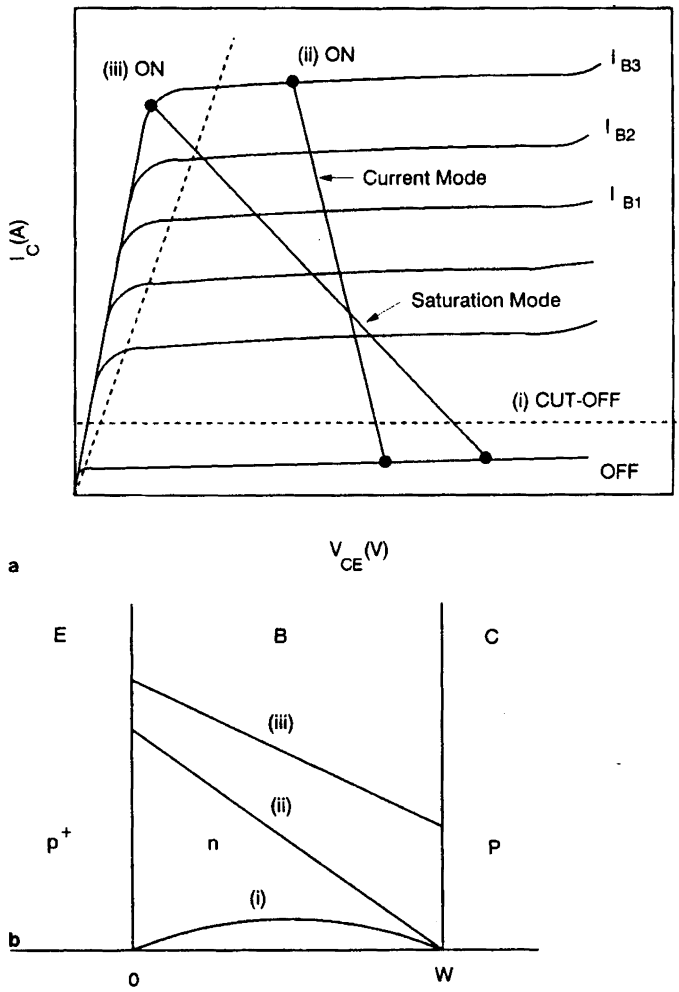


FIGURE 14.13. (a) Operation regions and switching modes of a silicon switching transistor and (b) distribution of minority carrier densities in the base for (i) cutoff, (ii) active, and (iii) saturation modes. After Moll,[5] by permission, © IEEE–1954.

configuration, current amplification (h_{FE}) is achieved. As shown in Figure 14.13a, for the current switch mode, the large collector current I_C flowing through the load resistance is switched by controlling the smaller base current I_B at the input. The static "on" and "off" states can be analyzed using the modified Ebers–Moll model shown in Figure 14.14.

The operation of a switching transistor is determined by its output characteristic curve, as illustrated in Figure 14.13a. In the cutoff region, the collector current is off and both the emitter and collector junctions are reverse-biased. In the active region, the emitter junction is forward-biased and the collector junction is reverse-biased.

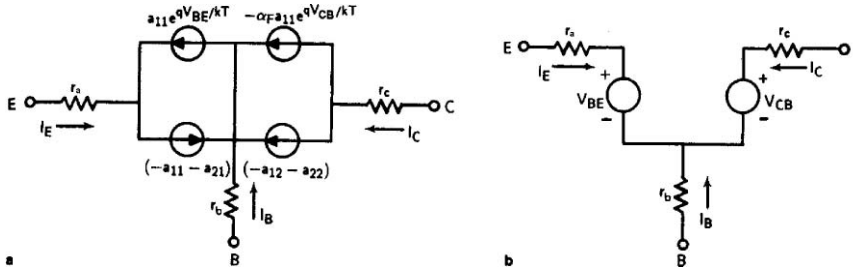


FIGURE 14.14. Equivalent circuit diagrams of a switching transistor (a) in regions (i) and (ii) and (b) in region (iii). After Ebers and Moll,[4] by permission, © IEEE–1961.

In the saturation region, the emitter and collector regions are both forward-biased. The minority carrier density distributions in the base region corresponding to (i) cutoff, (ii) active, and (iii) saturation regions are shown in Figure 14.13b.

The switch-off condition of a switching transistor for all switching modes may be obtained by extending the load line into the cutoff regime of the transistor. Therefore, the operating mode of a switching transistor is determined mainly by the dc current at the switch-on condition and the location of the operating point. The most common mode of operation for a switching transistor is the saturation mode. The transistor is nearly open-circuited between the emitter and the collector terminals during the switch-off condition, and is short-circuited during the switch-on condition. The current-mode operation is suitable for high-speed switching applications, since the delay associated with the excursion of the transistor into the saturation regime is eliminated.

In the design of a switching transistor, two important factors must be considered: the switching time and the current gain. The switching time is normally controlled by the minority carrier lifetime, which controls the charge storage time in the base and the collector. For example, gold (Au) is commonly used in silicon switching transistors to shorten its switching time, because Au impurity introduces a mid-gap acceptor level ($E_{Au}^{-} = E_c - 0.55$ eV) in silicon. The Au acceptor center is known as the most effective recombination center in silicon. Thus, by doping silicon transistors with a high concentration of Au impurities, the minority carrier lifetime can be drastically reduced, and hence the switching speed of a Au-doped Si BJT can be greatly enhanced. Finally, it should be noted that the current gain of a switching transistor may be improved by lowering the doping density in the base region of the transistor.

The switching behavior of a BJT may be analyzed using the Ebers–Moll model discussed in the previous section. Using (14.53) and (14.54), the four coefficients a_{11}, a_{12}, a_{21}, and a_{22} may be related to the measurable parameters I_{EO}, I_{CO}, α_F, and α_R, which are given by

$$a_{11} = \frac{-I_{EO}}{(1-\alpha_F\alpha_R)}, \quad a_{12} = \frac{\alpha_R I_{CO}}{(1-\alpha_F\alpha_R)}, \quad a_{21} = \frac{\alpha_F I_{EO}}{(1-\alpha_F\alpha_R)}, \quad a_{22} = \frac{-I_{CO}}{(1-\alpha_F\alpha_R)}, \tag{14.68}$$

where I_{EO} denotes the reverse saturation current of the emitter junction with collector opened and I_{CO} is the reverse saturation current of the collector junction with emitter opened; α_F and α_R denote the forward and reverse common-base current gains, respectively. For switching operation, the collector junction is reverse-biased in the cutoff and active regimes, and (14.53) and (14.54) reduce to

$$I_E = \frac{-I_{EO}\, e^{qV_{EB}/k_BT}}{(1-\alpha_F\alpha_R)} + \frac{(1-\alpha_F)I_{EO}}{(1-\alpha_F\alpha_R)}, \tag{14.69}$$

$$I_C = \frac{\alpha_F I_{EO}\, e^{qV_{EB}/k_BT}}{(1-\alpha_F\alpha_R)} + \frac{(1-\alpha_R)I_{CO}}{(1-\alpha_R\alpha_R)}. \tag{14.70}$$

The equivalent circuit of a switching transistor described by (14.69) and (14.70) is shown in Figure 14.14a.[4] It is noted that the emitter resistance r_e, base resistance r_b, and collector resistance r_c are included in the equivalent circuit shown in Figure 14.14a to account for the finite resistances in each region of the transistor. As for the saturation regime, both the emitter and collector junctions are under forward-bias conditions, and the C-B and E-B junction voltages can be derived from (14.53) and (14.54) in terms of the emitter and collector currents. This yields

$$V_{EB} = \left(\frac{k_BT}{q}\right) \ln[-(I_E + \alpha_R I_C)/I_{EO} + 1], \tag{14.71}$$

$$V_{CB} = \left(\frac{k_BT}{q}\right) \ln[-(I_C + \alpha_F I_E)/I_{CO} + 1]. \tag{14.72}$$

Figure 14.14b shows the equivalent circuit of a switching transistor operating in the saturation regime. Equations (14.69) through (14.72) may be used to analyze the nonlinear large-signal switching characteristics of a switching BJT.

In order to characterize a switching transistor, several key parameters such as the current-carrying capability, maximum open-circuit voltage, on and off impedances, as well as the switching time must be considered. The current-carrying capability is determined by the maximum power dissipation allowed in the transistor. The maximum open-circuit voltage is determined by the breakdown or punch-through voltage. The impedance during on and off conditions can be determined from (14.69) through (14.72) using proper boundary conditions. For example, for a common-base configuration, the on and off impedances of the transistor are given respectively by

$$Z_C(\text{on}) = \frac{V_C}{I_C} = \left(\frac{k_BT}{qI_C}\right) \ln[-(I_C + \alpha_F I_E)/I_{CO}], \tag{14.73}$$

$$Z_C(\text{off}) = \frac{V_C}{I_C} = \frac{V_C(1-\alpha_F\alpha_R)}{I_{CO} - \alpha_F I_{EO}}. \tag{14.74}$$

Equation (14.73) shows that the *on*-state impedance varies inversely with the collector current. The *on*-state impedance is very small when the collector current is large. On the other hand, the *off*-state impedance is very large when the reverse saturation currents I_{EO} and I_{CO} are small.

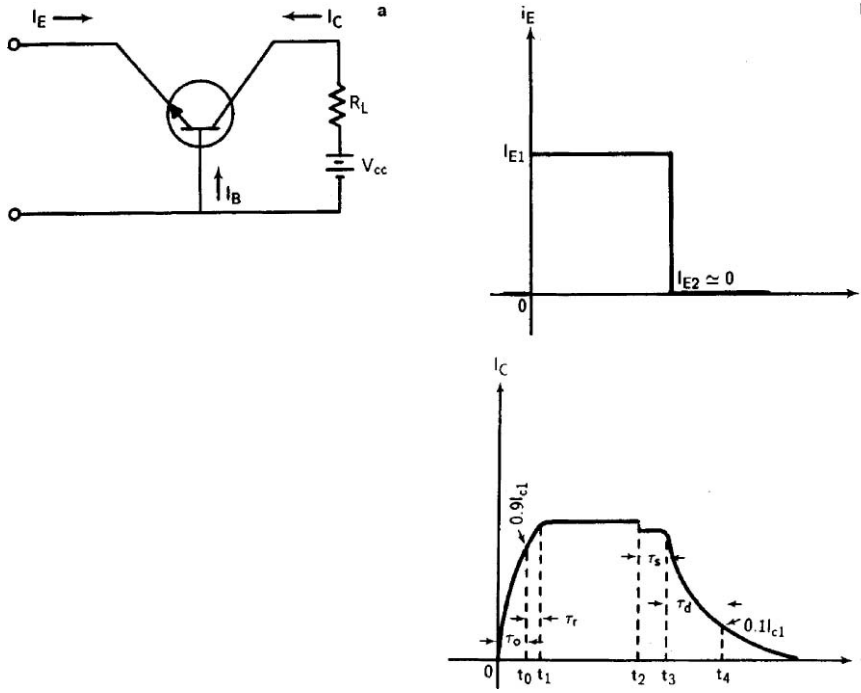


FIGURE 14.15. (a) Circuit diagram of an n-p-n switching transistor, (b) input emitter current pulse, and (c) collector output current response. Here τ_0 is the turn-on delay time, τ_r is the rise time, τ_s is the storage time, and τ_d is the decay time. After Moll,[5] by permission, © IEEE–1954.

Let us analyze the switching behavior of a transistor switch. Figure 14.15a shows the circuit diagram of an n^+-p-n BJT operating in the common-base configuration.[5] The transistor is assumed to be driven by a square current pulse from the emitter terminal whose waveform is shown in Figure 14.15b. The corresponding output collector current response is shown in Figure 14.15c. In the time interval from $t = 0$ to $t = t_1$, the transistor is turned on and the transient is determined by the transistor parameters in the active regime. At time t_1, the operating point of the transistor is in the saturation regime. The time required for the current to reach 90% of its saturation current (i.e., $I_{C1} = V_{CC}/R_L$) is called the turn-on time τ_0. At time t_2, the emitter current is reduced to zero (i.e., $I_{E2} \approx 0$) and the turnoff transient begins. From time t_2 to t_3 the minority carrier density in the base region is large. This corresponds to operation in regime III, except that the minority carrier density decays toward 0. During time τ_1, the collector is in the low-impedance state, and the collector current is determined mainly by the external circuit parameters. At time t_3, the carrier density near the collector junction is close to zero. At this point, the collector junction impedance increases rapidly and the transistor begins to operate in the active regime (II). The time interval τ_1

is called the carrier storage time. After time t_3, the transient behavior is calculated from the active regime parameters. At time t_4, the collector current has decayed to 10% of its peak value. The time between t_3 and t_4 is called the decay time τ_d.

The turn-on time τ_0 can be determined from the transient response in the active regime. From a step input current pulse I_{E1}, the Laplace transform is given by I_{E1}/S. If the common-base current gain is expressed in terms of $\alpha_F/(1 + j\omega/\omega_N)$, where ω_N is the alpha cutoff frequency at which $\omega/\omega_N = 0.707$, then the Laplace transform of the current gain is equal to $\alpha_F/(1 + s/\omega_N)$. Thus, the Laplace transform of the collector current can be expressed as

$$I_C(s) = \frac{\alpha_F I_{EI}}{(1 + s/\omega_N)}, \tag{14.75}$$

and the inverse transform of (14.75) can be written as

$$I_C(t) = \alpha_F I_{E1}(1 - e^{-\omega_N t}). \tag{14.76}$$

If one sets $I_{C1} = V_{CC}/R_L$ as the saturation value of the collector current, then τ_0 is obtained by setting $I_C = 0.9I_{C1}$ in (14.76), which yields

$$\tau_0 = \left(\frac{1}{\omega_N}\right) \ln\left[\frac{I_{E1}}{(I_{E1} - 0.9I_{C1}/\alpha_F)}\right], \tag{14.77}$$

where τ_0 is the time constant for the collector current to reach 90% of its peak value. Similarly, the storage time τ_1 and decay time τ_d for the common-base configuration can be written, respectively, by

$$\tau_1 = \frac{(\omega_N + \omega_I)}{\omega_N \omega_I (1 - \alpha_F \alpha_R)} \ln\left[\frac{(I_{E1} - I_{E2})}{(I_{C1}/\alpha_N - I_{E2})}\right], \tag{14.78}$$

$$\tau_d = \left(\frac{1}{\omega_N}\right) \ln\left[\frac{(I_{C1} - \alpha_F I_{E2})}{(0.1I_{C1} - \alpha_F I_{E2})}\right], \tag{14.79}$$

where ω_1 is the inverted alpha cutoff frequency, while I_{E1} and I_{E2} (≈ 0) are the peak and bottom of the emitter input current pulse. It is seen that the turnoff time is equal to the sum of τ_1 and τ_d. From (14.78) and (14.79) it is noted that both switching times (i.e., turn-on time τ_0 and turnoff time $\tau_1 + \tau_d$) are inversely proportional to the cutoff frequency of the transistor. Thus, in order to increase the switching speed, one must increase the cutoff frequency of the transistor. Since the cutoff frequency for most switching transistors is limited by the collector storage capacitance, it is important that this capacitance be kept at its minimum value.

The dc characteristics of a BJT were described in Section 14.3. We now discuss the ac characteristics and frequency response of a BJT when a small ac signal voltage or current is superimposed on the dc value. If a BJT is operating as an amplifier in common-emitter configuration, then the E-B junction is under forward-bias and the C-B junction is under reverse-bias conditions. The equivalent circuit of the BJT under low-frequency operation is shown in Figure 14.16a, where v_{EB} denotes the ac voltage applied to the E-B junction, $g_{EB}(= i_B/v_{EB})$ is the input conductance, and $g_m(= i_c/v_{EB})$ is the transconductance. At higher frequencies, additional circuit

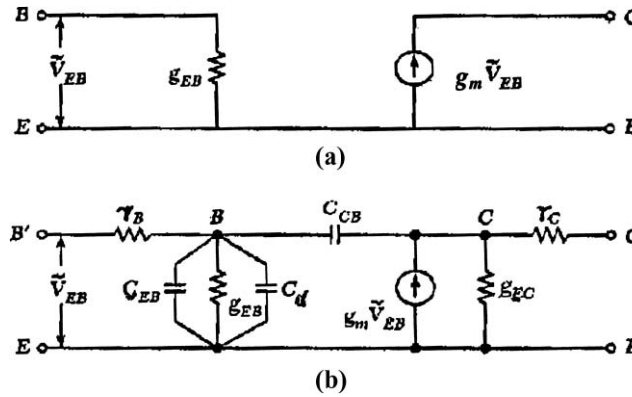


FIGURE 14.16. (a) The equivalent circuit of a BJT under low-frequency operation, and (b) the high-frequency equivalent circuit with added capacitance and conductance to (a).

elements should be added to account for the base-width modulation effect (i.e., finite output conductance $g_{EC}\,(=i_c/v_c)$), the depletion capacitance and diffusion capacitances, base resistance r_B, and collector resistance r_C. Figure 14.16b shows the high-frequency equivalent circuit with added capacitances and conductances to Figure 14.16a. It is noted that the transconductance g_m and input conductance g_{EC} are dependent on common base current gain, α. At low frequency, α is a constant $(=\alpha_0)$ and independent of operating frequency. However, the value of α will decrease after reaching a critical frequency. The frequency dependence of the common-base current gain α can be expressed by

$$\alpha = \frac{\alpha_0}{1 + j(f/f_\alpha)}, \tag{14.80}$$

where α_0 is the dc common-base current gain ($\approx$1) and f_α is the common-base cutoff frequency. At $f = f_\alpha$ the magnitude of α reduces to $0.707\alpha_0$ (3 dB down). The common-emitter current gain β can be related to α by the following expression:

$$\beta = \frac{\alpha}{1-\alpha} = \frac{\beta_0}{1 + j(f/f_\beta)}, \tag{14.81}$$

where f_β is the common-emitter cutoff frequency ($f_\beta = (1-\alpha_0)f_\alpha$), which is much smaller than f_α. Another cutoff frequency f_T known as the unit gain (i.e., $\beta = 1$) cutoff frequency of the BJT can be related to f_β and f_α by

$$f_T = f_\beta\left(\beta_0^2 - 1\right)^{1/2} = \beta_0(1-\alpha_0)f_\alpha \approx \alpha_0 f_\alpha, \tag{14.82}$$

which shows that f_T is slightly smaller than f_α. The cutoff frequency f_T can also be expressed as $(2\pi\tau_T)^{-1}$, where τ_T represents the total time of carrier transit from the emitter to the collector, which includes the emitter delay time τ_E, the base transit time τ_B, and the collector transit time τ_C. The base transit time τ_B due to

hole drift through the base region can be expressed by

$$\tau_B = \frac{W^2}{2D_p}. \tag{14.83}$$

Equation (14.83) is valid, provided that the recombination loss in the base region is negligible. It is noted that to improve frequency response of the BJTs, the transit time of the minority carriers across the base region must be short. Therefore, high-frequency BJTs are designed with a very small base width. For high-frequency silicon BJTs, n-p-n structures, are preferred over p-n-p structures, since the electron diffusion length is much larger than the hole diffusion length (i.e., $L_n \approx 3L_p$). Another approach to reduce the base transit time is to use a graded base with a built-in field to assist minority carriers to move across the base faster toward the collector and hence reduce the base transit time.

14.7. Advanced Bipolar Junction Transistors

In order to increase the current gain of a BJT, the emitter region is usually doped very heavily and the base region is kept very thin. Most silicon BJTs are fabricated using the polysilicon-emitter structure heavily doped in situ with phosphorous impurities on the lightly doped ion-implanted base at a temperature low enough (630°C) to prevent dopant diffusion. Common-emitter current gains in excess of 10^4 and emitter Gummel number greater than 10^{14} cm^{-4} have been achieved in silicon BJTs with polysilicon emitter. Other emitter structures such as the MIS tunnel junction emitter transistor with very high current gain have also been reported in the literature. Polysilicon has been widely used in bipolar technology for the emitter and base contacts in advanced silicon BJTs. Vertical scaling of silicon BJTs can be greatly simplified using polysilicon emitter contacts, since the base saturation current and emitter junction depth can be effectively reduced. The self-aligned polysilicon-emitter BJT is rapidly becoming the dominant bipolar structure used in very large scale integrated circuits (VLSIs). The advantages of using the polysilicon-emitter structure over the conventional metal-emitter contact structure include superior process yields, higher packing densities, and better device performance.

Fabrication of BJTs using the polysilicon-emitter structure is quite different from that of the conventional metal-contacted emitter BJTs. For example, after the emitter window is opened, a polysilicon layer is deposited and doped by ion implantation or, alternatively, an in situ doped polysilicon layer is deposited onto the underlying emitter region. This polysilicon layer serves as the emitter contact and at the same time as a dopant source for the underlying emitters during postimplant activation annealing. Process yields are enhanced because the polysilicon layer prevents implantation damage from the underlying emitter. Figure 14.17 shows a self-aligned silicon BJT with a polysilicon-contacted emitter.[6] The polysilicon emitter is formed by arsenic implantation, and the polysilicon is selectively etched to form the emitter contact. The structure is then oxidized, resulting in a thicker

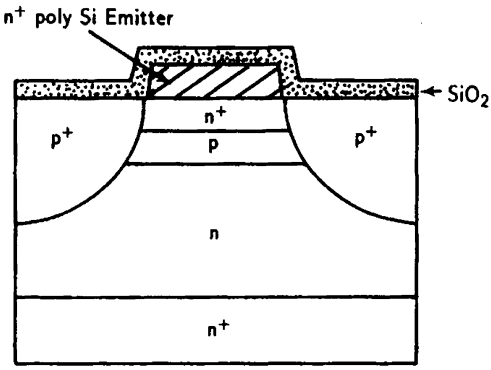


FIGURE 14.17. Cross-section view of a self-aligned silicon BJT with polysilicon emitter contact. After Cuthbertson and Ashburn,[6] by permission, © IEEE–1985.

oxide layer over the polysilicon and a thinner oxide layer over the silicon. The p^+-base contact region is formed using high-energy boron implantation. A high-temperature diffusion step is used to produce the emitter region and the extrinsic base region. Packing densities are increased substantially by realizing the self-aligned structures. Device performance is greatly improved by the self-aligned structure and reduction of the base current, since the former reduces the device parasitics while the latter is traded for low base resistance. As a result, the speed–power performance of a polysilicon-emitter BJT is improved substantially. Using polysilicon-emitter BJTs, emitter-coupled logic (ECL) circuits with propagation delay times in the sub-100 ps have been reported.

14.8. Thyristors

When an extra p-n junction is added to a p-n-p or an n-p-n BJT, an n^+-p-n-p or p^+-n-p-n four-layer thyristor is formed. A thyristor is a semiconductor device that exhibits bistable characteristics and can be switched between a low-impedance, high-current on-state condition to a high-impedance, low-current off-state condition. The operation of a thyristor is very similar to the operation of a BJT in that both electrons and holes participate in the transport process. Typical doping densities in a p_1^+-n_1-p_2-n_2^+ four-layer structure are 10^{19} cm^{-3} for the p_1^+ region, 5×10^{14} cm^{-3} in the n_1 region, 10^{16}–10^{17} cm^{-3} in the p_2 region, and 10^{19} cm^{-3} in the n_2^+ region.

The schematic diagrams of a p-n-p-n device with two, three, and four terminals are shown in Figures 14.18 a–c, respectively. The device consists of three junctions, J_1, J_2, and J_3 (i.e., p_1^+-n_1, n_1-p_2, and p_2-n_2^+), in series. The contact electrode connected to the outer p_1 layer is called the anode, and the contact electrode connected to the outer n_2^+ layer is called the cathode. Figure 14.18a shows the two-terminal p-n-p-n diode with the gate terminal opened. If a gate electrode is connected to the inner p_2 layer to form a three-terminal p^+-n-p-n^+ device, then the device is called a semiconductor-controlled rectifier or a thyristor. This is shown in Figure 14.18b. An additional gate electrode may be connected to the inner n_1 layer

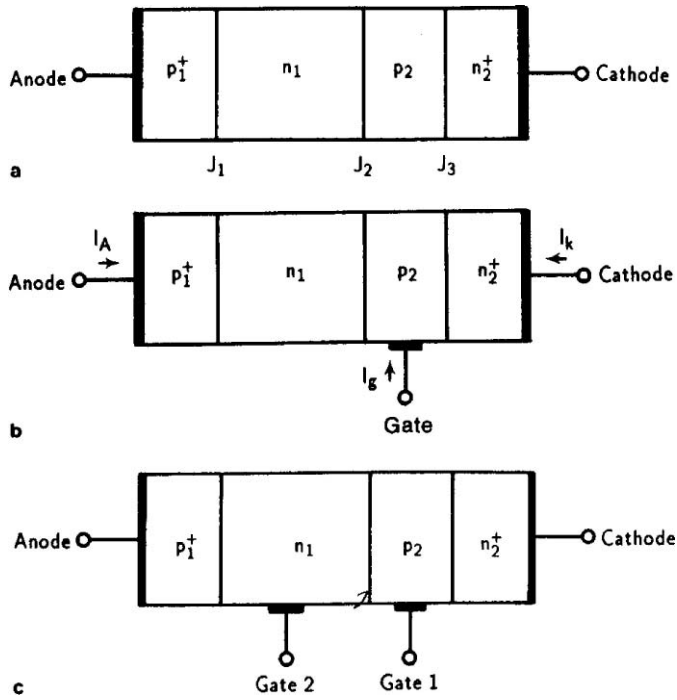


FIGURE 14.18. (a) Schematic diagrams of a two-terminal p^+-n-p-n diode, (b) three-terminal thyristor (SCR) with a controlled gate, and (c) four-terminal p-n-p-n device with two controlled gates. The device has three junctions, J_1, J_2, and J_3, in series. The current gain α_1 is for the p-n-p transistor, and α_2 is for the n-p-n transistor. Under the forward-blocking condition, the center junction J_2 is reverse-biased and serves as a common collector for the p-n-p and n-p-n transistors.

of a p^+-n-p-n diode with two gate electrodes, as shown in Figure 14.18c. If no gate electrode is provided, then the device is operated as a two-terminal p^+-n-p-n^+ Shockley diode as shown in Figure 14.18a.

The current–voltage (I–V) characteristics of a typical p-n-p-n thyristor are shown in Figure 14.19a. It is noted that there are four distinct regions shown in this plot. In region I ($0 \Rightarrow 1$) at low-bias voltages, junctions J_1 and J_3 are forward-biased and junction J_2 is reverse-biased. Therefore, the external voltage drop is almost entirely across the J_2 junction, and the device behaves like a reverse-biased p-n junction diode. In this region, the device is in the forward-blocking or high-impedance, low-current off state. In region I, the forward breakover occurs when $dV/dI = 0$, and a breakover voltage V_{bo} and a switching current I_s can be defined in this region. These parameters are shown in Figure 14.19a. Region II ($1 \Rightarrow 2$) is the negative differential resistance region in which the current decreases with increasing applied voltage. In region III ($2 \Rightarrow 3$), the current increases rapidly as the applied voltage increases slowly. In this region, junction J_2 is forward-biased, and the voltage drop

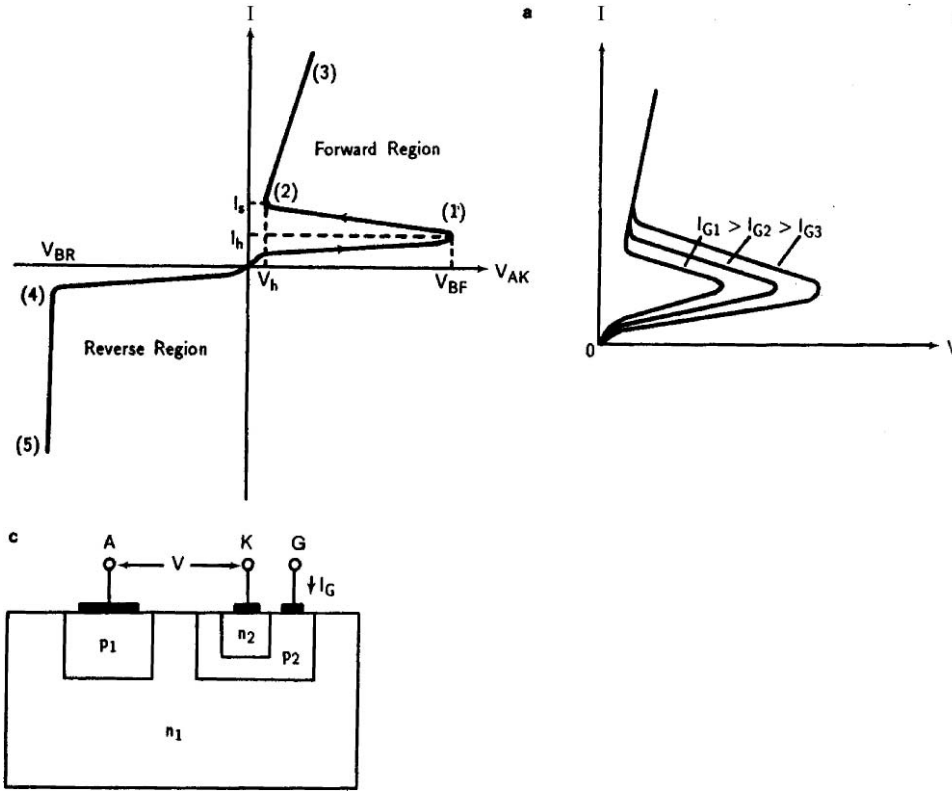


FIGURE 14.19. (a) Current–voltage ($I-V$) characteristics of a thyristor showing the forward and reverse regions. Region (1) forward-blocking or "off" state (high impedance, low current); (2) negative resistance regions; (3) forward-conducting or "on" state (low impedance, high current); (4) reverse-blocking state; (5) reverse-breakdown region. (b) The effect of gate current on the current–voltage characteristics of a thyristor (SCR). (c) A low-power SCR device structure.

across the device is that of a single p-n junction diode. The device is in the low-impedance, high-current on state. When the current flow in the diode is reduced, the device will remain in the on state until it reaches a current level I_h. The current I_h and its corresponding voltage V_h are called the holding current and holding voltage, respectively. When the current drops below I_h, the diode switches back to its high-impedance state and the cycle repeats. If a negative-bias voltage (in region $0 \Rightarrow 4$) is applied to the p_1^+ terminal and a positive voltage to the n_2^+ terminal, then both J_1 and J_3 junctions become reverse-biased. Zener or avalanche breakdown may occur when the applied reverse-bias voltage is large enough to cause the breakdown of junctions J_1 and J_3. This region is usually avoided in thyristor operation. A thyristor operating in the foward-bias region is thus a bistable device that can switch from a high-impedance, low-current off state to a low-impedance, high-current on state.

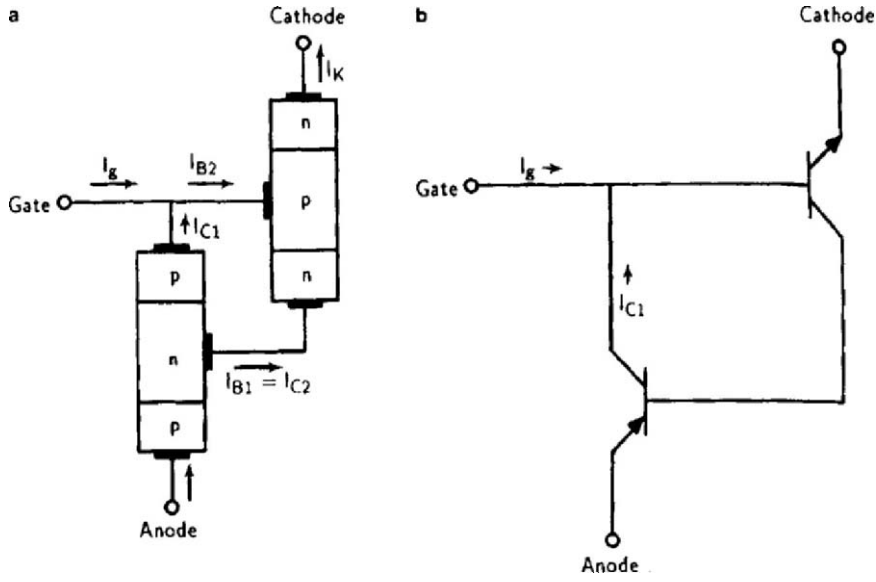


FIGURE 14.20. (a) Two-transistor approximation of a three-terminal thyristor. (b) Same as (a) using transistor symbols.

Thyristors are the most widely used four-layer p-n-p-n devices with applications ranging from speed control in home appliances to switching and power inversion in high-voltage transmission lines. In p-n-p-n diode operation, it is necessary to increase the external applied voltage so that junction J_2 is in the avalanche multiplication region. The breakover voltage of a p-n-p-n diode is fixed during fabrication. However, the shape of the *I–V* characteristic curve can be controlled by using a third terminal or a gate in the p_2 region of the thyristor as shown in Figure 14.19c. A thyristor can be fabricated using standard silicon planar technology; the p_1 and p_2 regions are formed using thermal diffusion (or implantation) of boron dopant followed by diffusion of phosphorus impurity to form the n_2 region to complete the four-layer p-n-p-n structure as shown in Figure 14.17. The p_1-n_1-p_2 structure is known as the lateral transistor and the n_2-p_2-n_1 structure as the vertical transistor. The gate electrode is connected to the p_2 region to control the *I–V* characteristics of the thyristor. Figure 14.19b shows the *I–V* characteristics of a typical silicon-controlled rectifier (SCR) under different gate currents I_g.

The predominant effects of increasing I_g in an SCR device are an increase in the off current and a decrease in both the breakover voltage and the holding current. These effects can be explained qualitatively in terms of the two-transistor equivalent circuit shown in Figures 14.20a and b. In the off state, the device behaves essentially like a normal n-p-n transistor with a p-n-p transistor acting as an emitter follower having a very small forward current gain. Increasing the gate current I_g will increase the collector current and hence the anode current of the n-p-n transistor. The larger anode current will result in an increase of the transistor

current gain. When $\alpha_1 + \alpha_2 = 1$, the avalanche multiplication factor decreases and the breakover voltage decreases. In the on state, the flow of gate current will again increase the value of α. Thus, the holding current can reach a lower value before it switches back to the off state.

In general, the basic *I–V* characteristics of a thyristor can be best explained using a two-transistor analogue developed by Ebers in which the n base of a p-n-p transistor is connected to the emitter of an n-p-n transistor to form a four-layer p-n-p-n device. Figure 14.20a shows a three-terminal thyristor and Figure 14.20b is its equivalent circuit representation. It is noted from Figure 14.20b that the collector current of the n-p-n transistor provides the base drive for the p-n-p transistor, while the collector current and the gate current of the p-n-p transistor supply the base drive for the n-p-n transistor. Thus, a regeneration condition occurs when the total loop gain is greater than one. The base current I_{B1} of the p-n-p transistor is equal to the collector current I_{C2} of the n-p-n transistor, and is given by

$$I_{B1} = I_{C2} = (1 - \alpha_1)I_A - I_{CO1}, \tag{14.80}$$

where I_A is the anode current of the p-n-p transistor and α_1 is the dc common-base current gain. The collector current of the n-p-n transistor is given by

$$I_{C2} = \alpha_2 I_K + I_{CO2}, \tag{14.81}$$

where α_2 is the dc common-base current gain and $I_K = I_A + I_g$ is the cathode current of the n-p-n transistor. Solving (14.80) and (14.81), one obtains

$$I_A = \frac{(\alpha_2 I_g + I_{CO1} + I_{CO2})}{(1 - \alpha_1 - \alpha_2)}, \tag{14.82}$$

which predicts the dc characteristics of a thyristor up to the breakover voltage, and the device behaves like a p-i-n diode beyond the breakover voltage. It is noted that all the current components in the numerator of (14.82) are very small except when $\alpha_1 + \alpha_2$ approaches unity. At this point, the denominator of (14.82) becomes zero, and I_A increases without limit. As a result, the forward breakover or switching takes place when $dV_{AK}/dI_A = 0$. The transistor current gain is seen to increase with collector voltage and collector current at low current level. The effect of collector voltage on α is particularly pronounced as it approaches the avalanche voltage. Therefore, as the voltage across the thyristor increases, the collector current and values of α in the two equivalent transistors will also increase. When $\alpha_1 + \alpha_2$ approaches unity, I_A increases sharply, which, in turn, increases the value of α. When the sum of avalanche-enhanced α value is equal to one (i.e., $\alpha_1 + \alpha_2 = 1$), breakover will occur. Because of the regenerative nature of these processes, the device is eventually switched to its on state. Upon switching, the current flowing through the thyristor must be limited by the external load resistance, or the device will be destroyed when the applied voltage becomes too large. In the on state, all three junctions of the device are forward-biased and normal transistor action is no longer effective. The voltage across the device is nearly equal to the sum of the three saturation junction voltage drops, on the order of 1 V for a silicon thyristor. In order to keep the device in its low-impedance on

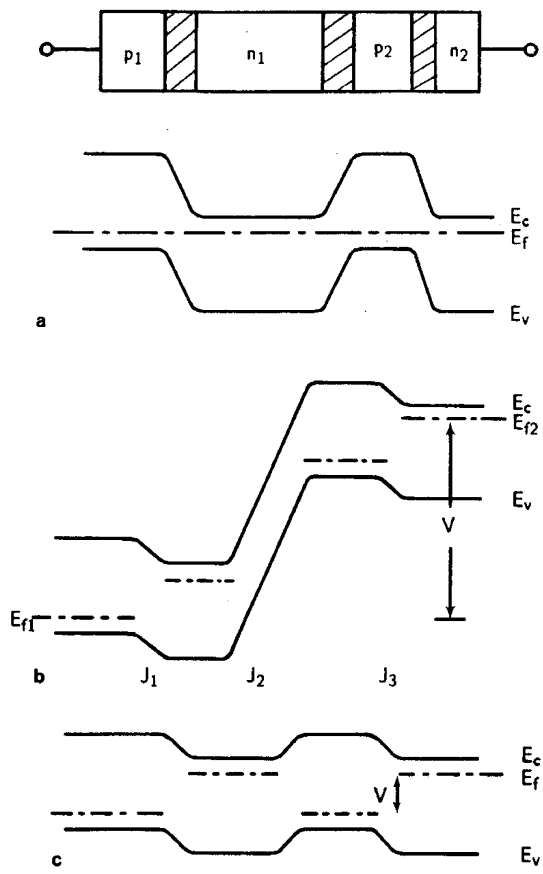


FIGURE 14.21. Energy band diagrams of a p-n-p-n diode in (a) equilibrium conditions, (b) forward off state, where most of the voltage drop is across the J_2 junction, (c) forward on state, in which all three junctions are forward-biased.

state, the condition that $\alpha_1 + \alpha_2 = 1$ must be satisfied. In this case, the holding current corresponds to the minimum current in which $\alpha_1 + \alpha_2 = 1$ is satisfied. Further reduction of current will result in the device being switched back to the high-impedance off state.

Figure 14.21 shows the energy band diagrams of a p-n-p-n thyristor under different bias conditions: Figure 14.21a is for equilibrium conditions, Figure 14.21b is for the forward off state, in which most of the voltage drop is across junction J_2, and Figure 14.21c is for the forward on state, in which all three junctions are forward-biased. In practice, when a positive voltage is applied to the anode to turn the thyristor from the off state to the on state, the junction capacitance across J_2 is charged. This charging current flows through the emitter junctions of the two transistors. If the rate of change of applied voltage with time is large, the charging current may be large enough to increase the α value of the two transistors sufficiently to turn on the device. This rate effect may reduce the breakover voltage to half or less than half of its static value. The voltage at which an SCR device goes from the "on" to the "off" state is usually controlled by a small gate signal. In a

low-power SCR, the gate electrode can be used to turn the device to the on and off states. However, for high-power SCRs, once the device is in the on state, the gate circuit has little effect on the device operation.

The SCR is usually a large-area device since it needs to handle a large amount of current. As a result, lateral gate current flow can give rise to a substantial voltage drop across the device, and the current-crowding effect tends to turn on the periphery of the device first. This turn-on condition may propagate through the entire device. During the turn-on transient, the anode current passes through the small peripheral area momentarily, and the high current density could cause the device to burn out. To prevent this problem, an interdigitated structure is often used to reduce the lateral effect.

14.9. Heterojunction Bipolar Transistors

14.9.1. Introduction

Heterojunction bipolar transistors (HBTs) are currently being used in a wide range of communications products including power amplifiers, voltage-controlled oscillators (VCO) and mixers, wireless security systems, wireless local area networks (WLAN), satellite communication systems, high-power radar transmit/receive (T/R) modules, and high-speed analog/digital (A/D) ICs. Recent development of high-speed HBT technologies are based on the InP/InGaAs, AlGaAs/GaAs, Si/SiGe, and GaInP/GaAs material systems. The GaInP/GaAs has recently emerged as an alternative to AlGaAs/GaAs material system for HBT power applications. The GaInP/GaAs system has a larger valence band discontinuity than AlGaAs/GaAs, which results in higher emitter injection efficiency and higher device gain. Furthermore, because of the smaller conduction band discontinuity of GaInP/GaAs, transport can be improved at higher current levels because the injected electrons from the emitter to the base at these levels are less likely to transfer to the L-valley of the GaAs base. Other distinct advantages of GaInP/GaAs over AlGaAs/GaAs are the absence of donor-related traps (DX centers) and the etching selectivity of GaInP with respect to GaAs.

The advantages offered by the InP HBT technology have opened up a range of new applications for high-speed mixed-signal and digital ICs. Modern digital communications, instrumentation, electronics warfare, and radar systems require high-speed digital and mixed-signal ICs operating at frequencies from DC to 100 GHz. Broadband requirements place severe constraints on available semiconductor technologies and design expertise. Indeed, commercial off-the-shelf digital and mixed-signal ICs based on SiGe or GaAs technologies are generally available only at speeds up to 13 GHz. For speeds over 13 GHz, InP HBT technology could pave the way for the development of 100 GHz digital and mixed signal ICs applications.

Silicon–germanium (SiGe) HBTs have received much attention since IBM refined its process and began offering a foundry 8-inch SiGe line for the

fabrication of Si/SiGe HBTs and ICs on silicon substrates. SiGe HBTs have found applications in many microwave and mixed-signal products, where they can offer high performance and cost-effective solutions that are not available on a silicon platform. However, the Si/SiGe HBT structure remains a low-power configuration. The high-frequency performance exhibited by SiGe HBTs is largely a result of decreased minority carrier transit time through the base layer. This is achieved by thinning the SiGe base layer, using a graded Ge_xSi_{1-x} base layer for the built-in field to push electrons across the base, and increasing the doping density in the base to lower the base resistance.

14.9.2. Device Structures and Fabrication Technology

In this section, the device structure, fabrication technology, operation principle, current–voltage (*I–V*) behavior, and performance characteristics such as current gain and cutoff frequency of a heterojunction bipolar transistor (HBT) are described. Device characteristics and applications of HBTs fabricated from Si/SiGe, GaAs/AlGaAs, and InGaAs/InAlAs material systems will be presented.

The concept of a heterojunction bipolar transistor (HBT) was first proposed by Shockley, and the basic device theory describing the operation principles of an HBT was subsequently developed by Kroemer. In spite of the great potential for using HBTs in high-speed digital and microwave circuit applications, the technology for fabricating HBTs did not exist until the 1970s. With the advances of MBE and MOCVD growth techniques of III-V epitaxial layers, significant progress in HBT device fabrication technology has been made in recent years, although it is still not as mature as the technology for FETs. For example, frequency divider circuits using AlGaAs/GaAs HBTs with clock frequencies exceeding 20 GHz and a maximum oscillation frequency f_{max} of 105 GHz have been reported recently. The main motivation for using the HBT structure is to overcome some of the limitations found in homojunction bipolar transistors (BJTs). The advantages of an HBT over a BJT include using (i) a wide-band-gap emitter to suppress minority carrier back injection, (ii) a lightly doped emitter to reduce the E-B junction capacitance, and (iii) a heavily doped base to lower the base resistance. As a result both the speed and frequency performance of the HBT can be significantly improved over that of the conventional BJT. In this section, we will present the device structure, operation principles, and dc characteristics of an AlGaAs/GaAs HBT.

In an HBT, the current path (i.e., speed-limiting factor) is perpendicular to the surface and the epilayers. Therefore, to first order, the speed of an HBT is governed mainly by the thickness of the epilayers. Since the epilayer thickness can be easily made much smaller by the MBE or MOCVD technique than the horizontal lithography dimensions, for a given horizontal dimension there is a higher speed potential for the HBT structure than for MESFETs. The HBT using an AlGaAs/GaAs material system has shown great promise for high-speed device applications. The use of a wide-band-gap $Al_xGa_{1-x}As$ emitter for the HBT results in an injection efficiency of close to unity even if the doping density in the GaAs base region is much higher than that of the emitter. This provides an extra degree of freedom in

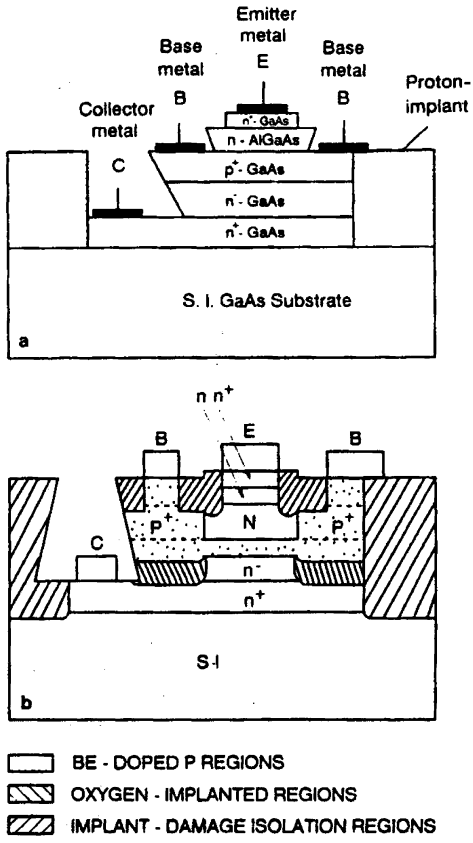


FIGURE 14.22. Schematic cross-sectional view of an AlGaAs/GaAs n-p^+-n-n^+ HBT: (a) with self-aligned process and (b) fabricated by ion-implanted process. After Asbeck,[8] by permission, IEEE–1988.

transistor design, which helps to achieve high-speed operation in such a device. Its major limitations include technological problems related to reproducible and stable processing and the device physics related to gain degradation mechanisms. Historically, the AlGaAs/GaAs emitter-up HBTs have been fabricated on emitter, base, and collector epilayers grown sequentially by the MBE or MOCVD technique, with ohmic contacts being made on the emitter, base, and collector regions by sequential etching. Etching through the emitter to the base and the E-B-n^- collector to the n^+ collector usually leads to steps in the GaAs surface ranging from 0.4 to 1.0 μm in depth. Although high-quality HBTs can be readily fabricated in this manner, the resulting mesa structure is a severe topographical obstacle to integrating these HBTs with a multilevel metal system into a densely packed integrated circuit. High levels of integration have been achieved with HBTs using a planar HI^2L technology, which relies on an emitter-down AlGaAs/GaAs structure with implanted base and extrinsic p^+ base regions. Other advantages of III-V HBTs over silicon BJTs include possible transient electron velocity overshoot, radiation hard, and compatibility with optoelectronic integrated circuits (OEICs).

Figure 14.22a shows a cross-sectional view of an AlGaAs/GaAs HBT fabricated with a self-aligned base process, and Figure 14.22b shows the AlGaAs/GaAs

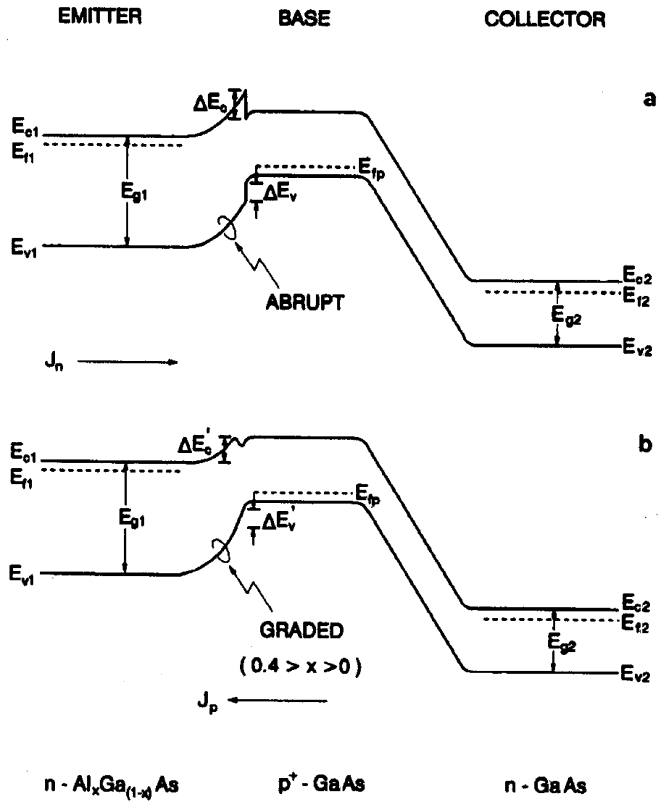


FIGURE 14.23. Energy band diagrams for an AlGaAs/GaAs HBT: (a) with an abrupt E-B junction and (b) with a graded E-B junction. The electron injection is from the wide-band-gap n-AlGaAs emitter region ($n \approx 5 \times 10^{17} cm^{-3}$) into the narrow band gap GaAs base region ($p \approx 10^{19}\ cm^{-3}$).

HBT fabricated by using an ion-implanted process. The advantages of an ion-implanted process include low base contact resistance, flexibility of layer structure, and low C-B capacitance, and the problems associated with this process are dopant diffusion, anneal uniformity, and parasitic base resistance. As for the self-aligned base process, the advantages include a simpler, faster, and low-temperature process, while etch control, higher base contact resistance, and lower current gain are some of the problems associated with this process. Figure 14.23a, b shows the energy band diagram of an AlGaAs/GaAs HBT with an abrupt E-B junction, and Figure 14.23b is the energy band diagram for an AlGaAs/GaAs HBT with a graded E-B junction. The effects of using the graded E-B junction shown in Figure 14.23b include (i) reducing space-charge recombination in the E-B junction, (ii) increasing injection electron velocity, (iii) being less effective in suppressing hole injection from the base to the emitter, and (iv) being more susceptible to base dopant diffusion.

The typical dopant densities and layer thicknesses for the AlGaAs/GaAs HBT structure shown in Figures 14.22a and b are as follows: The device structure consists of a 0.2 μm heavily doped (3×10^{18}cm^{-3}) n$^+$-GaAs cap layer grown on top of the wide-band-gap $Al_{0.3}Ga_{0.7}As$ emitter layer to reduce the emitter contact resistance, a 0.1 μm $Al_{0.3}Ga_{0.7}As$ emitter layer of dopant density around 5×10^{17}cm^{-3}, a 0.1 μm p$^+$-type GaAs base layer with dopant density 1×10^{19} cm^{-3}, and a 0.3 μm n-type GaAs collector layer with dopant density 10^{17} cm^{-3} grown on top of an n$^+$-GaAs buffer layer with dopant density 3×10^{18} cm^{-3}. These GaAs/AlGaAs active layers were grown on a semi-insulating GaAs substrate using the MBE or MOCVD technique.

14.9.3. Current Gain and Device Parameters

The current gain expression for an HBT can be derived using the theory developed by Kroemer for a wide-band-gap emitter HBT.[7] For example, the electron current injected from emitter to base (i.e., J_n) and the hole current (J_p) injected from base to emitter for an n-p-n AlGaAs/GaAs graded E-B junction HBT shown in Figure 14.23b can be expressed as

$$J_n = q\left(\frac{D_n}{W_B}\right)\left[N_E \exp\left(-\Delta E_c'/k_B T\right)\right], \quad J_p = q\left(\frac{D_p}{W_E}\right)\left[N_B \exp\left(-\Delta E_v'/k_B T\right)\right]. \tag{14.83}$$

From (14.83) one can estimate the maximum current gain for the HBT shown in Figure 14.23b from the ratio of the electron and hole current density, which is given by

$$\beta_{max} \approx \frac{J_n}{J_p} \approx \frac{N_E v_{nB}}{N_B v_{pE}} \exp\left[-\left(\Delta E_c' - \Delta E_v'\right)/k_B T\right] = \frac{N_E v_{nB}}{N_B v_{pE}} \exp\left(\Delta E_g/k_B T\right). \tag{14.84}$$

From (14.84) it is seen that a very high value of β_{max} can be achieved even when N_E is smaller than N_B. To obtain a current gain of $\beta > 100$ in the AlGaAs/GaAs HBT structure shown in Figure 14.23b with a base-to-emitter dopant density ratio (N_B/N_E) of 50 to 100, the value of $\Delta E_v' - \Delta E_c' = \Delta E_g$ should be equal to or greater than 0.24 eV, which corresponds to a wide-band-gap $Al_{0.22}Ga_{0.78}$ As (i.e., with 22% of AlAs) emitter. A typical AlAs mole fraction used in an AlGaAs/GaAs HBT is about 25%. It is noted that $\Delta E_c'(= \Delta E_g' + \Delta E_v')$ is the band gap discontinuity in the valence band edge of the wide-band-gap AlGaAs emitter. It is clear that a substantial increase in current gain may be achieved in an HBT as a result of the exponential increase of β with $\Delta E_v'$. A decrease in β_{max} due to the decrease of potential barrier for holes may be partially compensated by the increase of electron velocity in the base caused by the ballistic injection of electrons from the spike-notch structure at the conduction band edge of the wide-band-gap emitter near the E-B junction (see Figure 14.23a). Smoothing out the conduction band

spike can be achieved by grading the composition of the wide-band-gap emitter near the heterointerface of the E-B junction, as shown in Figure 14.23b.

Since the emitter injection efficiency of an HBT can be made very high, its current gain is essentially equal to the base transport factor. For an n-p^+-n HBT, this is given by

$$\beta \approx \frac{\tau_n}{\tau_B}, \tag{14.85}$$

where τ_n and τ_B denote the electron lifetime in the base and the transit time across the base, respectively. For a uniformly doped base the electron transport across the base is by diffusion and $\tau_B \approx W_B^2/2D_n$, while for a graded composition base the transport of electrons in the base is by drift and $\tau_B \approx W_B/\mu_n\mathcal{E}$. Thus, in order to obtain a high current gain, τ_B should be as small as possible. For example, for a sufficiently short base HBT with $W_B = 0.1$ μm, $\beta > 10^3$ can be obtained even if τ_n in the base is on the order of 1 ns. From (14.85) it is interesting to note that the current gain in an HBT does not depend on the emitter doping level, and is sensitive to the base doping density only through the variation of τ_n with the base doping density N_B. Therefore, it is possible to shape the doping profiles of an HBT such that the emitter doping density N_E is smaller than the base doping density N_B. As a result, the base spreading resistance $r_{b'b}$ and the emitter depletion capacitance C_{TE} can be greatly reduced. The base-spreading resistances $r_{b'b}$ for a circular and a rectangular geometry are given respectively by

$$r_{b'b} = \frac{1}{8\pi\mu_p Q_B} \quad \text{(circular)}, \tag{14.86}$$

$$r_{b'b} = \frac{1}{12\,(h/l)\,\mu_p Q_B} \quad \text{(rectangular)}, \tag{14.87}$$

where

$$Q_B = q\int_0^{W_b} N_B(x)\,dx \tag{14.88}$$

is the Gummel number. The emitter junction transition capacitance is given by

$$C_{TE} = A_E\sqrt{\frac{q\epsilon_0\epsilon_s N_E}{2\,(V_{bi} - V_{BE})}}. \tag{14.89}$$

The above equations can be used to improve the high-frequency performance of an HBT, such as increasing the cutoff frequency f_T and power gain G. This will be discussed next.

14.9.4. Current–Voltage Characteristics

In this section, the behavior of the collector current (I_C) and base current (I_B) as well as the current gain of a single heterojunction AlGaAs/GaAs HBT is discussed. In general, the current–voltage (I–V) behavior for an AlGaAs/GaAs HBT is similar to that of silicon BJTs but with a few distinct differences. Figure 14.24

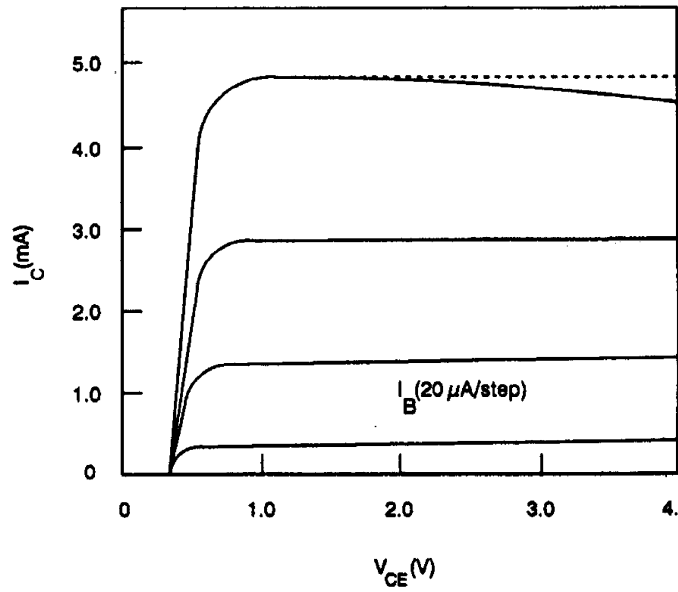


FIGURE 14.24. Collector current versus collector–emitter bias voltage for an AlGaAs/GaAs HBT with an emitter rea of $2 \times 3.5\ \mu m^2$; I_B steps: 20 μA. After Asbeck,[8] by permission, IEEE–1988.

shows the collector current versus collector–emitter bias voltage with base current as a parameter for the single-heterojunction AlGaAs/GaAs HBT used in digital and analog-to-digital (A/D) converter circuits. The HBT has an emitter area of $2 \times 3.5\ \mu m^2$. Several distinct features that are absent in a Si BJT are displayed in this figure. First, the nonzero offset voltage V_{CE} to produce positive I_C for the HBT is due to the difference in the turn-on voltage of the E-B junction and C-B junction. Second, there exists a negative differential output conductance at a higher I_B and V_{CE}, which is attributed to the heating effect at a higher current level or higher temperature. In general, the current gain of an HBT decreases with increasing temperature. Third, the dc current gain increases with increasing collector current ($\approx I_c^{1/2}$) and becomes saturated at high collector current. To understand the basic mechanisms governing the current conduction in an HBT, we next analyze the collector current and base current separately.

Figure 14.25 shows the Gummel plot (I_C, I_B versus V_{BE}) for the HBT shown in Figure 14.24. As shown in this figure, the ideality factor for the collector current is equal to unity for low to medium values of V_{BE}, implying that the diffusion current is the dominant component, while the diode ideality factor for I_B at low V_{BE} is equal to 2, indicating that the recombination current is dominant in the base. At high V_{BE} the series resistance effect becomes dominant for both I_C and I_B.

The collector current for an HBT can be explained using the Moll–Ross–Kroemer relation. If the collector current is base transport limited, then the collector

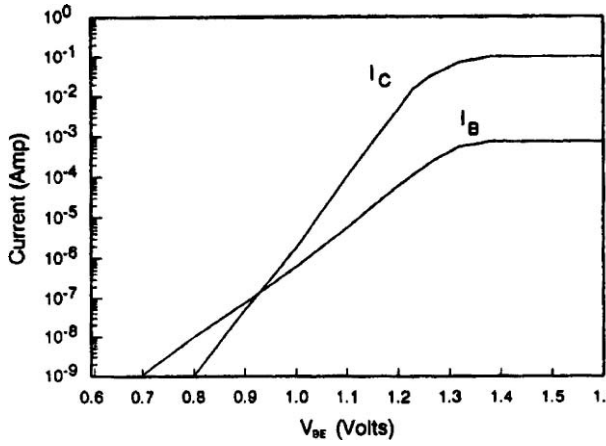


FIGURE 14.25. The collector and base current versus emitter-base bias voltage (Gummel plot) for the AlGaAs/GaAs HBT shown in Figure 14.24.

current density J_C can be expressed by[8]

$$J_C = \frac{q D_n n_{ie}^2 \exp(q V_{BE}/k_B T)}{\int_0^{W_B} p(x)\,dx}. \tag{14.90}$$

The integral in the denominator represents the number of impurity atoms per unit area (cm^{-2}) in the base, and is known as the Gummel number. Therefore, a large collector current can be realized with a smaller Gummel number, which corresponds to a narrow base width.

The base current I_B in an AlGaAs/GaAs HBT is more complex than that of a silicon BJT due to the use of a wide-band-gap AlGaAs emitter and a narrow-gap GaAs base. In general, deep-level defects such as DX centers in AlGaAs play an important role in controlling the recombination current in the E-B junction of the HBT. For example, the base current of an HBT may consist of four components: (i) recombination current in the base, (ii) recombination current in the E-B junction space-charge region, (iii) recombination current in the emitter, and (iv) periphery current. A general expression for these current components is given by

$$I_B \approx \exp(q V_{BE}/n k_B T), \tag{14.91}$$

where n is the diode ideality factor, which may vary between 1 and 2. When the recombination current is dominant in the base due to the short minority carrier lifetimes, the value of n is equal to unity and the current gain $\beta\,(=I_C/I_B)$ is constant. If recombination in the E-B junction space-charge region dominates due to the high density of deep-level centers, such as in the case of the graded E-B junction HBT shown in Figure 14.23b, then the value of n is equal to 2, and β increases with I_C and decreases with increasing temperature. If recombination in the emitter is dominant, then the value of n is equal to unity and β decreases

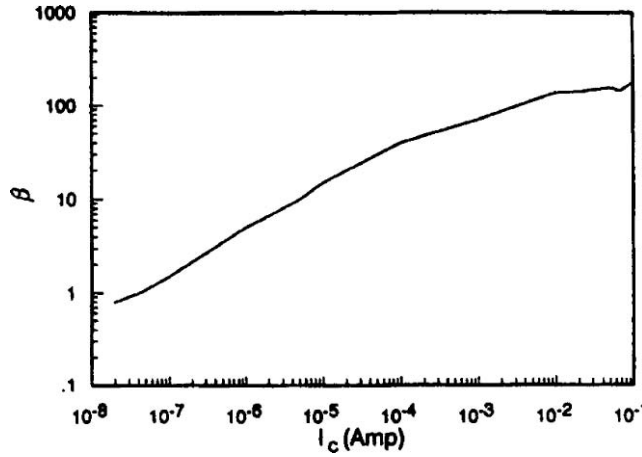


FIGURE 14.26. The dc current gain versus collector current for the AlGaAs/GaAs HBT shown in Figure 14.24.

with increasing temperature. Finally, the periphery current is attributed to the high surface recombination velocity around the emitter edge if the AlGaAs emitter surface is not properly passivated. In general, the current gain of an AlGaAs/GaAs HBT scales with the length of the emitter but not the area. Figure 14.26 shows the current gain versus collector current for the HBT shown in Figure 14.24. The results show that β greater than 100 can be achieved at higher collector current (e.g., $I_C \geq 10$ mA) for this device.

14.9.5. High-Frequency Performance

The cutoff frequency f_T is an important figure of merit for assessing the performance of an HBT in high-speed applications. The value of f_T for an HBT can be calculated using the expression

$$\frac{1}{2\pi f_T} = \tau_E + \tau_C + \tau_B + \tau_{TC}, \tag{14.92}$$

where

$$\tau_E = r_e (C_{TE} + C_{DE}) \approx \frac{4k_B T}{q I_E} C_{TE}(0) \tag{14.93}$$

is the emitter capacitance charging time, r_e is the E-B junction resistance, and C_{DE} is the emitter diffusion capacitance. The collector charging time τ_C is given by

$$\tau_C = r_{cc'} C_{TC}, \tag{14.94}$$

where

$$C_{TC} = A_c \sqrt{\frac{q\varepsilon_0 \varepsilon_s N_c}{2(V_{bi} + V_{CB})}} \tag{14.95}$$

is the C-B junction depletion capacitance and $r_{cc'}$ is the collector series resistance. Thus, to reduce τ_C, the doping density between the collector region and the collector contact should be as large as possible so that $r_{cc'}$ can be minimized. The effective base transit time τ_B is related to the effective electron velocity v_n and base width W_B by

$$\tau_B = \frac{W_B}{v_n} \approx \frac{W_B^2}{2D_n}, \tag{14.96}$$

where $D_n (= k_B T \mu_n / q)$ is the electron diffusion constant in the base. For an AlGaAs/GaAs HBT with a p^+ GaAs base of $W_B = 50$ nm and $v_n = 1 \times 10^7$ cm/sec, the value of τ_B is found to be 0.5 psec. The transit time of carriers across the C-B junction τ_{TC} is given by

$$\tau_{TC} = \frac{x_c}{v_s}, \tag{14.97}$$

where x_c is the depletion layer width of the C-B junction and v_s is the saturation velocity of carriers in the C-B junction. Finally, the power gain of an HBT can be written as

$$G = \frac{f_T}{8\pi f^2 r_{bb'} C_{TC}}. \tag{14.98}$$

Equation (14.98) shows that the power gain of an HBT is directly proportional to the cutoff frequency f_T and varies inversely with the parasitic base resistance and C-B junction capacitance. It is evident that high electron mobility in GaAs is essential for high-frequency performance of the HBT, because an increase in μ_n will lower the values of both τ_B and $r_{cc'}$, which in turn will increase f_T and hence the power gain G. A value of f_T equal to 75 GHz can be achieved for an AlGaAs/GaAs HBT with a 1.2 μm emitter width. In addition to high electron mobility, the lower doping density in the wide-band-gap AlGaAs emitter region and the higher doping density of the GaAs base region will result in a smaller E-B junction capacitance and a smaller base-spreading resistance. These two factors are essential for high-speed and high-frequency operation of the HBT. An AlGaAs/GaAs HBT with a very short base width (≤ 0.1 μm) can have a current gain of several thousands or higher, provided that the electron lifetime τ_n in the base is on the order of a nanosecond.

The high electron mobility and high base doping density in the GaAs base region will have additional beneficial effects on the performance of an AlGaAs/GaAs HBT. For example, parasitic mechanisms such as emitter current crowding and base widening in the collector region can be greatly reduced with the HBT structure shown in Figure 14.23a. The additional advantage is the increase of critical current density when the base widening becomes important. This critical current density

is related to the electron mobility by

$$J_{\text{bwc}} = q\mu_{\text{n}} N_{\text{dc}} \frac{V_{\text{CB}}}{W_{\text{c}}}, \tag{14.99}$$

where W_{c} is the width of the collector region. Another important consideration in the design of an HBT is the emitter current crowding effect, which becomes important when the emitter current density exceeds J_{ec}, given by

$$J_{\text{ec}} = \frac{8}{l^2} D_{\text{pb}} Q_{\text{b}} h_{\text{FE}}. \tag{14.100}$$

This shows that a higher base doping density (i.e., a higher Q_{b}) will reduce the emitter current crowding in the HBT.

In many circuit applications in which a large load capacitance is required, BJTs are preferred over field-effect transistors (FETs) because of their large current-carrying capability, high transconductance, and excellent threshold voltage control. The main advantage for developing an HBT is to reduce the base resistance $r_{\text{bb}'}$, which severely limits the high-speed performance of a BJT in the bipolar digital and microwave circuits. For example, the maximum oscillation frequency f_{max} for an HBT is given by

$$f_{\text{max}} = \frac{1}{4\pi r_{\text{bb}'} C_{\text{TC}} \tau_{\text{EC}}}, \tag{14.101}$$

which clearly shows that f_{max} is controlled by the base resistance $r_{\text{bb}'}$. The collector junction capacitance (C_{TC}) can be reduced using a smaller collector junction area, and τ_{EC} is the total emitter-to-collector delay time, which is given by

$$\tau_{\text{EC}} = \tau_{\text{E}} + \tau_{\text{B}} + \tau_{\text{TC}} + \tau_{\text{C}}. \tag{14.102}$$

Since τ_{E} is inversely proportional to the emitter current density, a large emitter current will improve the frequency response. The base resistance $r_{\text{bb}'}$ can also have a profound effect on the noise and performance of the HBT. Values of $r_{\text{bb}'}$ can be reduced by increasing the base width W_{B} and base-doping density N_{B}. Increasing base width W_{B} is not desirable since it will increase the base transit time τ_{B}, which in turn will reduce the base transport factor γ and current gain β. It is well known that increasing N_{ab} in a BJT will increase the unwanted carrier injection from the base into the emitter, which in turn will reduce the emitter injection efficiency. In an n-p-n HBT, however, due to the presence of two different band gap materials, the energy barriers for injection of electrons and holes are quite different. The barrier is larger for holes and the injection efficiency is nearly independent of the dopant density in the base. Furthermore, the gain is only limited by the base transport factor. As a result, the base region of an HBT can be heavily doped without significantly affecting the current gain. In fact, the dopant densities in the emitter and collector regions can be adjusted to minimize the junction capacitance and series resistance of an HBT. This is a very attractive feature for the HBT.

The conduction band spike at the E-B junction shown in Figure 14.23a is due to the abrupt transition at the AlGaAs/GaAs interface. This conduction band spike can be smoothed out using a compositional grading across the interface (i.e., by

changing the aluminum molar fraction x gradually in the $Al_xGa_{1-x}As$ layer). The spike can be used for near-ballistic injection of electrons into the base region to reduce the base transit time τ_B (e.g., from 1 to 0.2 psec for a base width of 0.1 μm). Without ballistic injection, the best reported unit current gain cutoff frequency $f_T\,(=1/2\pi\,\tau_{EC})$ for an HBT is about 40 GHz with an emitter width of 1.6 μm. Therefore, with ballistic injection, further improvement in f_T is possible. To further reduce τ_{EC}, the device area and parasitic of HBTs will have to be reduced and the current level increased. An important factor for increasing the speed of an HBT is the reduction of base resistance $r_{bb'}$.

Additional advantages of HBTs over BJTs include (1) the suppression of hole injection into the collector in saturating logic, (2) emitter–collector interchangeability leading to an improvement in VLSI circuit design, packing density, and interconnects, and (3) better control of the emitter–collector offset voltage. The collector of an HBT can also use a wide-band-gap AlGaAs material to form a double heterojunction transistor (DHBT). Besides III-V semiconductor HBTs, several new types of silicon HBTs using materials such as hydrogenated-amorphous silicon (a-Si:H and a-SiC:H) and hydrogenated microcrystalline silicon (μc-Si:H) as wide-band-gap emitters have been reported recently. Among these devices, the μc-Si:H n-p-n silicon HBT has the best overall performance characteristics. The device shows a much higher common emitter current gain than the conventional homojunction BJT. Figure 14.27 presents a comparison of the unit current gain cutoff frequency f_T as a function of collector current for an AlGaAs/GaAs HBT and a Si-BJT with similar geometries. The results clearly show that the former has a much higher f_T than the latter. In addition to AlGaAs/GaAs HBTs, AlGaAs/InGaAs p-n-p HBTs have been fabricated using carbon-doped material grown by the nonarsine MOVPE technique with $f_{max} = 39$ GHz and $f_T = 18$ GHz achieved. Operating in the common-base mode, this HBT has achieved a 0.5 W output power with 8 dB gain at 10 GHz.

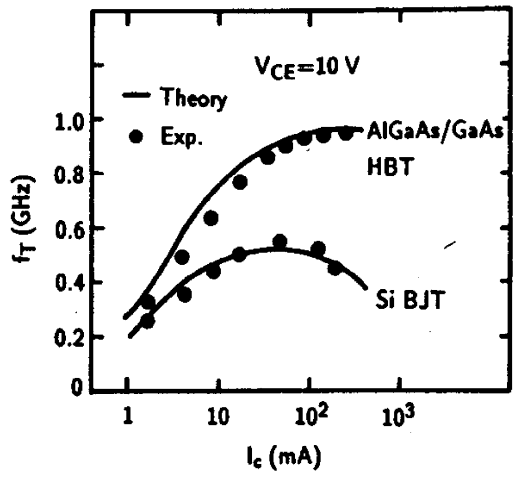
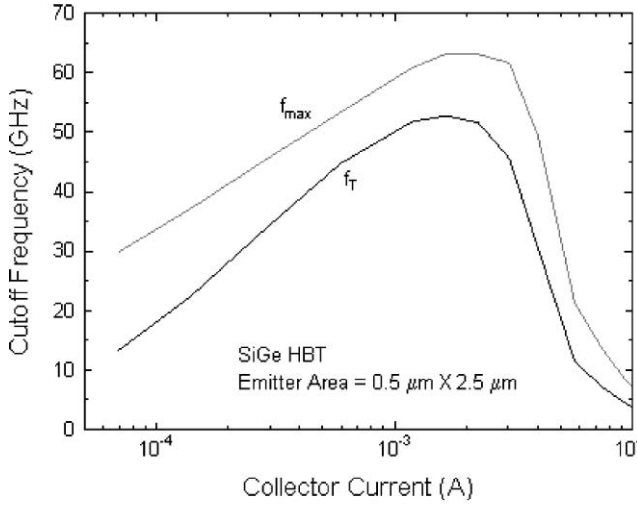


FIGURE 14.27. Comparison of unit current gain cutoff frequency f_T versus collector current for an AlGaAs/GaAs HBT and a Si BJT with similar geometry. After Beilbe et al. (8), by permission, IEEE–1980.

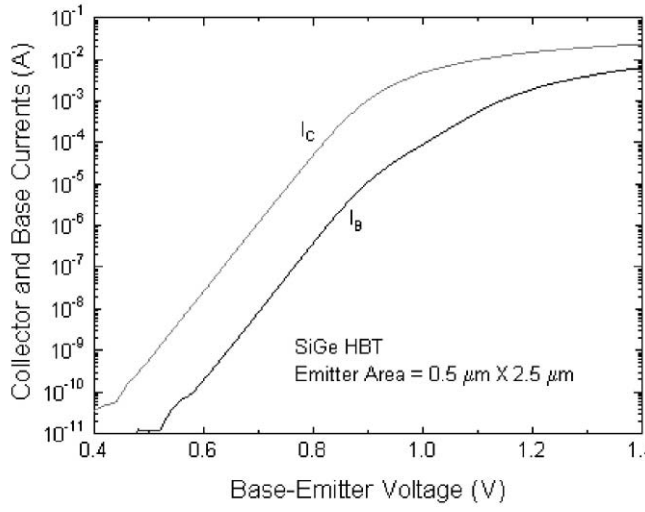
In the past decade, using molecular beam epitaxy (MBE) and ultra-high-vacuum chemical vapor deposition (UHV-CVD) techniques, high-performance Si/Ge_xSi_{1-x} HBTs have been developed for a wide variety of applications including high-speed communication and BiCMOS integration, and have been used as a key component in modern analog/mixed-signal/RF and high-frequency devices. This latest development has offered new promise for silicon-based HBTs to compete directly with III-V semiconductor HBTs and HEMTs for high-speed and high-frequency circuit applications. The epitaxial-base, in situ doped SiGe HBT technology has been developed by IBM for high-frequency applications. In research, IBM has achieved values of f_T of up to 120 GHz for the SiGe HBT. To suit a wide range of applications, two variants of the SiGe HBTs have been developed at IBM: a standard version for general applications (with $f_{T.} = 48$ GHz and $f_{max} = 70$ GHz; $\beta = 80$) and a high-breakdown version (with $f_{T.} = 30$ GHz) for power applications, which are able to tolerate a moderate tradeoff in f_T. Both variants can be mixed in any combination in the same circuit if desired. In addition to enabling a variety of high-frequency applications, the SiGe HBT also offers great advantages at the lower frequencies (e.g., 1.8 or 2.4 GHz) of today's hottest wireless applications. In 2004, Intel released the 90 nm SiGe HBT technology with f_T above 200 GHz. The tremendous headroom in speed may be traded for very low power. If high current is desired, the SiGe HBT can easily achieve values for I_C in excess of 1.6 $mA/\mu m^2$ of emitter area, with near perfect ideality and flat beta over 7 orders of magnitude. Figure 14.28a shows the plots of f_T and f_{max} versus collector current I_C for a GeSi HBT with emitter area of 0.5 μm × 2.5 μm developed by IBM. Figure 14.28b shows the collector and base currents versus base–emitter voltage (the Gummel plot) for the GeSi HBT shown in Figure 14.27a.

Figure 14.29 shows a comparison of unity current gain cutoff frequency (f_T) as a function of the device-critical dimension for various high-speed HBTs and PHEMT technologies. The 1-μm InP HBT technology has the highest f_T value (145 GHz) compared to Si/SiGe HBT and InGaP/GaAs HBT and GaAs PHEMT technologies. The key attributes make InP HBT technology ideal for high-speed digital and mixed-signal ICs with low to medium levels of integration include the following: (i) High-speed InP HBTs exhibit the highest cutoff frequency (f_T) of all commercial semiconductor technologies. Indeed, they are significantly faster than competing technologies at similar or smaller critical dimensions (see Figure 14.29). (ii) High reproducibility: Great progress has been made in the high-volume production of GaAs-based HBT circuits in the past decade, and this expertise can be translated to InP HBT technology. For example, the reproducibility of InP HBT turn-on voltage is typically a few millivolts, compared with hundreds of millivolts achieved with GaAs PHEMT.

Modern digital communications, instrumentation, electronics warfare, and radar systems require high-speed digital and mixed-signal ICs operating at frequencies from D.C. to 100 GHz. Broadband requirements place severe constraints on available semiconductor technologies and design expertise. Indeed, commercial off-the-shelf digital and mixed-signal ICs based on SiGe or GaAs technologies



(a)



(b)

FIGURE 14.28. (a) Cutoff frequencies f_T and f_{max} for a Si/SiGe HBT reported by IBM and (b) the collector- and base-current versus emitter–base voltage for the same Si/SiGe HBT shown in (a).

are generally available only at speeds up to 13 GHz. The advanced InP HBT technology enables digital and mixed-signal ICs at D.C. to 100 GHz frequencies. The advantages offered by InP HBT technology have opened up a range of new applications for high-speed mixed-signal and digital ICs.

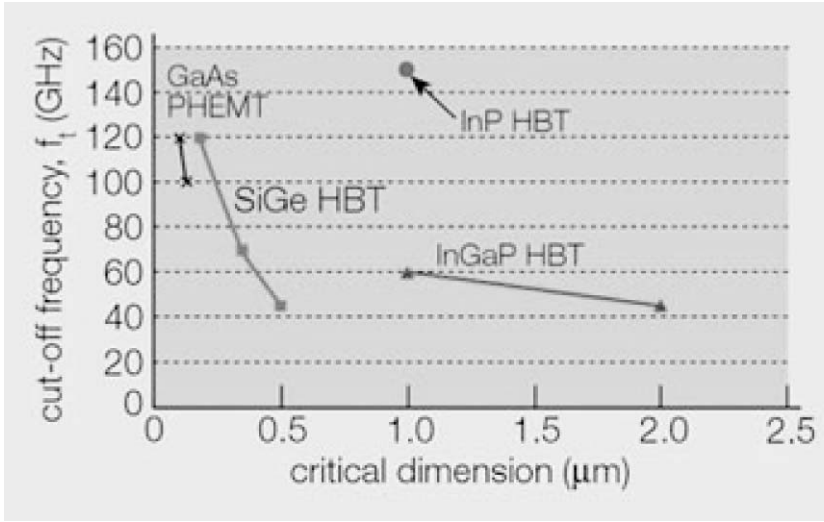


FIGURE 14.29. Comparison of unity current gain cutoff frequency (f_T) as a function of the device-critical dimension (emitter size) for various high-speed HBTs and PHEMT technologies.

Problems

14.1. (a) Plot the energy band diagram of an n-p-n transistor in thermal equilibrium and in the normal active mode of operation.
(b) Draw a schematic diagram of an n-p-n transistor, and show all the current components in the three regions of the transistor.

14.2. Plot the minority carrier density profiles for an n^+-p-n BJT for the following cases:
(a) The E-B junction is forward-biased and the C-B junction is reverse-biased.
(b) Both the E-B and C-B junctions are reverse-biased.
(c) Both the E-B and C-B junctions are forward-biased.
Plot minority carrier distributions in the base region for the cases $W_B \ll L_{nb}$ and $W_B > L_{nb}$, assuming $W_E > L_{pe}$ and $W_C > L_{pc}$.

14.3. Consider a double-diffused silicon p-n-p planar transistor, where the impurity profile after the base diffusion is given by

$$N_D(x) = \frac{Q_0}{\sqrt{\pi Dt}} \exp\left(\frac{x^2}{4Dt}\right),$$

where $Q_0 = 10^{16}$ cm^{-2}, $t = 5$ h, $D = 5 \times 10^{-14}$ cm^2/sec, and N_A (substrate) $= 10^{15}$ cm^{-3}.
(a) Calculate the collector junction depth for the above transistor.
(b) If the emitter junction is obtained by an additional short period of boron diffusion that yields an emitter junction depth of 1 μm, what is the base

dopant density $N_B(0)$ near the emitter junction, assuming that the collector junction depth remains unchanged after the emitter diffusion?

14.4. Using (14.10) through (14.16) show that the base current I_B for an n^+-p-n transistor can be expressed approximately by

$$I_B = qA'n_i^2 \left[\left(\frac{D_{pe}}{N_{de}W_e} + \frac{D_{nb}W_b}{2N_aL_{nb}^2} \right) \times (e^{qV_{EB}/k_BT} - 1) + \left(\frac{x_E}{2\tau_0 n_i} \right) e^{qV_{EB}/k_BT} \right]$$

and the inverse common-emitter current gain is given by

$$\frac{1}{h_{FE}} = \frac{N_{ab}W_bD_{pe}}{N_{de}W_bD_{nb}} + \frac{W_b^2}{2L_{nb}^2} + \left(\frac{N_{ab}W_bx_E}{2D_{nb}n_i\tau_0} \right) e^{-V_{EB}/k_BT},$$

where x_E is the emitter depletion layer width. Note that the second term in the square brackets of the above equation represents the recombination current in the depletion region of the forward-biased emitter junction. It is further assumed that the base width W_b is much smaller than the electron diffusion length L_{nb} in the base region (i.e., $W_b \ll L_{nb}$).

14.5. An interesting physical insight into the base transport factor β_T can be obtained if (14.33) is expressed in terms of the transit time τ_B of the minority carriers through the base.

(a) Show that the base transport factor for a p-n-p transistor can be written as

$$\beta_T = \frac{1}{1 + \tau_B/\tau_{pb}},$$

where $\tau_B = W_b^2/2D_{pb}$ and τ_{pb} denote the base transit time and hole lifetime in the n-base region, respectively, where W_b is the base width.

(b) Explain the physical significance of the result given in (a).

14.6. When a BJT is operating in the normal active mode, the C-B junction is reverse-biased. The collector voltage, which determines the depletion layer width of the C-B junction, can thereby affect the actual base width.

(a) Plot the minority carrier charge (i.e., holes) in the base region of a p-n-p transistor for two values of V_{CB}, assuming that I_E is kept constant. How is the base transport factor affected by this base width modulation (i.e., by the change of V_{CB})?

(b) Derive an expression of the output resistance for the transistor given in (a).

(c) For a planar silicon p^+-n-p transistor with $L_{pb} = 12\ \mu m$, $W_b = 1\ \mu m$, $V_{CB} = 10$ V, $I_C = 1$ mA, $N_D = 10^{16}\ cm^{-3}$, and $N_A = 5 \times 10^{15}\ cm^{-3}$, calculate the output resistance for this transistor using the result derived in (b).

14.7. The Gummel number can be calculated from the denominator of (14.27) if the impurity profile in the base region is known. Calculate the Gummel number of a silicon p-n-p transistor with

(a) Uniformly doped base with $N_D = 5 \times 10^{16}$ cm^{-3} and a base width of 1 μm.
(b) $N_d(x) = N_0 e^{-x/w}$, where $N_0 = 10^{18}$ cm^{-3} with a 1 μm base width.

14.8. (a) Show that the general expression for the base transport factor of a p$^+$-n-p BJT with an arbitrary base impurity doping profile can be expressed by

$$\beta_T = 1 - \left(\frac{1}{L_{pb}^2}\right) \int_0^W \left[\frac{1}{N_D} \int_x^W N_D(x)\right] dx.$$

Note that the above equation will reduce to (14.33) if the base doping profile is uniform.
(b) Using the expression for the base transport factor defined by (a), find values of the base transport factor for the base impurity dopant profiles given by (a) and (b) of Problem 14.7.

14.9. If the space-charge recombination current is negligible, show that the exact expression for the common-emitter output characteristics of a BJT is given by

$$-V_{CE} = \left(\frac{k_B T}{q}\right) \ln\left(\frac{-I_{CO} + \alpha_F I_B - I_C(1-\alpha_F)}{-I_{EO} + I_B + I_C(1-\alpha_R)}\right) + \left(\frac{k_B T}{q}\right) \ln\left(\frac{\alpha_R}{\alpha_F}\right).$$

14.10. (a) There are three possible ways of keeping a BJT switch in the off state. These include (i) the open base ($I_B = 0$), (ii) the B-E junction shorted ($V_{BE} = 0$), and (iii) the B-E junction reverse-biased ($V_{BE} < 0$). Draw the equivalent circuit diagrams of a p-n-p BJT for these three cases showing the polarity of V_{CC}, V_{CB}, and V_{EB}, the current flow, and the load resistance R_L.
(b) Find an expression for I_{CEO} (the open-base, collector–emitter leakage current) in terms of I_{CBO} (the open emitter, collector–base leakage current) and the forward current gain α_F for case (i).
(c) Find an expression for I_{CES} (the shorted base, collector–emitter leakage current) in terms of I_{CBO}, α_F, and α_R for case (ii).
(d) Find an expression for I_{CER} (the E-B junction reverse-biased) in terms of I_{CBO}, α_F, and α_R for case (iii).
(e) If $\alpha_F = 0.99$ and $\alpha_R = 0.1$, calculate the ratio of the leakage current to I_{CBO} for cases (i), (ii), and (iii).

14.11. (a) Construct the energy band diagram of an n$^+$-$In_{0.51}Ga_{0.49}P$/p-GaAs heterojunction diode, assuming that the conduction band offset $\Delta E_c = 0.21$ eV and the valence band offset $\Delta E_v = 0.25$ eV.
(b) Plot the energy band diagram for an n$^+$-$Al_{0.3}Ga_{0.7}As$/p-GaAs/n-GaAs HBT with an abrupt E-B interface by including the effect of the energy band offsets in the diagram.

References

1. P. G. Jespers, "Measurements for Bipolar Devices," in: *Process and Device Modeling for Integrated Circuit Design* (F. Van de Wiele, W. L. Engl, and P. G. Jespers, eds.), Noordhoff, Leyden (1977).
2. H. K. Gummel and H. C. Poon, "An Integral Charge Control Model of Bipolar Transistors," *Bell Syst. Tech. J.* **49**, 827 (1970).
3. E. J. McGrath and D. H. Navon, "Factors Limiting Current Gain in Power Transistors," *IEEE Trans. Electron Devices* **ED-24,** 1255 (1977).
4. J. J. Ebers and J. L. Moll, "Large Signal Behavior of Junction Transistors," *Proc. IRE* **49**, 834 (1961).
5. J. L. Moll, "Large-Signal Transient Response of Junction Transistors," *Proc. IRE* **42**, 1773 (1954).
6. A. Cuthbertson and P. Ashburn, "Self-Aligned Transistors with Polysilicon Emitters for Bipolar VLSI," *IEEE Trans. Electron Devices* **ED-32**, 242 (1985).
7. H. Kromer, *Proc. IRE*, 45,1535 (1957).
8. P. M. Asbeck, *IEEE IEDM Short Course: Heterostructure Transistors,* New York (1988).

Bibliography

J. Bardeen and W. H. Brattain, "The Transistor, A Semiconductor Triode," *Phys. Rev.* **74**, 230 (1948).

A. Bar-Lev, *Semiconductors and Electronic Devices*, 2nd ed., Prentice-Hall, Englewood Cliffs (1984).

E. I. Carroll, *Power Electronics for Very High Power Applications, 7th Int. Conf. Power Electronics and Variation Speed Drives*, p. 218 (1998).

C. Y. Chang and Francis Kai, *GaAs High Speed Devices*, Wiley, New York (1994).

C. Y. Chang and S. M. Sze, *ULSI Devices*, Wiley, New York (2000).

C. Y. Chang and S. M. Sze, *ULSI Devices*, Wiley & Sons, Inc., New York, 2000.

M. F. Chang, P. M. Asbeck, K. C. Wang, G. J. Sullivan, N. H. Sheng, J. A. Higgins, and D. L. Miller, *IEEE Elec. Dev. Lett.,* **EDL-8**, 303 (1987).

J. Early, "Effects of Space-Charge Layer Widening in Junction Transistors," *Proc. IRE,* **40**, 1401 (1952).

J. J. Ebers and J. L. Moll, "Large Signal Behavior of Junction Transistors," *Proc. IRE,* **49**, 834 (1961).

P. E. Gray, D. DeWitt, A. R. Boothroyd, and J. F. Gibbons, *Physical Electronics and Circuit Models of Transistors*, p. 145, SEEC Vol. II, Wiley, New York (1964).

P. E. Gray and C. L. Searle, *Electronic Principles: Physics, Models and Circuits*, Wiley, New York (1969).

A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York (1967).

C. T. Kirk, "A Theory of Transistor Cutoff Frequency Falloff at High Current Density," *IEEEE Trans. Electron Devices* **ED-9**, 164 (1962).

S. Konaka, Y. Yamamoto, and T. Sakai, "A 30 ps Bipolar IC Using Super Self-Aligned Process Technology," *IEEE Trans. Electron Devices* **ED-33**, 526 (1986).

H. F. Lips, "Technology Trends or HVDC Thyristor Valves," 1998 In *Conf. Power Syst. Tech.Proc.*, **1**, 446 (1998).

B. S. Meyerson, SiGe based mixed-signal technology for optimization of wired and wireless telecommunications, IBM J. RES. DEVELOP, vol. 44 (3), May (2000).

R. S. Muller and T. I. Kamins, *Device Electronics and for Integrated Circuits*, 2nd edition, Wiley, New York (1986).

D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*, 3rd edition, McGraw-Hill, New York (2003).

G. W. Neudeck, *Semiconductor Microdevices and Materials*, New York: Holt, Rinehart, & Winston, (1986).

J. F. A. Nijs, *Advanced Silicon and Semiconducting Silicon Alloy Based Materials and Devices*, Institute of Physics Publishing (1994).

T. H. Ning and R. D. Isaac, "Effect of Emitter Contact on Current Gain of Silicon Bipolar Devices," *IEEE Trans. Electron Devices* **ED-27,** 2051 (1980).

H. K. Park, K. Boyer, C. Clawson, G. Eiden, A. Tang, Y. Yamaguchi, and J. Sachitano, "High-Speed Polysilicon Emitter–Base Bipolar Transistor," *IEEE Electron Devices Lett.* **ED1-7**, 658 (1986).

B. K. Rose, "*Evaluation of Modern Power Semiconductor Devices and Future Trends of Converters,*" *IEEE Trans. Ind. Appl.,* **28(2)**, 403 (1992).

K. Schonenber et al., *A 200 mm SiGe HBT BiCMOS Technology for Mixed Signal Applications*, Proc. of the 1995 Bipolar/BiCMOS Circuits and Technology Meeting, BCTM'95 p. 89 (1995).

W. Shockley, "The Theory of p–n Junctions in Semiconductors and p–n Junction Transistors," *Bell Syst. Tech. J.* **28**, 435 (1949).

M. Shur, *Physics of Semiconductor Devices*, Englewood Cliffs, NJ: Prentice Hall (1990).

B. G. Streetman and S. Banerjee, *Solid State Electronic Devices,* 5th ed. Upper Saddle River: Prentice Hall (2000).

S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1981).

S. M. Sze, *High Speed Semiconductor Devices*, Wiley, New York (1990).

S. M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd. Edition, Wiley, New York (2002).

F. D. Taylor, *Thyristor Design and Realization*, Wiley Interscience, New York, 1993.

M. Vora, Y. L. Ho, S. Bhanre, F. Chien, G. Bakker, H. Hingarh, and C. Schmitz, *A Sub-100 Picosecond Bipolar ECL Technology*, IEDM Tech. Digest, p. 34, 1985.

E. S. Yang, *Microelectronic Devices*, McGraw-Hill, New York (1988).

J. S. Yuan, *SiGe, GaAs, and InP Heterojunction Bipolar Transistors*, Wiley, New York (1999).

15
Metal-Oxide-Semiconductor Field-Effect Transistors

15.1. Introduction

The metal-oxide-semiconductor (MOS) system is by far the most important device structure used in advanced integrated circuits (ICs) such as microprocessors and semiconductor memory chips. The present VLSI (very large scale integration) and ULSI (ultra-large-scale integration) digital circuits are based almost entirely on n-channel MOS field-effect transistors (MOSFETs) and complementary MOSFETs (CMOSFETs). The MOS structure is a basic building block for several key IC active components, namely, MOS field-effect transistors (MOSFETs), insulated-gate field-effect transistors (IGFETs), and charge-coupled devices (CCDs). Most commerically available MOSFETs and CCDs are fabricated from the Si–SiO_2 system. The MOSFETs consume very low power and can be easily scaled down for ULSI circuit applications. Therefore, it is pertinent to devote this chapter for silicon-based MOS capacitors, MOSFETs, and CCDs. Advanced FETs and other types of high-speed devices fabricated from III-V compound semiconductors will be described in Chapter 16.

As discussed in Chapter 11, the operation of a junction field-effect transistor (JFET) is based on the control of channel current by a reverse-bias p-n junction gate. In contrast to a JFET, the channel current of a MOSFET is controlled by the voltage applied across the gate electrode through a thin gate oxide grown on top of the channel. The current–voltage (I–V) characteristics of a MOSFET are very similar to those of a JFET. However, there are several advantages of a MOSFET over a JFET including lower power consumption, simpler structure, smaller size, higher packing density, higher yield, and higher compatiblity with VLSI technologies.

In this chapter, the basic device theories and general characteristics of silicon-based MOS capacitors, MOSFETs, and CCDs are presented. Section 15.2 describes the physical properties of the surface space-charge region and capacitance–voltage (C–V) behavior of an ideal MOS capacitor. The oxide charges and interface traps associated with the Si–SiO_2 interface of a nonideal silicon MOS capacitor are discussed in Section 15.3. Section 15.4 is concerned with basic device

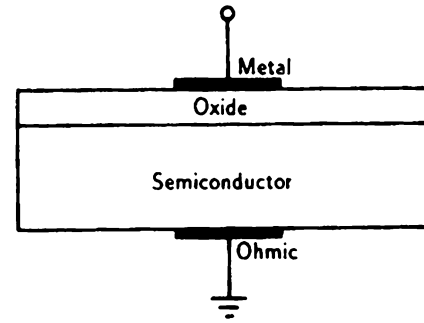


FIGURE 15.1. Cross-sectional view of an MOS capacitor.

physics, current–voltage characteristics, small-signal device parameters, and the equivalent circuit of a MOSFET. Some of the problems associated with a scaled-down MOSFET used in VLSI circuits are also discussed in this section. Section 15.5 presents the advanced MOSFET device structures and characteristics based on SOI (silicon-on insulator) technology for ULSI circuit applications. Finally, the operation principles and electrical characteristics of CCDs are discussed in Section 15.6.

15.2. An Ideal Metal-Oxide-Semiconductor System

In this section, the formation of a surface space-charge region and energy band diagrams for an ideal MOS capacitor under different bias conditions are discussed. The MOS structure has been used extensively for investigating the physical and electrical properties of a semiconductor surface as well as for various IC applications. Since the reliability and stability of a MOSFET and a CCD are closely related to the conditions of the semiconductor surface, understanding the physical and electrical properties of a semiconductor surface is essential for improving the performance of MOS devices. Although extensive studies of the Si–SiO_2 interface have been reported in the literature, new physical phenomena associated with the use of an ultrathin oxide layer in scaled-down MOSFETs need to be studied, and investigation of the top and bottom interface properties of the SOI MOSFETs has also been widely reported recently.

Figure 15.1 shows a cross-sectional view of a simple MOS capacitor. The energy band diagrams for an ideal MOS structure with n- and p-type semiconductor substrates under equilibrium conditions ($V = 0$) are illustrated in Figures 15.2a and b, respectively. An ideal MOS system is defined by the conditions that (i) the work function difference between the metal and the semiconductor is assumed equal to zero in thermal equilibrium conditions (i.e., $\phi_{ms} = 0$ at $V = 0$), (ii) the flat-band condition prevails, (iii) at any given bias condition, an equal amount of charge with opposite sign can exist only in the bulk semiconductor and at the metal–insulator interface, and (iv) no dc current can flow through an insulator (i.e.,

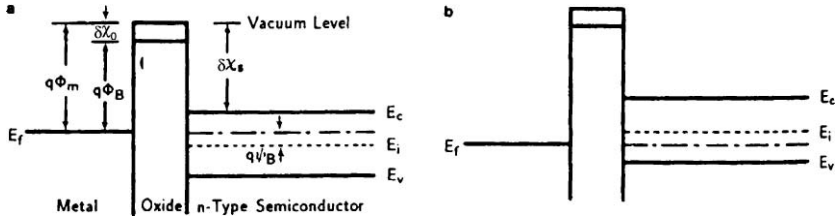


FIGURE 15.2. Energy band diagrams for an ideal MOS capacitor with (a) n-type and (b) p-type semiconductor substrates under equilibrium conditions ($V = 0$).

infinite oxide resistance). Condition (i) can be described by

$$\phi_{ms} = \begin{cases} \phi_m - \left(\chi_s + \frac{E_g}{2q} - \varphi_B\right) = 0 & \text{for n-type,} \quad (15.1) \\ \phi_m - \left(\chi_s + \frac{E_g}{2q} + \varphi_B\right) = 0 & \text{for p-type,} \quad (15.2) \end{cases}$$

where ϕ_m is the metal work function, χ_s is the electron affinity of the semiconductor, E_g is the energy band gap, φ_B is the bulk potential, and q is the electronic charge. As shown in Figure 15.2a, χ_0 denotes the electron affinity of the oxide, φ_B is the potential barrier between the metal and oxide, E_f is the Fermi level, and E_i is the intrinsic Fermi level. When a bias voltage is applied to an ideal MOS capacitor, three different surface charge conditions (i.e., accumulation, depletion, and inversion) can be created in the semiconductor surface; these are illustrated in Figures 15.3a–c for a metal-oxide p-type semiconductor structure. When a negative voltage is applied to the metal gate, the valence band bends upward and moves closer to the Fermi level. This results in an exponential increase in the majority carrier density (holes) at the semiconductor–oxide interface and the semiconductor surface is in accumulation, as shown in Figure 15.3a. When a small positive voltage is applied to the metal gate electrode, the valence band bends downward and the semiconductor surface becomes depleted; this is shown in Figure 15.3b. Finally, if a large positive voltage is applied to the metal gate, the valence band bends downward even more and the Fermi level moves above the intrinsic Fermi level. In this case, an inversion layer is formed at the semiconductor surface, as shown in Figure 15.3c. Therefore, depending on the polarity and the applied bias voltage, an accumulation, depletion, or inversion region can be created at the semiconductor surface of an MOS device. If the MOS structure is formed on an n-type semiconductor substrate, similar surface conditions to those of p-type substrates can be obtained, provided that the polarity of the applied bias voltage is changed. The charge distributions under different bias conditions are also shown on the right-hand side of Figures 15.3a–c. We shall next discuss the physical properties of the surface space-charge region and the high- and low-frequency capacitance–voltage (C–V) behavior for an ideal MOS capacitor.

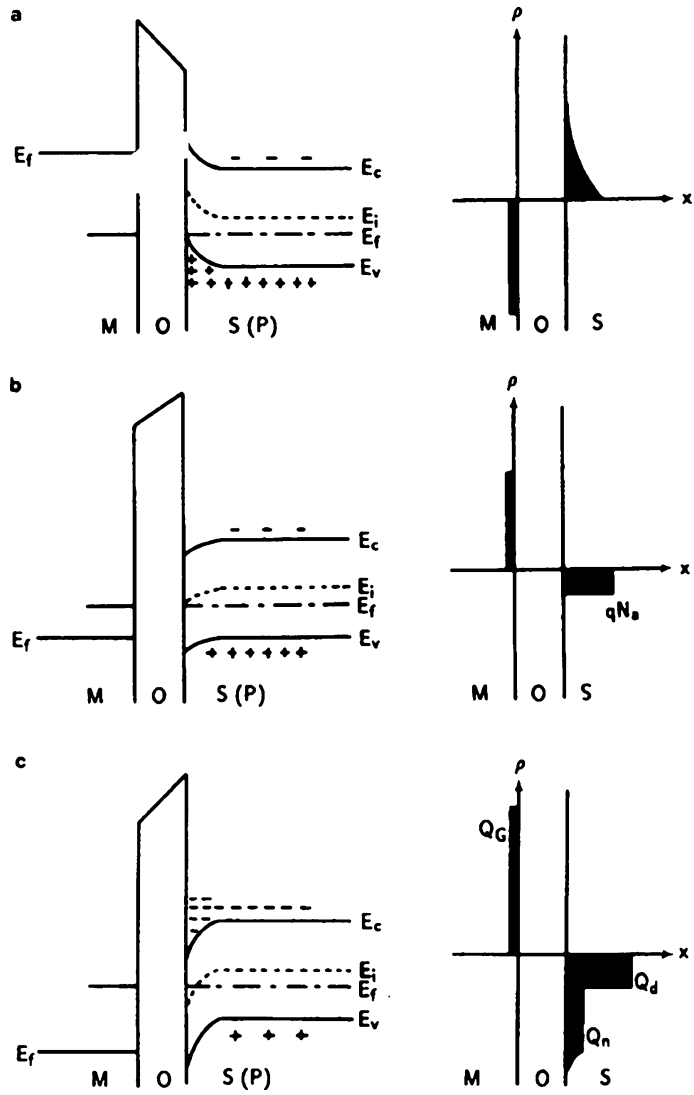


FIGURE 15.3. Energy band diagrams and charge distribution for a p-type MOS capacitor under bias conditions: (a) accumulation ($V < 0$), (b) depletion ($V > 0$), and (c) inversion ($V \gg 0$).

15.2.1. Surface Space-Charge Region

In order to predict the capacitance versus applied voltage (C–V) characteristics of an ideal MOS capacitor, we first derive the expressions for the space-charge density and electric field, which depend on the surface potential of the semiconductor. Figure 15.4 shows the energy band diagram for a p-type semiconductor

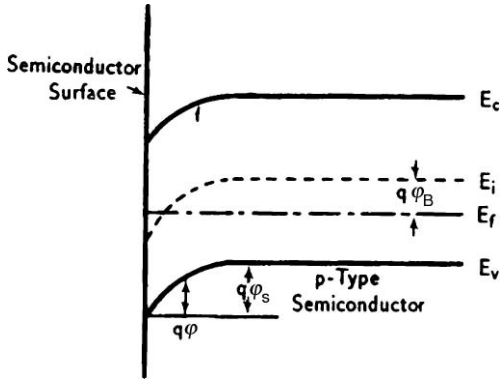


FIGURE 15.4. Energy band diagram at the surface of a p-type semiconductor. The potential Ψ is measured with respect to the intrinsic Fermi level and is equal to zero in the bulk semiconductor; $\Psi_B = (E_f - E_i)/q$ denotes the bulk potential. At the semiconductor surface, $\Psi = \Psi_s$, where Ψ_s is the surface potential. Accumulation occurs when $\Psi_s < 0$, depletion occurs when $\Psi_B > \Psi_s > 0$, and inversion occurs when $\Psi_s > \Psi_B$.

surface. The potential φ is measured with respect to the intrinsic Fermi level (e.g., $q\varphi_B = E_f - E_i$) in the bulk, which reduces to zero inside the bulk semiconductor. At the semiconductor surface, $\varphi = \varphi_s$, where φ_s is called the surface potential.

In a bulk semiconductor, electron and hole densities as a function of potential φ can be expressed by

$$p_p = p_{p0} \exp\left(-\frac{q\varphi}{k_B T}\right), \tag{15.3}$$

$$n_p = n_{p0} \exp\left(\frac{q\varphi}{k_B T}\right), \tag{15.4}$$

where p_{p0} and n_{p0} denote the equilibrium densities of holes and electrons in a p-type semiconductor, respectively. It is noted that φ is positive when the band bends downward. At the semiconductor surface, the densities of electrons and holes are given respectively by

$$p_s = p_{p0} \exp\left(-\frac{q\varphi_s}{k_B T}\right), \tag{15.5}$$

$$n_s = n_{p0} \exp\left(\frac{q\varphi_s}{k_B T}\right). \tag{15.6}$$

Equations (15.5) and (15.6) relate the carrier density at the semiconductor surface to the surface potential φ_s. Depending on the polarity and magnitude of the surface potential, different surface conditions can be established. These include (i) for $\varphi_s < 0$, accumulation of holes results, with the band bending upward; (ii) for $\varphi_s = 0$, the flat-band condition is obtained; (iii) for $\varphi_B > \varphi_s > 0$, the depletion of holes results, with the band bending downward; and (iv) for $\varphi_s > \varphi_B$, an inversion region is created near the surface, with the band bending downward. In general, the potential and electric fields as a function of distance from the interface to the bulk semiconductor can be obtained by solving Poisson's equation

$$\frac{d^2\varphi}{dx^2} = q\left(N_D^+ - N_A^- + p_p - n_p\right), \tag{15.7}$$

where N_D^+ and N_A^- denote the ionized donor and acceptor densities, respectively. Since the potential φ is zero in the bulk one obtains $N_D^+ - N_A^- = n_{p0} - p_{p0}$. Now substituting (15.3) and (15.4) into (15.7) and using the condition that $(N_D^+ - N_A^-) = (n_{p0} - p_{p0})$, the electric field as a function of distance from the surface into the bulk of the semiconductor can be expressed by

$$\mathcal{E} = \pm \frac{\sqrt{2} k_B T}{q L_D} \left[\left(e^{-q\varphi/k_B T} + q\varphi/k_B T - 1 \right) + \frac{n_{p0}}{p_{p0}} \left(e^{-q\varphi/k_B T} + q\varphi/k_B T - 1 \right) \right]^{\frac{1}{2}}, \tag{15.8}$$

where the plus sign is for $\varphi > 0$ and the minus sign for $\varphi < 0$, and L_D is the extrinsic Debye length for holes. The space charge per unit area required to produce this electric field can be obtained using Gauss's law, which is

$$Q_S = -\varepsilon_0 \varepsilon_r \mathcal{E}_s = \pm \frac{\sqrt{2} \epsilon_0 \epsilon_r k_B T}{q L_D} \left[\left(e^{-q\varphi_s/k_B T} + q\varphi_s/k_B T - 1 \right) + \frac{n_{p0}}{p_{p0}} \left(e^{-q\varphi_s/k_B T} + q\varphi_s/k_B T - 1 \right) \right]^{\frac{1}{2}}, \tag{15.9}$$

where $\mathcal{E}_s$ is the electric field at the surface and φ_s is the surface potential. Detailed derivation of the above equations as well as the variation of the space-charge density with the surface potential for p-type silicon can be found in the classic paper by Garrett and Brattain.[1]

It is interesting to note that the onset of strong inversion in an MOS device occurs at a surface potential given approximately by

$$\varphi_{si} \approx 2\varphi_B = \left(\frac{2k_B T}{q} \right) \ln \left(\frac{N_A}{n_i} \right), \tag{15.10}$$

where φ_B is the bulk potential.

15.2.2. Capacitance–Voltage Characteristics

In an ideal MOS capacitor, the effects due to interface traps, oxide charges, and work function difference are negligible. The energy band diagram for an ideal MOS device formed on a p-type silicon substrate is shown in Figure 15.3b for $V > 0$. The charge distributions in the bulk semiconductor and across the metal-oxide and oxide-semiconductor interfaces are shown in Figure 15.3b. From the charge-neutrality condition one obtains

$$Q_M = Q_n + q N_A W_d = Q_s, \tag{15.11}$$

where Q_M is the charge per unit area in the metal, Q_n is the charge per unit area in the inversion region, $q N_A W_d$ is the number of ionized acceptors per unit area in the space-charge region of width W_d, and Q_s is the total charge per unit area in the bulk semiconductor. The electric field and potential distribution for an ideal MOS capacitor are shown in Figures 15.5a and b, respectively.

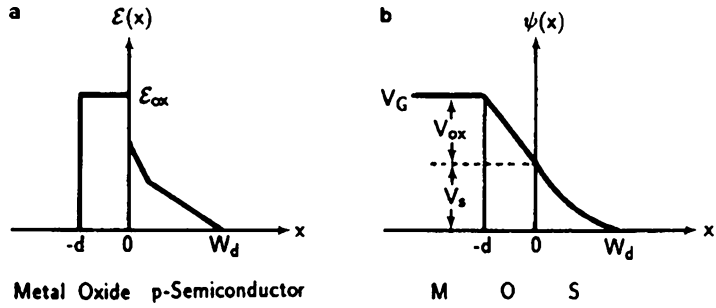


FIGURE 15.5. (a) Electric field distribution and (b) potential distribution of an ideal MOS capacitor under inversion conditions. The energy band diagram and charge distribution are shown in Figure 14.4.

If the work function difference between the metal and the semiconductor is neglected, then the applied voltage across the MOS capacitor is equal to the sum of the voltage drops across the oxide and semiconductor. This can be expressed as

$$V = V_{ox} + \varphi_s, \tag{15.12}$$

where V_{ox} is the potential drop across the oxide and is given by

$$V_{ox} = \mathcal{E}_{ox} d_{ox} = \frac{Q_s}{C_{ox}}. \tag{15.13}$$

It is noted that $C_{ox} = \varepsilon_{ox}\varepsilon_0/d_{ox}$ is the oxide capacitance per unit area. The total capacitance per unit area C is equal to the series combination of the oxide capacitance C_{ox} and the depletion layer capacitance C_d ($= \varepsilon_s\varepsilon_0/W_d$), namely,

$$C = \frac{C_{ox}C_d}{C_{ox} + C_d}. \tag{15.14}$$

Since C_d depends on the applied voltage, the total capacitance of the MOS capacitor is a function of the applied bias voltage. Figure 15.6 shows the low- and high-frequency small-signal capacitance versus applied voltage (C–V) plot for an ideal MOS capacitor formed on a p-type substrate. At high frequencies (typically 1 MHz), an accumulation of holes occurs near the semiconductor surface when a large negative-bias voltage is applied to the metal electrode, and a strong inversion region is formed near the semiconductor surface when a large positive-bias voltage is applied to the metal gate. A depletion region is created below the semiconductor surface when a small positive-bias voltage is applied to the MOS capacitor. It is seen that in the strong accumulation region (i.e., $V \ll 0$), C_d becomes very large and the total capacitance is equal to the oxide capacitance C_{ox}. This corresponds to the maximum capacitance of the MOS capacitor. In the strong inversion region (i.e., $V \gg 0$), the depletion layer reaches a maximum width and remains constant for further increase in the applied-bias voltage. Thus, the total capacitance in the strong inversion region is also constant. If the applied voltage becomes more positive than the flat-band voltage, then holes are pushed away from the semiconductor surface

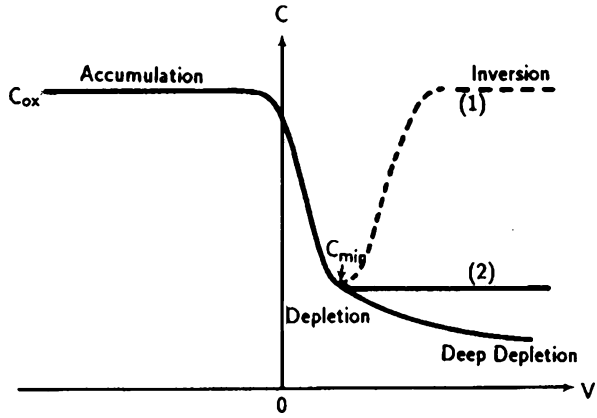


FIGURE 15.6. High- and low-frequency capacitance versus voltage (C–V) curves for a p-type MOS capacitor: (1) low-frequency C–V curve, (2) high-frequency C–V curve, and (3) high-frequency C–V curve in deep depletion.

and the surface becomes depleted. In this region, the depletion layer width varies with the applied voltage, and the total capacitance is also a function of the applied voltage. Of particular interest in the depletion region is the total capacitance per unit area under the flat-band condition (i.e., $\varphi_s = 0$), which is given by

$$C_{FB} = \frac{1}{d_{ox}/\varepsilon_{ox}\varepsilon_0 + L_D/\varepsilon_s\varepsilon_0}, \tag{15.15}$$

where $L_D = \sqrt{2k_BT\varepsilon_s\varepsilon_0/q^2N_A}$ is the extrinsic Debye length. For an ideal MOS capacitor, by neglecting interface traps, oxide charges, and the work function difference, the flat-band capacitance occurs at $V = \varphi_s = 0$. It is noteworthy that a depletion region is formed in the device when the surface potential φ_s is greater than zero but smaller than φ_B, where $\varphi_B = (k_BT/q)\ln(N_A/n_i)$ is the bulk potential. The weak inversion region begins at $\varphi_s = \varphi_B$, and the onset of strong inversion occurs at $\varphi_s \approx 2\varphi_B$.

The high-frequency C–V behavior (typically at 1 MHz) for a silicon MOS capacitor shown in Figure 15.6 can be explained using the one-sided abrupt junction approach. When the silicon surface is depleted, the number of ionized acceptors in the depletion region is equal to $-qN_AW_d$, where W_d is the depletion layer width. In this case, the potential distribution in the depletion region is a quadratic function of distance, and by solving Poisson's equation one obtains

$$\varphi = \varphi_s\left(1 - \frac{x}{W_d}\right)^2, \tag{15.16}$$

where φ_s is the surface potential, which is given by

$$\varphi_s = \frac{qN_AW_d^2}{2\varepsilon_s\varepsilon_0}. \tag{15.17}$$

From (15.16) and (15.17), it is seen that both φ_s and W_d will increase with increasing applied voltage. Eventually, the strong inversion condition is reached at $\varphi_{si} \approx 2\varphi_B$. When strong inversion occurs, the depletion layer width reaches a maximum. This maximum depletion layer width can be derived from (15.10) and (15.17), yielding

$$W_{d\,max} = \sqrt{\frac{2\varepsilon_0\varepsilon_s\varphi_{si}}{qN_A}} = \sqrt{\frac{4k_BT\varepsilon_0\varepsilon_s \ln(N_A/n_i)}{q^2N_A}}. \tag{15.18}$$

The threshold (or turn-on) voltage V_{TH} at the onset of strong inversion is given by

$$V_{TH} = \frac{Q_s}{C_i} + 2\varphi_B, \tag{15.19}$$

where $Q_s = qN_AW_{d\,max}$ is the total charge in the depletion region under strong inversion. The corresponding total capacitance at the onset of strong inversion is

$$C'_{min} = \frac{1}{1/C_{ox} + W_{d\,max}/\varepsilon_s\varepsilon_0}, \tag{15.20}$$

where C'_{min} is the minimum capacitance at the onset of strong inversion. Therefore, the high-frequency C–V behavior can be predicted using the above equations for different bias voltages. Values of the capacitance measured in the strong inversion region depend on the ability of minority carriers to follow up the applied ac signals in the small-signal capacitance measurements. This is usually accomplished by the low-frequency C–V measurements in which the generation–recombination rates of minority carriers can keep up with the small ac signals. The simplest case arises when both the dc gate-bias voltage and the small-signal measuring voltage are changed very slowly such that the semiconductor is near equilibrium. In this case, the signal frequency is low enough that the inversion layer population can follow it. The measured capacitance is equal to the stored charge on either side of the oxide, and hence its value is equal to the oxide capacitance C_{ox}. Under this condition, the C–V curve follows the low-frequency behavior as shown in Figure 15.6. The capacitance, which is equal to the oxide capacitance C_{ox} in the accumulation region, decreases while the surface is depleted, and moves back up to C_{ox} when the surface becomes inverted. For silicon MOS devices, the onset of low-frequency C–V behavior occurs for $f \leq 100$ Hz. In general, the capacitance in the inversion region increases from C_{min} to C_{ox} as the signal frequency decreases from the high-frequency regime to the quasistatic regime. Another interesting capacitance behavior shown in Figure 15.6 is the deep-depletion region. This corresponds to the experimental situation in which both the gate voltage and the small ac signals vary at a faster rate than the minority carrier generation–recombination rates in the surface depletion region. In this case the inversion layer cannot form and the depletion region becomes wider than $W_{d\,max}$. Consequently, deep depletion is used to describe this region. The deep-depletion phenomenon can be relaxed by using higher-bias voltages or by illuminating the MOS device during the C–V scan. A final note on the high-frequency C–V curve of the MOS capacitor is that the

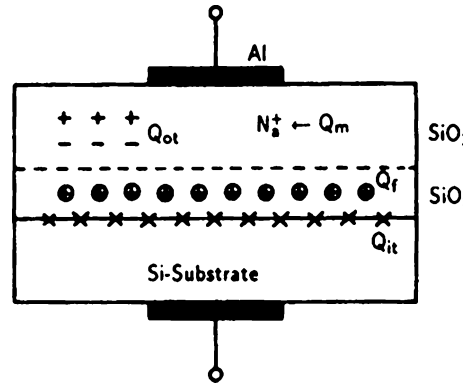


FIGURE 15.7. The thermally grown silicon MOS structure shown in the figure comprises a fixed oxide charge, an oxide trap charge, a mobile ion charge, and an interface at the Si–SiO_2 interface.

series resistance of the semiconductor substrate can also affect the capacitance value in the accumulation region under high-frequency conditions. It is generally observed that in the accumulation region, values of the capacitance will decrease with increasing frequencies when the series resistance effect becomes important.

15.3. Oxide Charges and Interface Traps

The oxide charges and interface traps play an important role in affecting the physical and electrical properties of an MOS device. To illustrate the importance of these charges, consider an Al–SiO_2–Si MOS capacitor structure as shown in Figure 15.7. For a thermally grown SiO_2 layer on silicon substrate, the transition from silicon to the stoichiometric SiO_2 is sharp. The transition region consists of SiO_x, where x may vary between 1 and 2. From X-ray photospectroscopy (XPS) measurements, this region has been found to be approximately 10 Å thick. A tail of silicon atoms bonding to only three oxygen atoms extends about 30 Å into the SiO_2 layer.

There are four major types of charges in SiO_2 and at the Si–SiO_2 interface that need to be considered. These are the mobile ionic charges (Q_m), the oxide trapped charges (Q_{0t}), the fixed oxide charges (Q_f), and the interface trapped charges (Q_{it}). The mobile ionic charges are usually caused by sodium or potassium ions, which become mobile at high temperatures under an applied electric field. These positively charged ions could migrate from the bulk of the oxide layer to the Si–SiO_2 interface over a period of time, slowly increasing the oxide charge there. The oxide trapped charges arise from defects in the oxide. These defects can be structural, chemical, or impurity related. The defects, initially neutral, capture electrons or holes and become negatively or positively charged. Since very little current flows through the oxide layer during normal device operation, the traps usually remain neutral. However, if carriers are injected into the oxide, or ionizing radiation travels through the oxide, these traps can become charged. Both Q_m and Q_{0t} are distributed randomly throughout the oxide layer. The exact nature of the

fixed oxide charges is not known. However, more than one type of defect may cause fixed oxide charges. Although 90% of the charges are located within 30–40 Å of the Si–SiO_2 interface, fixed charges are not mobile and are independent of the applied voltage. The density of Q_f is highly dependent on the process used to create the oxide layer and on the orientation of the silicon substrate. Finally, interface traps are generally caused by trivalent silicon, which occurs when silicon atoms bond to only three oxygen atoms instead of four. This defect is amphoteric. As the Fermi level rises from the valence band toward the mid-gap, the interface traps capture electrons and become neutral. As the Fermi level rises toward the conduction band, the traps accept additional electrons and become negatively charged. It is also possible for positive charges near the interface to induce interface traps. The energies of the traps vary continuously throughout the silicon band gap. Therefore, the probability of trapping an electron is dependent on the applied bias voltage. The positions of the oxide charges and the interface traps in a Si–SiO_2 system are shown in Figure 15.7.

15.3.1. Interface Trap Charges

Interface states in a Si–SiO_2 system are referred as fast states, which can exchange charges with silicon in a very short period of time. They may be created by trivalent silicon (i.e., the so-called P_b center), excess oxygen, or impurities. The density profile of these interface traps across the forbidden gap of silicon is generally found to be of a "U" shape, relatively flat near the mid-gap and increasing very rapidly toward the band edges. Since the interface trap states are distributed across the silicon forbidden gap, it is important to find out the distribution of the interface trap density across the band gap. The density of the interface trap states can be expressed by

$$D_{it} = \left(\frac{1}{q}\right)\left(\frac{dQ_{it}}{dE}\right), \tag{15.21}$$

where D_{it} has units of $cm^{-2}\cdot eV^{-1}$ and Q_{it} is the total number of interface charges per unit area (C/cm^2). Integrating (15.21) once with respect to energy E from the valence band edge E_v to the conduction band edge E_c yields the total interface state density per unit area in the forbidden gap.

When a voltage is applied to an MOS capacitor, the interface trap levels will move up or down with the valence and conduction bands, while the Fermi level remains constant. A change of interface trap charge density will cause a change in capacitance, and hence will alter the C–V curve of an ideal MOS device. Figure 15.8a shows the equivalent circuit of an MOS capacitor when the interface state traps are included, where C_{ox} denotes the oxide capacitance, C_d is the depletion layer capacitance, C_{it} is the capacitance associated with the interface traps, and R_{it} is the resistance associated with the interface traps. Quantities C_{it} and R_{it} are a function of the surface potential. The product $R_{it}C_{it} = \tau_{it}$ is defined as the interface state lifetime, which determines the frequency behavior of the interface traps. The parallel branch of the equivalent circuit shown in Figure 15.8a can be converted

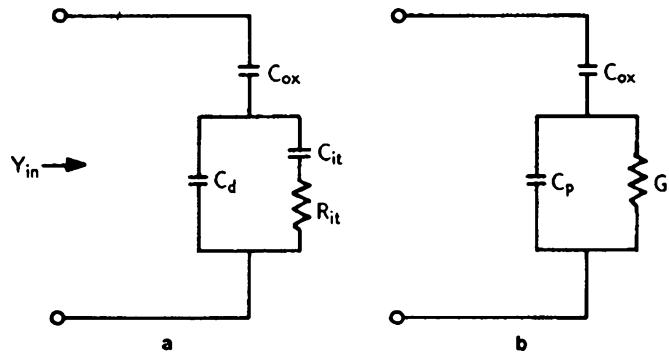


FIGURE 15.8. Equivalent circuit of an MOS capacitor taking into account the interface states effect: (a) C_{it} and G_{it} are the capacitance and conductance associated with the interface traps, respectively; C_d is the depletion capacitance and C_{ox} is the oxide capacitance. Quantities C_p and G_p shown in (b) are given by $C_p = C_p + C_{it}/(1 + \omega^2\tau^2)$. and $G_p = C_{it}\omega^2\tau/(1 + \omega^2\tau^2)$.

into a parallel frequency-dependent capacitance C_p and a parallel conductance G_p equivalent circuit as shown in Figure 15.8b, with both components given by

$$C_p = C_d + \frac{C_{it}}{1 + \omega^2\tau_{it}^2}, \tag{15.22}$$

$$G_p = \frac{C_{it}\omega^2\tau_{it}}{1 + \omega^2\tau_{it}^2}. \tag{15.23}$$

Therefore, the input admittance of an ideal MOS capacitor can be written as

$$Y_{in} = G_{in} + j\omega C_{in}, \tag{15.24}$$

where G_{in} and C_{in} are given by

$$G_{in} = \frac{\omega^2 C_{it}\tau_{it}C_{ox}^2}{(C_{ox} + C_d + C_{it})^2 + \omega^2\tau_{it}^2\,(C_{ox} + C_d)^2}, \tag{15.25}$$

$$C_{in} = \frac{C_{ox}}{(C_{ox} + C_d + C_{it})}\left[C_d + C_{it}\frac{(C_{it} + C_d + C_{ox})^2 + \omega^2\tau_{it}^2 C_d\,(C_{ox} + C_d)}{(C_{it} + C_d + C_{ox})^2 + \omega^2\tau_{it}^2\,(C_{ox} + C_d)^2}\right]. \tag{15.26}$$

It is noted that both the input conductance G_{in} and input capacitance C_{in} given by (15.25) and (15.26) contain similar information with regard to the interface state traps, which can be determined using either the capacitance or conductance measurements. It can be shown that for MOS devices, when the interface state density is low, the conductance technique is more accurate than the capacitance method. On the other hand, the capacitance technique can give a rapid evaluation of the flat-band shift and the total interface trap charge Q_{it}. Figures 15.9a–c show the stretch-out of the C–V curves of an MOS capacitor owing to the increase in the interface trap charges: (a) the stretch-out of the high-frequency C–V curve for a p-type MOS capacitor owing to the interface trap charges, and the shift of the

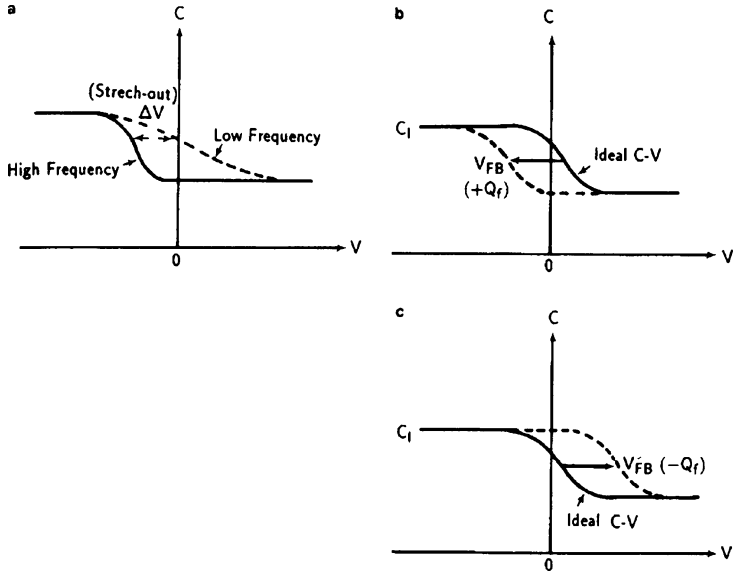


FIGURE 15.9. (a) The stretch-out of the high-frequency C–V curve for a p-type MOS capacitor due to the interface trap charges, and the shift of the C–V curves due to (b) the positive fixed charge and (c) the negative fixed charge in the oxide.

C–V curves owing to (b) the positive fixed charge and (c) the negative fixed charge in the oxide. At high frequencies (i.e., $\omega\tau \gg 1$), the interface traps cannot follow the ac signals. As a result, the expression for the capacitance given by (15.26) reduces to (15.14). In this case, the effect of interface traps on the high-frequency capacitance curve is negligible. As shown in Figure 15.9, in the presence of interface trap charges the ideal MOS C–V curve will stretch out along the voltage axis since more charges on the metal gate are needed for a given surface potential.

Measurements of the interface trap density using capacitance and conductance techniques in a Si–SO_2 device have been reported extensively in the literature.[2] Figure 15.10 shows the distribution of interface trap densities in the forbidden gap of silicon for thermally oxidized silicon along the (111) and (100) orientations. The results clearly show that the interface trap density at mid-gap is about one order of magnitude higher for the (111) orientation than for the (100) orientation. This result has been correlated to the difference in the available dangling bonds per unit area on (111) and (100) silicon surfaces [e.g., the density of dangling bonds is 11.8×10^{14} cm^{-2} for a (111) silicon surface and 6.8×10^{14} cm^{-2} for a (100) surface].

15.3.2. Oxide Charges

The oxide charges in SiO_2 films will be considered next. There are three different types of oxide charges that could be presented in the SiO_2. These include the oxide trap charge Q_{ot}, the mobile ionic charge Q_m, and the fixed oxide charge Q_f. As discussed earlier, the fixed oxide charge is located within 30 Å of the Si–SiO_2

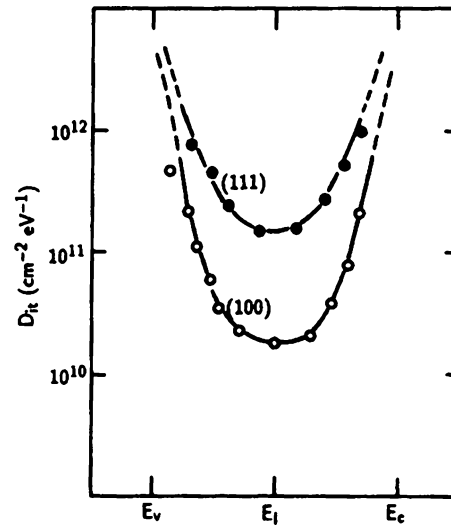


FIGURE 15.10. Distribution of interface trap densities in the forbidden gap of silicon for a thermally grown silicon dioxide along the (111) and (100) orientations. After White and Cricchi,[3] by permission, © IEEE–1972.

interface. It cannot be charged or discharged over a wide range of the surface potential. The density of the fixed oxide charge is virtually independent of the oxide thickness and the type or density of impurities in the bulk silicon. Its charge state is generally positive and depends on the oxidation annealing conditions and on the orientation of silicon crystal. The physical origins of the fixed oxide charge have been attributed to the trivalent silicon and nonbridging oxygen (excess oxygen) near the Si–SiO_2 interface. Figure 15.9b shows the shift of a high-frequency C–V curve along the voltage axis when the positive or negative fixed charges are present near the Si–SiO_2 interface of a p-type silicon substrate. The voltage shift is measured with respect to an ideal C–V curve when the fixed charge is equal to zero. For a negative fixed charge, the C–V curve shifts to more positive bias voltages, and a positive fixed charge shifts the C–V curve toward more negative bias voltages with respect to the ideal C–V curve. The reason for the shift of the C–V curve in the presence of fixed oxide charges is due to the fact that charge neutrality in a practical MOS capacitor requires that every negative charge on the metal gate must be compensated by an opposite charge in the oxide and in the bulk silicon substrate. This implies that in the presence of a positive fixed oxide charge in the oxide, the net shallow ionized donor density in silicon must be reduced, which, in turn, will decrease the depletion layer width. Thus, the capacitance will be higher than that of the ideal case for all values of the applied gate voltages in the depletion and weak inversion regions. The result is a shift of the C–V curve toward a more negative bias region for the positive fixed charges and toward a more positive gate bias for the negative fixed charges. The magnitude of the C–V shift due to fixed oxide charges with respect to the ideal C–V curve can be estimated using the expression

$$\Delta V_{\mathrm{f}} = \frac{Q_{\mathrm{f}}}{C_{\mathrm{ox}}}, \tag{15.27}$$

where Q_f is the fixed oxide charge density and C_{ox} is the oxide capacitance per unit area.

Mobile ionic charges such as sodium ions (which are present in thermally grown SiO_2) can cause surface instability of passivated silicon devices. Reliability problems of silicon devices operating at high temperatures and high bias conditions may be related to the trace contamination of sodium ions in these devices. Mobile ionic charges due to sodium can move in and out of the oxide with changes in biasing and temperature conditions. These problems can be reduced or eliminated if a cap layer of Si_3N_4, Al_2O_3, or phosphosilicate glass is used as a sodium barrier layer. The shift of the C–V curve due to the mobile ionic charge can be calculated using the expression

$$\Delta V_m = \frac{Q_m}{C_{ox}}, \tag{15.28}$$

where Q_m is the mobile ionic charge per unit area at the Si–SiO_2 interface.

Oxide trap charges can also cause a voltage shift in the ideal C–V curve. The oxide traps are associated with defects created either by impurities or by radiation damage in the oxide layer. They are usually neutral and become charged when electrons or holes are captured by the oxide traps. The voltage shift due to the oxide trap charges can be calculated from

$$\Delta V_{0t} = \frac{Q_{ot}}{C_{ox}}, \tag{15.29}$$

where Q_{0t} is the net oxide trap charge per unit area in the SiO_2. Therefore, the total voltage shift of the ideal C–V curve can be expressed by

$$\Delta V_t = \Delta V_f + \Delta V_m + \Delta V_{0t} = \frac{Q_0}{C_{ox}}, \tag{15.30}$$

where $Q_0 = Q_f + Q_m + Q_{0t}$ is the sum of the effective net oxide charges per unit area in the SiO_2 layer.

From the above analysis, it is clear that oxide charges play an important role in the stability and reliability of an MOS capacitor. Therefore, effective control of the oxide charges in silicon MOS devices is essential for stable device operation.

Finally, the metal–semiconductor work function difference affects the flat-band voltage shift. The work function difference between a metal and an n-type semiconductor is obtained from (15.1), which reads

$$\phi_{ms} = \phi_m - \left(\chi_s - \frac{E_g}{2q} - \varphi_B\right). \tag{15.31}$$

For an ideal MOS device, ϕ_{ms} is zero. If ϕ_{ms} and Q_0 are not equal to zero, then the measured C–V curve of the MOS capacitor will be shifted from the ideal C–V curve by an amount given by

$$V_{FB} = \phi_{ms} - \frac{Q_0}{C_{ox}}, \tag{15.32}$$

where V_{FB} is known as the flat-band voltage shift, and $Q_0 = Q_f + Q_m + Q_{0t}$ is the total oxide charge. If the mobile ionic charge Q_m and the oxide trap charge Q_{0t} are negligible, then the flat-band voltage will reduce to

$$V_{FB} = \phi_{ms} - \frac{Q_f}{C_{ox}}. \tag{15.33}$$

It is noted that the work function difference ϕ_{ms} can have a significant influence on both the surface potential and voltage shift. For example, for an Al–SiO_2–Si MOS capacitor with $\phi_m = 4.1$ eV and $\phi_s = 4.35$ eV, the work function difference ϕ_{ms} is found to be equal to $-0{\cdot}25$ eV. This work function difference must be included in the calculation of voltage shift in the measured C–V curve.

15.4. MOS Field Effect Transistors

In this section we present the basic principles, device structure, and characteristics of a long-channel metal-oxide-semiconductor field-effect transistor (MOSFET). The MOSFET is a unipolar device in which the current conduction is due to the majority carriers. The silicon MOSFET is probably the most important active component used in a wide variety of silicon IC applications such as microprocessors, logic and memory chips, power devices, and many other digital ICs. The MOSFET is usually referred to as the FET formed on the Si–SiO_2 system. Other acronyms such as IGFET (Insulated-Gate FET) or MISFET (Metal–Insulator–Semiconductor FET) have also been used for FETs formed on different insulators or semiconductors. Although semiconductors such as Si, GaAs, and InP and insulators such as SiO_2, Al_2O_3, and Si_3N_4 have been used in the fabrication of IGFETs, the majority of IGFETs used in present-day VLSI technologies are almost entirely based on the Si–SiO_2 system. Therefore, only silicon-based MOSFETS will be presented in this section, while MESFETs (MEtal–Semiconductor FETs) and modulation-doped FETs (MODFETs or HEMTs) using III-V compound semiconductors (e.g., GaAs and InP) will be discussed in Chapter 16. In addition to the conventional MOSFETs, advanced MOSFETs using SOI technology will also be discussed in this section.

15.4.1. General Characteristics of a MOSFET

Figure 15.11 shows the structure of an n-channel silicon MOSFET. The basic structure of an n-channel MOSFET consists of a p-type silicon substrate, the heavily doped n^+ source and drain regions formed by ion implantation or by thermal diffusion on the p-substrate, and a thin gate oxide (MOS structure) deposited on the p-substrate that serves as the gate electrode to control the current flow through the channel region underneath the gate oxide and between the source and drain regions. The gate electrode deposited on top of the SiO_2 gate oxide is formed using either heavily doped polysilicon or a combination of polysilicon and silicide metal. In addition to the gate oxide, a much thicker field oxide surrounding the MOSFET is also deposited on the outer edge of the device to isolate it from other devices on the same IC chip. A conducting channel between the source and drain

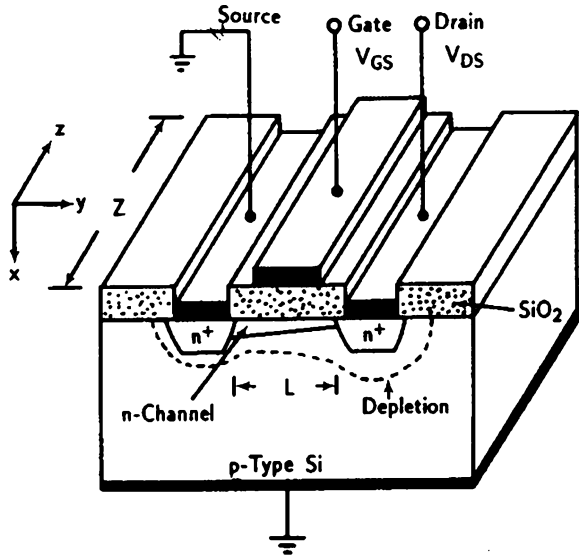


FIGURE 15.11. Device structure of an n-channel silicon MOSFET.

can be formed using a buried implanted layer or a channel that can be induced by applying a gate voltage. The distance between the metallurgical junctions of the source and drain regions is defined as the channel length L (i.e., along the y-direction), and the channel width along the z-direction is designated as Z. The gate oxide thickness is denoted by d_{ox} (≤ 100 nm), and N_A is the substrate dopant density. The n^+-p junctions formed in the source and drain regions are electrically isolated from one another when the gate voltage is equal to zero. If a positive voltage is applied to the gate electrode, an n-type inversion layer is induced at the surface of the semiconductor, which in turn creates a conducting channel between the source and drain regions. When the semiconductor surface under the gate oxide is inverted, and a voltage is applied between the source and drain junctions, electrons can enter the inverted channel from the source junction and leave at the drain junction. Similarly, a p-channel MOSFET can be fabricated using an n-type substrate and implantation of boron to form heavily doped p^+ source and drain regions. In a p-channel MOSFET, holes are the majority carriers that flow through the channel between the source and drain regions of the device.

Under thermal equilibrium conditions (i.e., $V = 0$), if a work function difference and oxide charges exist in the Si–SiO_2, then an inverted surface or channel between the source and drain regions may be formed in the MOSFET. In this case, the device is called a *depletion-mode* MOSFET because a negative bias voltage must be applied to the gate in order to deplete the carriers from the channel region to reduce the channel conductance. This type of MOSFET is also known as a normally on depletion-mode MOSFET. However, in most MOSFETs, a positive-bias voltage must be applied to the gate to induce a channel under the gate oxide of the device. This type of device is usually called an *enhancement-mode* MOSFET

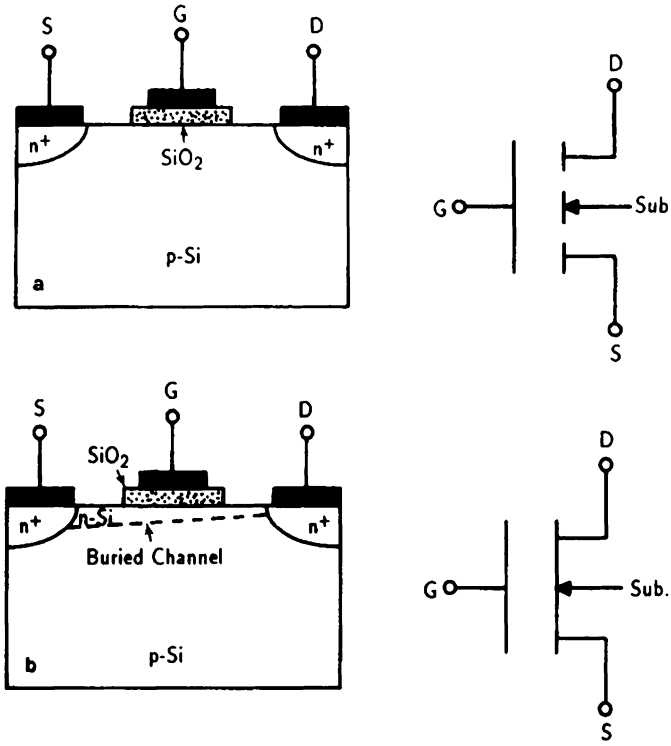


FIGURE 15.12. Cross-sectional views and circuit symbols of (a) an enhancement-mode (normally off) n-channel MOSFET and (b) a depletion mode (normally on) n-channel MOSFET.

or a normally off MOSFET. Figure 15.12 shows the cross-sectional views of (a) an enhancement-mode n-channel MOSFET and (b) a depletion-mode n-channel MOSFET. Enhancement-mode MOSFETs are more widely used in IC applications than depletion-mode MOSFETs. The depletion-mode MOSFET is also called the buried-channel MOSFET because the channel conduction occurs inside the bulk silicon.

The MOSFET is a four-terminal device with electrical contacts to the source, drain, gate, and substrate. Under normal operating conditions, the source and substrate terminals are connected to a common ground. However, when a bias voltage is applied to the substrate it can also change the channel conductance of the MOSFET.

15.4.2. Channel Conductance

The channel conductance is a very important parameter for MOSFET operation. As shown in Figure 15.11, when a positive voltage is applied to the gate electrode of a MOSFET, an inversion layer is formed in the semiconductor surface under the

gate oxide. The inversion layer provides a conducting path between the source and drain, and is known as the channel. The channel conductance can be calculated using the expression

$$g_{\mathrm{I}} = \frac{Z}{L}\int_0^{x_{\mathrm{I}}} q\mu_{\mathrm{n}} n_{\mathrm{I}}(x)\,\mathrm{d}x, \tag{15.34}$$

where $n_{\mathrm{I}}(x)$ is the density of electrons in the inversion (n) channel, Z/L is the gate width to gate length ratio, μ_{n} is the electron mobility in the channel, q is the electronic charge, and x_{I} is the channel depth along the x-direction. The total charge per unit area Q_{I} in the n-inversion channel is obtained by integrating $n_{\mathrm{I}}(x)$ over the channel depth. This can be expressed by

$$Q_{\mathrm{I}} = -\int_0^{x_{\mathrm{I}}} q n_{\mathrm{I}}(x)\,\mathrm{d}x. \tag{15.35}$$

Solving (15.34) and (15.35), the channel conductance becomes

$$g_{\mathrm{I}} = -\left(\frac{Z}{L}\right)\mu_{\mathrm{n}} Q_{\mathrm{I}}. \tag{15.36}$$

It is worth noting that Q_{I} is a function of the applied gate voltage and that the induced mobile charges in the channel become the current carriers in the device. The threshold voltage V_{TH} is defined as the gate voltage required to achieve strong inversion in the MOSFET, and is given by

$$V_{\mathrm{TH}} = -\frac{Q_{\mathrm{B}}}{C_{\mathrm{ox}}} + \varphi_{\mathrm{si}}, \tag{15.37}$$

where Q_{B} is the bulk charge and $\varphi_{\mathrm{si}} \approx 2\varphi_{\mathrm{B}}$ is the surface potential under strong inversion. If a voltage is established in the channel by the applied voltages from the source and drain electrodes, then Q_{B} under strong inversion can be written as

$$Q_{\mathrm{B}} = -\sqrt{2q\varepsilon_0\varepsilon_{\mathrm{s}} N_{\mathrm{A}}(V_{\mathrm{c}} + \varphi_{\mathrm{si}})}. \tag{15.38}$$

Equation (15.38) shows that Q_{B} depends on the applied voltage V_{c} in the channel. If the effects of the work function difference and oxide charges are included, then the threshold voltage given above must be modified, and (15.37) becomes

$$V'_{\mathrm{TH}} = \phi_{\mathrm{ms}} + \varphi_{\mathrm{si}} - \frac{Q_0}{C_{\mathrm{ox}}} - \frac{Q_{\mathrm{B}}}{C_{\mathrm{ox}}} = V_{\mathrm{FB}} + \varphi_{\mathrm{si}} + \sqrt{2q\varepsilon_{\mathrm{s}}\varepsilon_0\varphi_{\mathrm{si}}/C_{\mathrm{ox}}}, \tag{15.39}$$

where ϕ_{ms} is the metal–semiconductor work function difference and Q_0/C_{ox} is the voltage shift due to the oxide charges. Quantity V_{FB} is the flat-band voltage defined by (15.32). Furthermore, if a bias voltage is applied to the substrate (body) of the MOSFET, then the threshold voltage becomes

$$V''_{\mathrm{TH}} = V_{\mathrm{FB}} + \varphi_{\mathrm{si}} + \frac{\sqrt{2q\varepsilon_{\mathrm{s}}\varepsilon_0 N_{\mathrm{A}}(\varphi_{\mathrm{si}} + V_{\mathrm{sb}})}}{C_{\mathrm{ox}}}, \tag{15.40}$$

where V_{sb} is the substrate bias voltage. Beyond strong inversion, the charge condition in the surface region becomes

$$Q_s = Q_I + Q_B = Q_I - qN_A W_{d\,max}, \tag{15.41}$$

where Q_I and Q_B denote the channel and bulk charges, respectively; $W_{d\,max}$ is the maximum depletion layer width at the onset of strong inversion and is given by (15.18). The applied gate voltage V_{GS}, corresponding to strong inversion, is given by

$$V_{GS} = -\frac{Q_s}{C_{ox}} + \varphi_{si}. \tag{15.42}$$

As an approximation, the channel charge Q_I may be related to the threshold voltage V_{TH} by

$$Q_I = -C_{ox}(V_{GS} - V_{TH}). \tag{15.43}$$

Therefore, the channel conductance g_I can be expressed in terms of the gate voltage and threshold voltage. From (15.43), the channel conductance is given by

$$g_I = -\frac{Z\mu_n Q_I}{L} = \frac{Z}{L}\mu_n C_{ox}(V_{GS} - V_{TH}). \tag{15.44}$$

Equation (15.44) accurately describes the channel conductance in the strong inversion region (i.e., $V_G > V_{TH}$). In this region, the channel conductance is a linear function of the applied gate voltage. In fact, most MOSFETs operate in this region.

15.4.3. Current–Voltage Characteristics

To analyze the current–voltage (I–V) relationship of a MOSFET, consider an n-channel silicon MOSFET as shown in Figure 15.11. For a long-channel silicon MOSFET, typical channel length may vary between 3 and 10 μm. With the advances in electron-beam lithography technology, silicon MOSFETs with submicron ($L < 0.13$ μm) channel length have been developed recently. In fact, for ULSI design submicron device geometries are routinely employed in present-day IC layouts.

A qualitative description of the current–voltage characteristics for an n-channel silicon MOSFET is given first. To facilitate the analysis, it is assumed that the source and substrate terminals are grounded and a gate voltage V_{GS} greater than the threshold voltage V_{TH} is applied to the gate electrode in order to induce an inversion surface channel. When a small drain voltage V_{DS} is applied to the drain electrode, current flows from the source to the drain electrodes via the inversion channel. For small V_{DS}, the channel acts as a variable resistor and the drain current I_{DS} varies linearly with V_{DS}. This corresponds to the *linear region* operation (i.e., $V_{GS} > V_{TH}$ and $V_{DS} \approx 0$). As the drain voltage continues to increase, the space-charge region under the channel and near the drain region widens and eventually reaches a pinch-off condition in which the channel depth x_I at $y = L$ becomes zero.

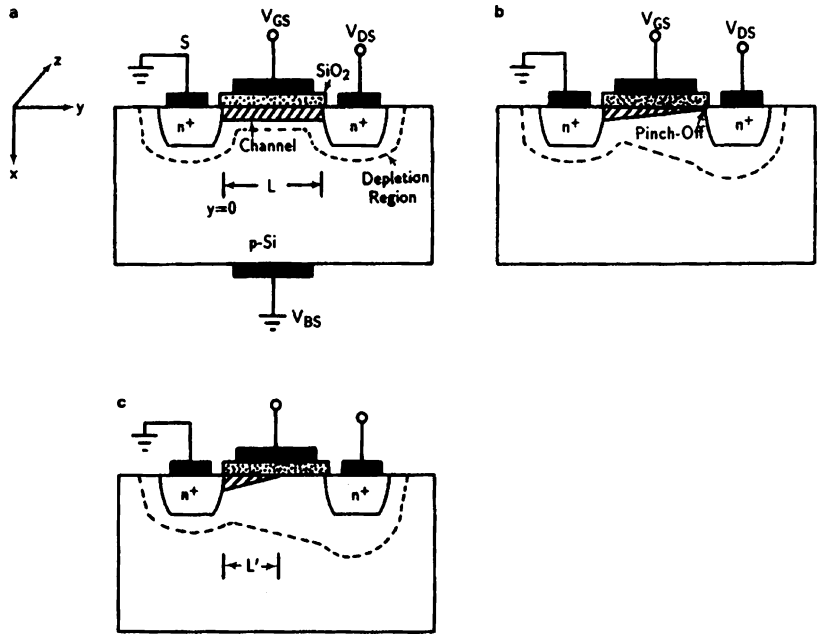


FIGURE 15.13. An n-channel MOSFET operating under different drain voltages: (a) linear region, $V_{GS} > V_{TH}$ and V_{DS} small, (b) onset of saturation $V_{GS} > V_{TH}$ and $V_{DS} = V_{Dsat}$, (c) beyond saturation $V_G > V_{TH}$ and $V_D > V_{Dsat}$.

Beyond the pinch-off point, the drain current remains essentially constant as the drain voltage continues to increase. This region is called the saturation region since the drain current becomes saturated. Figure 15.13 shows the schematic diagrams of a silicon MOSFET operating under different drain voltage conditions. The effect of drain voltage on the width of the depletion region across the source, channel, and drain junctions is clearly illustrated in this figure. Figure 15.13a shows the MOSFET operating in the linear region, and Figure 15.13b shows the onset of saturation operation. When the drain voltage is increased beyond saturation (i.e., $V_{DS} > V_{Dsat}$), the effective channel length L' is reduced (i.e., $L' < L$); however, as shown in Figure 15.13c, the saturation drain current that flows from the source to the drain remains unchanged.

For a long-channel MOSFET (i.e., $L \gg W_d$), the drain current versus drain voltage relationship can be derived from the gradual channel approximation. In this approximation, it is assumed that (i) the transverse electric field along the x-direction inside the channel is much larger than the longitudinal electric field in the y-direction (see Figure 15.13a), (ii) both electric fields are independent of each other, (iii) the effects of fixed oxide charges, interface traps, and metal–semiconductor work function difference are negligible, (iv) carrier mobility in the inversion layer is constant, and (v) current conduction in the channel consists of only the drift component. It should be noted that the transverse electric field in

the x-direction is to induce an inversion channel, while the longitudinal electric field in the y-direction is to produce a drain current that flows through the surface inversion channel.

(i) Linear region operation ($V_{DS} \approx 0$). The drain current versus drain voltage for a MOSFET operating in the linear region under strong inversion and small drain voltage conditions can be derived as follows. Let us consider a small incremental section along the channel in the y-direction as shown in Figure 15.13a for $V_{GS} > V_{TH}$. Under this bias condition, mobile carriers are induced in the inversion layer. If the channel voltage is equal to zero, the relationship between the mobile charge Q_I in the inversion layer and the gate voltage is given by (15.43). When a drain voltage V_{DS} is applied to the drain electrode (with source electrode grounded), a channel potential V_c is established along the y-direction in the inversion channel. Thus, the new expression for the induced charge Q_I in the channel is given by

$$Q_I(y) = -C_{ox}\left[V_G - V_{TH} - V_c(y)\right]. \tag{15.45}$$

The drain current I_{DS}, which is due to the majority carrier (electrons) flow, can be written as

$$I_{DS} = Z\mu_n Q_I \mathcal{E}_y. \tag{15.46}$$

Now, substituting $\mathcal{E}_y = -\mathrm{d}V_c/\mathrm{d}y$ and (15.45) into (15.46) and integrating over the distance from $y = 0$ to $y = L$ and potential from $V_c = 0$ to $V_c = V_{DS}$, one obtains

$$I_{DS} = \frac{C_{ox}\mu_n Z}{L}\left(V_{GS} - V_{TH} - \frac{V_{DS}}{2}\right)V_{DS}. \tag{15.47}$$

In deriving (15.47) it is assumed that V_{TH} is independent of V_c [see (15.37) and (15.38)]. This approximation could lead to a substantial error, because V_{TH} generally increases toward the drain region owing to the increase of bulk charge Q_B with the applied drain voltage. If (15.38) is used for Q_B, then a more accurate drain current versus drain voltage relationship can be derived, and the result is given by

$$I_{DS} = \frac{C_{ox}\mu_n Z}{L}\left\{\left(V_{GS} - \phi'_{ms} - \psi_{si} + \frac{Q_{ox}}{C_{ox}} - \frac{V_{DS}}{2}\right)V_{DS} - \frac{2}{3}\frac{\sqrt{2q\varepsilon_s\varepsilon_0 N_A}}{C_{ox}}\left[(V_{DS} + \varphi_{si})^{3/2} - \varphi_{si}^{3/2}\right]\right\}. \tag{15.48}$$

This equation shows that for a given gate voltage V_{GS}, the drain current initially increases linearly with drain voltage (i.e., linear region), and then gradually levels off, reaching a saturated value (i.e., saturation region). It is seen that the simplified expression given by (15.47) usually predicts a higher value of I_{DS} at large V_{DS} than predicted by (15.48). However, the simple expression given by (15.47) for I_{DS} offers better physical insight for the device operation, and hence it is easier to obtain a first-order prediction of the MOSFET's performance using (15.47) in a digital circuit design.

(ii) Saturation region operation ($V_{DS} > V_{Dsat}$). In the linear region, it is seen that the inversion layer is formed throughout the semiconductor surface between the source and the drain. As the drain voltage increases, the inversion layer at the drain side of the channel will gradually diminish. When the drain voltage is increased to the point such that the charge in the inversion layer at $y = L$ becomes zero, the so-called pinch-off condition is reached. The saturation drain voltage and drain current at the pinch-off point are designated by V_{Dsat} and I_{Dsat}, respectively. Beyond the pinch-off point, further increase of drain voltage will not increase the drain current significantly, and the saturation region is reached. This is shown in Figure 15.13b. Under the saturation condition, the channel charge at the drain side of the channel is reduced to zero (i.e., $Q_I = 0$ at $y = L$). The saturation drain voltage V_{Dsat} is obtained from (15.37) and (15.38) by setting $V = 0$, which yields

$$V_{D\,sat} = V_{GS} - V_{TH} = V_{GS} - \varphi_{si} + \frac{\sqrt{2q\varepsilon_0\varepsilon_s N_A \varphi_{si}}}{C_{ox}}. \tag{15.49}$$

The saturation drain current is obtained by substituting (15.49) (i.e., $V_{Dsat} = V_{GS} - V_{TH}$) into (15.47). Thus,

$$I_{D\,sat} = \frac{\mu_n Z C_{ox}}{2L} (V_{GS} - V_{TH})^2 . \tag{15.50}$$

Equation (15.50) predicts that the drain current in the saturation region is a quadratic function of the gate voltage. It is noted that (15.50) is valid at the onset of saturation. However, beyond this point the drain current can be considered a constant. Thus, (15.50) is still valid for $V_{DS} > V_{Dsat}$. Figure 15.14 shows typical drain current versus drain voltage curves for an n-channel silicon MOSFET under

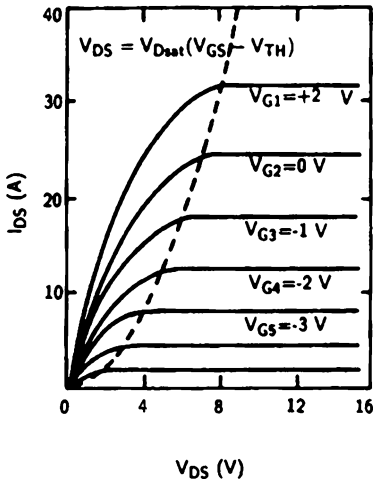


FIGURE 15.14. Drain current versus drain voltage curves for an n-channel MOSFET. The dashed line denotes the locus of the onset of current saturation under different gate bias voltages.

different gate bias conditions. The dashed line is the locus of the $I_{D\,sat}$ versus $V_{D\,sat}$ plot, which represents the onset of current saturation in the drain region.

(iii) Cutoff region operation ($V_{GS} \ll V_{TH}$). If the gate voltage is much smaller than the threshold voltage, then there would be no inversion layer formed in the channel region. Under this condition, the MOSFET acts like two p-n junction diodes connected back to back with no current flow in either direction between the source and drain electrodes. Thus, in the cutoff region, the MOSFET is open-circuited.

(iv) Subthreshold region operation. Another important region of operation for a MOSFET is known as the subthreshold region. In this region, the gate voltage is smaller than the threshold voltage, and the semiconductor surface is in weak inversion (i.e., $\varphi_s < 2\varphi_B$). The drain current in the weak inversion region is called the subthreshold current. The subthreshold region operation is particularly important for low-voltage and low-power applications such as switching devices used in digital logic and memory applications.

In the weak inversion region, the drain current is dominated by diffusion, and hence the drain current can be derived in a similar way to that by which the collector current is derived in a bipolar junction transistor (BJT) with a uniformly doped base (i.e., the MOSFET can be treated as an n^+-p-n BJT). It can be shown that in the subthreshold region, the drain current varies exponentially with gate voltage (i.e., $I_{DS} \approx e^{qV_{GS}/k_BT}$). For drain voltages greater than $3k_BT/q$, the drain current becomes independent of drain voltage. A detailed derivation of the drain current versus drain voltage expression for a silicon MOSFET operating in the subthreshold region can be found in (4).

15.4.4. Small-Signal Equivalent Circuits

The small-signal equivalent circuit for a MOSFET operating in a common-source configuration is shown in Figure 15.15. An ideal MOSFET has an infinite input resistance (R_i) and a current generator ($g_m V_{GS}$) at the output terminal of the device. However, in a practical MOSFET, several physical parameters must be included to reflect the nonideality and parasitic effects of the device.

In small-signal analysis two important device parameters must be considered, namely, the channel conductance g_d and the mutual transconductance g_m. In linear region operation, both parameters can be derived directly from (15.47), and the results are given by

$$g_d = \left.\frac{\partial I_{DS}}{\partial V_{DS}}\right|_{V_{GS}\,=\,\text{const}} = \frac{\mu_n Z C_{ox}}{L}\,(V_{GS} - V_{TH}) \tag{15.51}$$

and

$$g_m = \left.\frac{\partial I_{DS}}{\partial V_{GS}}\right|_{V_{DS}\,=\,\text{const}} = \frac{\mu_n Z C_{ox}}{L}\,V_{DS}. \tag{15.52}$$

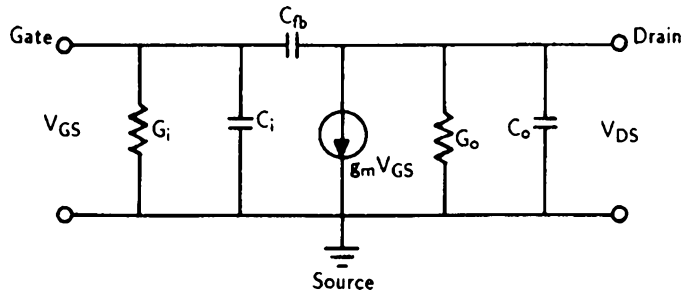


FIGURE 15.15. The small-signal equivalent circuit for a MOSFET in a common-source configuration.

It is seen that in the linear region, the drain conductance g_d is ohmic and depends linearly on the gate voltage (except at high gate voltages in which the carrier mobility decreases with increasing gate voltage). The inverse of the drain conductance is usually referred to as the "on" resistance. Thus, in linear region operation, the MOSFET is operating essentially as a voltage-controlled resistor in circuit applications.

In the saturation region, the mutual transconductance can be obtained by differentiating the saturation drain current given by (15.50) with respect to the gate voltage, which yields

$$g_m = \left.\frac{\partial I_{DS}}{\partial V_{GS}}\right|_{V_{DS}\,=\,\text{const}} = \frac{Z\mu_n C_{ox}}{L}(V_{GS} - V_{TH}). \tag{15.53}$$

Equation (15.53) shows that the transconductance varies linearly with gate voltage in the saturation region.

In the saturation region, the drain current remains constant for any drain voltage beyond the pinch-off point. This is true only for the ideal case. In practice, the drain resistance r_d has a finite value for $V_{DS} > V_{D\,sat}$, and hence one can define a drain resistance in the saturation region as

$$r_{d\,sat} = \left.\frac{\partial V_{DS}}{\partial I_{DS}}\right|_{V_{GS}\,=\,\text{const}}. \tag{15.54}$$

Therefore, from the slope of the I_{DS} versus V_{DS} plot, $r_{d\,sat}$ can be determined. It is noted that the small-signal equivalent circuit shown in Figure 15.15 includes all the device parameters discussed above. However, it is worth noting that the gate-to-drain capacitance C_{gd} is a key parameter that controls the high-frequency characteristics of a MOSFET, which is known as the Miller effect.

Another important parameter relating to the small-signal operation of a MOSFET is the maximum frequency in which its short-circuit current gain drops to unity. This frequency is usually referred to as the unity gain cutoff frequency f_T. Under the unit gain condition, the input current through the input capacitance

($=2\pi f_T C_g V_{GS}$) is equal to the output drain current ($\approx g_m V_{GS}$), as shown in Figure 15.15. Thus, in the linear region of operation (i.e., $V_{DS} \leq V_{D\,sat}$), the unity current gain cutoff frequency can be expressed by

$$f_T = \frac{g_m}{2\pi C_g} \approx \frac{\mu_n V_{DS}}{2\pi L^2}, \tag{15.55}$$

where C_g is the total gate capacitance. Equation (15.55) is obtained by using (15.52) for g_m and $C_g = ZLC_{ox}$. Thus, for small-signal high-frequency operation, short gate length and high electron mobility in the channel are highly desirable for a MOSFET.

In the saturation region, the unity current gain cutoff frequency is given by

$$f_T = \frac{g_m}{2\pi C_g} \approx \frac{v_s}{2\pi L}, \tag{15.56}$$

where $v_s = (\mu_n V_{D\,sat})/L$ is the saturation velocity in the channel. Equation (15.56) is obtained by using (15.53) for g_m and $C_g = ZLC_{ox}$.

In the saturation region operation, the MOSFET device may be used as an amplifier or a closed switch. A closed switch operates in the region where the gate voltage is smaller than the threshold voltage. Under this operation condition, no inversion layer is formed. As a result, the MOSFET behaves like two p-n junction diodes connected back to back, and no current is expected to flow in either direction. The device acts as an open-circuit switch in this region of operation.

The theoretical expressions presented in this section are valid for the long-channel MOSFET in which the channel length is much larger than the depletion layer width of the source and drain junctions. However, with recent advances in silicon VLSI technologies, reduction of channel lengths to less than a micron has become a reality, and gate lengths in the submicron region have also been widely used. As the size of MOSFETs continues to scale down, the channel length becomes equal to or less than the depletion layer width of the source and drain junctions, and hence departure from long-channel behavior occurs in short-channel devices. The short-channel effects are the results of the two-dimensional (2-D) potential distribution and the high electric field in the channel region. The gradual channel approximation used in analyzing the long-channel MOSFET presented in this section is no longer valid, and must be modified to take into account the short-channel effects. For example, the 2-D potential distribution will cause degradation of the subthreshold behavior, the dependence of threshold voltage on channel length and biasing voltages, as well as the failure of current saturation due to punch-through. As the electric field increases, the channel mobility becomes field-dependent, and eventually velocity saturation takes place. At high electric fields, carrier multiplication occurs near the drain region, leading to substrate current and parasitic bipolar transistor action. High fields can also cause hot-carrier injection into the oxide. This can lead to oxide charging, transconductance degradation, and a shift in the threshold voltage. In

short, in order to further the advancement of short-channel MOSFETs with submicron geometries, it is essential that these short-channel effects be eliminated or minimized.

15.4.5. Scaled-Down MOSFETs

In addition to the basic enhancement-mode and depletion-mode MOSFET structures discussed above, a variety of new MOSFET structures such as high-performance MOSFETs (HMOS), double-diffused MOSFETs (DMOS), vertical or V-shaped grooved MOSFETs (VMOS), U-shaped grooved MOSFETs (UMOS), Schottky-barrier source and drain MOSFETs, lightly doped drain (LDD) FETs, and thin-film transistors (TFTs) have been widely investigated. Furthermore, recent development of several new SOI technologies [e.g., Separation by IMplantation of OXygen (SIMOX) and Wafer Bonding (WB) techniques] enables the fabrication of MOSFETs and BJTs on these SOI substrates for various radiation-hard digital and low-power IC applications. Using these new structures, high-performance FETs have been developed for a wide variety of applications. Performance improvements include higher speed, lower power consumption, higher packing density, higher radiation tolerance, and higher power handling capabilities.

For VLSI applications, the dimensions and voltages used in conventional MOSFETs must be reduced drastically. One way to avoid the undesirable short-channel effects discussed in the foregoing section while maintaining the long-channel behavior of the MOSFETs is simply to scale down all dimensions and voltages of the long-channel MOSFETs. The basic idea underlying the scaling-down theory is to keep the electric field strength invariant while reducing the device sizes and voltages. Smaller device geometries will then be translated to shorter transit times (higher device speed) and lower voltages. Both vertical and horizontal dimensions must be scaled down accordingly. The source and drain junction depths must be reduced proportionally to prevent sidewall diffusion from encroaching on the effective channel diffusion length, and the substrate dopant density must be increased so that the depletion region can be scaled down accordingly; otherwise, punch-through between the source and drain may occur. In addition, the oxide thickness must also be scaled down to maintain the gate field at a reduced gate voltage and the height of the oxide steps on the surface must be reduced; otherwise, it may cause breaks in the thinner interconnects. New gate insulating materials with high dielectric constant are needed for submicron MOSFETs.

The scaling down of dimensions and device parameters in a MOSFET can be achieved as follows. If the channel length (L), channel width (Z), gate oxide thickness (d_{ox}), and voltages (V_{GS}, V_{DS}, V_B) (substrate voltage) are all scaled down by a factor $K (K > 1)$, and the substrate dopant density (N_A) is increased by the same factor, then the device parameters for the scaled-down MOSFETs can be modified from the long-channel MOSFETs according to the following

formulas:

$$C'_{\text{ox}} = \frac{\varepsilon_{\text{ox}}\varepsilon_0}{d_{\text{ox}}/K} = KC_{\text{ox}}, \tag{15.57}$$

$$W'_{\text{d}} = \sqrt{\frac{2\varepsilon_{\text{s}}\varepsilon_0(V_{\text{bi}}/K)}{qKN_{\text{A}}}} = \frac{W_{\text{d}}}{K}, \tag{15.58}$$

$$\varphi'_{\text{B}} = E_{\text{F}} - E_{\text{i}} = -k_{\text{B}}T \ln\left(\frac{KN_{\text{A}}}{n_{\text{i}}}\right), \tag{15.59}$$

$$Q'_{\text{B}} = -q(KN_{\text{A}})(W_{\text{d}}/K) \approx Q_{\text{B}}, \tag{15.60}$$

$$V'_{\text{TH}} = \phi_{\text{ms}} + \frac{2\varphi_{\text{B}}}{q} - \frac{(Q_{\text{ox}} + Q_{\text{B}})}{KC_{\text{ox}}} \approx \frac{V_{\text{TH}}}{K}, \tag{15.61}$$

$$I'_{\text{DS}} = \frac{\mu_{\text{n}}(KC_{\text{ox}})(Z/K)}{L/K}\left[(V_{\text{GS}}/K - V_{\text{TH}}/K)(V_{\text{DS}}/K) - \frac{1}{4}(V_{\text{DS}}/K)^2\right] = \frac{I_{\text{DS}}}{K}, \tag{15.62}$$

$$I'_{\text{D sat}} = \frac{\mu_{\text{n}}(KC_{\text{ox}})(Z/K)}{2L/K}(V_{\text{GS}}/K - V_{\text{TH}}/K)^2 = \frac{I_{\text{D sat}}}{K}, \tag{15.63}$$

$$g'_{\text{m}} = \frac{\partial(I_{\text{DS}}/K)}{\partial(V_{\text{DS}}/K)} = g_{\text{m}}, \tag{15.64}$$

$$C'_{\text{gs}} = \frac{2}{3}(L/K)(Z/K)(KC_{\text{ox}}) = \frac{C_{\text{gs}}}{K}, \tag{15.65}$$

$$t'_{\text{d}} = \frac{C_{\text{gs}}/K}{g_{\text{m}}} = \frac{t_{\text{d}}}{K}. \tag{15.66}$$

From (15.57) to (15.66) it can be shown that the device area (A) is reduced by a factor of K^2, power dissipation (IV) is reduced by a factor of K^2, and the power-delay product (IVt'_{d}) is reduced by a factor of K^3. The power dissipation per unit area (IV/A), however, remains unchanged, although considerable gain in the area and power-delay product (an important figure of merit) is expected by scaling down the MOSFET's dimensions. However, a drastic reduction in device size will also increase the importance of other secondary effects, such as narrow-channel and short-channel effects on the threshold voltage, etc. These effects must also be considered in the device modeling. Finally, weak inversion occurs at $V_{\text{GS}} < V_{\text{TH}}$, and hence subthreshold conduction must also be considered.

Figures 15.16a and b show a conventional long-channel MOSFET and a scaled-down MOSFET, respectively. In the scaled-down MOSFET, both drain voltage V_{DS} and threshold voltage V_{TH} are scaled down by a factor of K, and the number of devices per unit area and power dissipation per unit cell are increased by a factor of K^2. Also, the delay time due to transit across the channel is decreased by a factor of K. It is interesting to note that the subthreshold current remains essentially the same for both devices, since the subthreshold voltage swing remains the same. The junction built-in potential and surface potential at the onset of weak

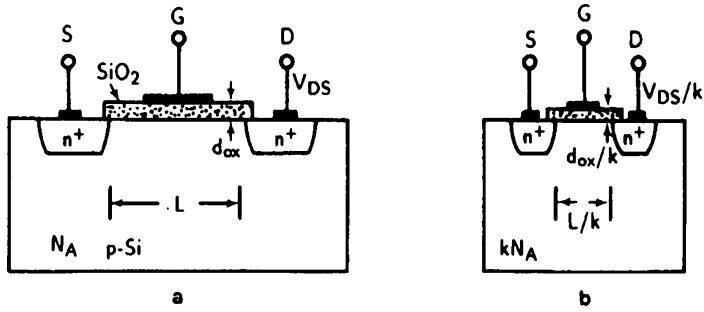


FIGURE 15.16. (a) A conventional long-channel MOSFET and (b) a scaled-down (by a factor K) short channel MOSFET.

inversion do not scale down with the dimensions, and change by only about 10% for a tenfold increase in substrate dopant density. Additionally, the range of gate voltages between depletion and strong inversion is about 0.5 V. However, the parasitic capacitance may not scale down, and the interconnect resistance will usually increase in a scaled-down MOSFET.

Another important circuit application using n- and p-channel MOSFET is the fabrication of CMOS and BiCMOS for logic and memory IC applications. The CMOS device refers to the complementary n- and p-channel MOSFET pairs fabricated on the same chip to form complementary MOS (CMOS) transistors. It is the most popular technology used in present-day IC design for logic circuits and memory cell applications. The reasons for the success of CMOS are due to its low power consumption and good noise immunity. In fact, currently only CMOS technology is used in advanced IC manufacturing because of the low power of dissipation requirement.

A CMOS inverter is used as the basic component of CMOS logic circuits. In a CMOS inverter, the gates of n- and p-channel MOSFETs are connected and serve as the input node of the inverter. The drain electrodes of the two transistors are also connected and serve as the output node of the inverter. The source and substrate contacts of n-channel MOSFETs are grounded, whereas those of p-channel MOSFETs are connected to the power supply. Both n- and p-channel MOSFETs are enhancement-mode FETs. The key feature of the CMOS inverter is that when the output is in a steady logic state only one transistor is on, and the current flow from the power supply to the ground is thus very low (equal to the leakage current of the device). In fact, there is significant current conduction during the short transient period when both transistors are on. Therefore, the power consumption in CMOS inverters is very low in the static state compared to other types of logic circuits.

Although CMOS devices have the advantages of low power consumption and high packing density, which make them suitable for manufacturing complex circuits (VLSI and ULSI), they suffer from low drive capability compared with bipolar technology, limiting their circuit performance. BiCMOS is a technology that

integrates both CMOS and bipolar device structures on the same chip. A BiCMOS circuit contains mostly CMOS devices, with a relatively small number of bipolar devices. The bipolar devices have better performance than their CMOS counterparts without consuming too much extra power. However, this performance enhancement is achieved at the expense of extra manufacturing complexity and costs. High-performance BiCMOS circuits have also been made on SOI wafers.

15.5. SOI MOSFETS

15.5.1. Introduction

Silicon-on-insulator (SOI) devices and circuits have progressed rapidly during the past decade. The advantages of SOI over bulk silicon have been demonstrated in terms of performance and reliability. Recent interest has turned to possible use of fully depleted (FD) SOI films for deep submicron circuits. SOI MOSFETs are free of some of the effects that tend to reduce the performance of their bulk counterparts. In particular, SOI MOSFETs offer reduced junction capacitances, a very small body factor, and hence a near-ideal subthreshold slope and high current drive. Other areas in which the capability of SOI technology has been identified include low-power and mixed-technology circuits. In this section, the basic properties, device structures, and characteristics of SOI MOSFETs are described.

Ultra-large-scale integration (ULSI) is the mainstream of the microelectronics industry. It is essentially in this high-priority arena that the very attractive arguments of SOI technology have to be materialized. Thin-film CMOS, BiCMOS, and complementary bipolar structures can overcome the severe technological limitations that bulk silicon is expected to encounter in scaling below 0.08 μm: the failure of LOCOS isolation, threshold voltage roll-off, increased subthreshold slope, hot-electron degradation, soft errors, and latch-up. In addition to being a high-speed, low-power process, new innovative structures for SOI CMOS devices can further increase the performance advantages over the equivalent bulk CMOS devices. Moreover, SIMOX-based BiCMOSs or complementary bipolar ICs may have circuit density comparable to that of MOS circuits. They will enjoy higher speed because of a denser layout and lower parasitic capacitances, and the simpler fabrication process will render them more cost-effective. Most of the commercial SOI wafers are produced by SIMOX and WB techniques.

Depending on the silicon film thickness and channel doping concentration of the SOI wafer, partially depleted (PD) and fully depleted (FD) SOI MOSFET devices can be fabricated on SOI wafers. The FD SOI MOSFET is fabricated on SOI wafer with ultrathin silicon film, and the PD SOI MOSFET uses a thicker silicon film on SOI with general characteristics similar to the bulk silicon MOSFET.

In a thin-film SOI MOSFET, the silicon film thickness is smaller than the maximum depletion layer width $x_{d\,max}$ from the Si–SiO_2 interface, where, $x_{d\,max} = \sqrt{4\varepsilon_0\varepsilon_s\varphi_F/qN_D}$; $\varphi_F = (k_BT/q)\ln(N_D/n_i)$ is the Fermi potential. In this case, the

silicon film is fully depleted at threshold, irrespective of the bias voltage applied to the back gate, with the exception of a thin inversion or accumulation layer formed at the back interface under large positive or negative bias conditions. Such a device is called an FD MOSFET. FD MOSFETs are virtually free of kink effect, if their back interface is not in accumulation. Among all types of SOI devices, FD MOSFETs with depleted back interface exhibit the most attractive features such as low electric field, high transconductance, excellent short-channel effect, and a near-ideal subthreshold slope.

In a thick-film SOI MOSFET, the thickness of silicon film is twice as large as the depletion layer width (i.e., $t_{\mathrm{Si}} > 2x_{\mathrm{d\,max}}$). In this case, there is no interaction between the depletion regions at the front and the back interfaces, and there exists a neutral region beneath the front depletion zone. Such a thick-film SOI device is called a PD SOI MOSFET. If the neutral region (called the body) is connected to the ground by a body contact or body tie, then the characteristics of this MOSFET will be the same as a bulk MOSFET device. If the body is left electrically floating, then two parasitic effects will occur in this device: one is the appearance of a kink effect in the output characteristics of the device, and the second is the presence of a parasitic, open-base bipolar transistor action (i.e., the so-called latch-up effect) between the source and drain of this MOSFET device.

Figure 15.17 shows a cross-sectional view of a thin-film n-channel FD SOI MOSFET device structure. The electrical characteristics of an FD SOI MOSFET are now described.

15.5.2. Electrical Characteristics

(i) The threshold voltage. The threshold voltage of an FD n-channel SOI MOSFET can be obtained by solving the Poisson equation ($\mathrm{d}^2\varphi/\mathrm{d}x^2 = qN_{\mathrm{a}}/\varepsilon_0\varepsilon_{\mathrm{r}}$) using the depletion approximation. By integrating the Poisson equation twice one obtains the potential $\varphi(x)$ as a function of depth x in the silicon film, which reads

$$\varphi(x) = \frac{qN_{\mathrm{A}}}{\varepsilon_0\varepsilon_{\mathrm{s}}}x^2 + \left(\frac{\varphi_{\mathrm{s2}} - \varphi_{\mathrm{s1}}}{t_{\mathrm{si}}} - \frac{qN_{\mathrm{A}}t_{\mathrm{si}}}{2\varepsilon_0\varepsilon_{\mathrm{s}}}\right)x + \varphi_{\mathrm{s1}}, \tag{15.67}$$

where φ_{s1} and φ_{s2} denote the potentials at the front and back of the Si–SiO_2 interfaces, respectively. The front- and back-gate voltages are given by $V_{\mathrm{g1}} = \varphi_{\mathrm{s1}} + \phi_{\mathrm{ox1}} + \phi_{\mathrm{ms1}}$ and $V_{\mathrm{g2}} = \varphi_{\mathrm{s2}} + \phi_{\mathrm{ox2}} + \phi_{\mathrm{ms2}}$; ϕ_{ms1} and ϕ_{ms2} are the front and back work function differences, respectively. Using these relationships, one can find the threshold voltage of the front interface by assuming that $\varphi_{\mathrm{s1}} = 2\varphi_{\mathrm{F}}$.

If the back surface is in accumulation, φ_{s2} is pinned to be approximately equal to zero. The thershold voltage $V_{\mathrm{th1\,acc2}}$ can be obtained from (15.67), where $V_{\mathrm{th1\,acc2}} = V_{\mathrm{g1}}$ is calculated at $\varphi_{\mathrm{s2}} = 0$, $Q_{\mathrm{inv1}} = 0$, and $\varphi_{\mathrm{s1}} = 2\varphi_{\mathrm{F}}$. One obtains

$$V_{\mathrm{th1,acc2}} = \phi_{\mathrm{ms1}} - \frac{Q_{\mathrm{ox1}}}{C_{\mathrm{ox1}}} + \left(1 + \frac{C_{\mathrm{si}}}{C_{\mathrm{ox1}}}\right)2\varphi_{\mathrm{F}} - \frac{Q_{\mathrm{depl}}}{2C_{\mathrm{ox1}}}. \tag{15.68}$$

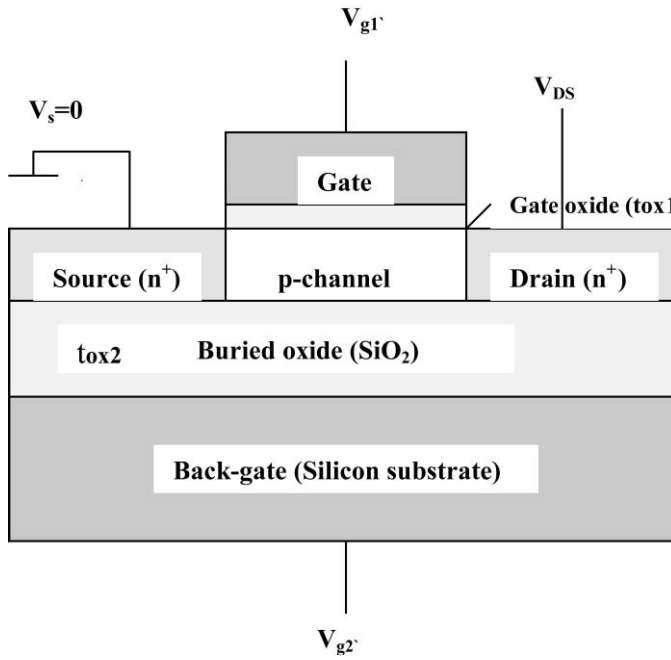


FIGURE 15.17. Cross-sectional view of a thin-film n-channel fully depleted (FD) SOI MOSFET showing some of the parameters used in the derivation of equations in this section.

If the back surface is inverted, $\varphi_{s2} = 2\varphi_F$ is pinned to approximately $V_{th1,inv2} = V_{g1}$ with $\varphi_{s2} = 2\varphi_F$, $Q_{inv1} = 0$, and $\varphi_{s1} = 2\varphi_F$. One then has

$$V_{th1,inv2} = \varphi_{ms1} - \frac{Q_{ox1}}{C_{ox1}} + 2\varphi_F - \frac{Q_{depl}}{2C_{ox1}}. \tag{15.69}$$

If the back surface is depleted, then φ_{s2} depends on the back-gate bias voltage V_{g2} and its values can vary between 0 and $2\varphi_F$. The back-gate voltage for which the back gate reaches accumulation with the front interface being at threshold, V_{g2}, is obtained by setting $\varphi_{s1} = 2\varphi_F$ and $\varphi_{s2} = 0$. Similarly, the value of the back-gate voltage for which the back interface reaches inversion, $V_{g2,inv}$, can be obtained by setting $\varphi_{s1} = 2\varphi_F$ and $\varphi_{s2} = 2\varphi_F$. When $V_{g2,acc} < V_{g2} < V_{g2,inv}$, the front threshold voltage is obtained by setting $\varphi_{s1} = 2\varphi_F$ and $Q_{inv} = 0$. One thus obtains

$$V_{th1,depl2} = V_{th1,acc2} - \frac{C_{si}C_{ox2}}{C_{ox1}(C_{si} + C_{ox2})}(V_{g2} - V_{g2,acc}). \tag{15.70}$$

The variation of threshold voltage V_{th1} for an FD SOI MOSFET as a function of the back-gate bias voltage V_{g2} is illustrated in Figure 15.18.

(ii) The body effect. In a bulk MOSFET the body effect is defined as the dependence of the threshold voltage on the substrate bias, while in an SOI MOSFET it is

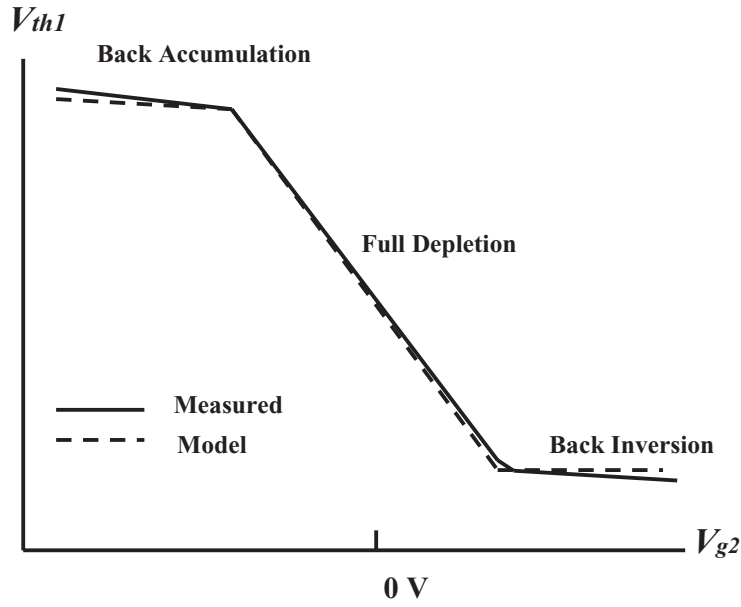


FIGURE 15.18. Variation of the front-gate threshold voltage with the back-gate bias for an SOI MOSFET.

defined as the dependence of the threshold voltage on the back-gate bias. In a bulk n-channel MOSFET, the threshold voltage can be expressed as

$$V_{\rm th} = \varphi_{\rm ms} + 2\varphi_{\rm F} - \frac{Q_{\rm ox}}{C_{\rm ox}} + \frac{Q_{\rm B}}{C_{\rm ox}} = \varphi_{\rm ms} + 2\varphi_{\rm F} - \frac{Q_{\rm ox}}{C_{\rm ox}} + \frac{\sqrt{2\varepsilon_0\varepsilon_{\rm s} q N_{\rm A}(2\varphi_{\rm F} - V_{\rm B})}}{C_{\rm ox}}. \tag{15.71}$$

If one introduces a body (or back-gate) effect parameter $\gamma = \sqrt{2\varepsilon_0\varepsilon_{\rm s} q N_{\rm A}}/C_{\rm ox}$ in (15.71), then the equation can be rewritten as

$$V_{\rm th} = \varphi_{\rm ms} + 2\varphi_{\rm F} - \frac{Q_{\rm ox}}{C_{\rm ox}} + \gamma\sqrt{2\varphi_{\rm F}} + \gamma(\sqrt{2\varphi_{\rm F} - V_{\rm B}} - \sqrt{2\varphi_{\rm F}}). \tag{15.72}$$

It is noted that the last term in (15.72) describes the body effect of the bulk or SOI MOSFETs. When a negative bias (with respect to the source) is applied to the substrate, the threshold voltage increases with the square root of the substrate bias. If one defines the threshold voltage at zero bias as $V_{\rm th0}$, then (15.72) becomes

$$V_{\rm th}(V_{\rm B}) = V_{\rm th0} + \gamma(\sqrt{2\varphi_{\rm F} - V_{\rm B}} - \sqrt{2\varphi_{\rm F}}). \tag{15.73}$$

For PD SOI MOSFETs, the back-gate effect can be neglected (i.e., $\gamma = 0$), because there is no coupling between the front gate and the back gate of the device. In a FD SOI MOSFET, the body effect can be described by

$$\frac{{\rm d}V_{\rm th1}}{{\rm d}V_{\rm g2}} = -\frac{C_{\rm si}C_{\rm ox2}}{C_{\rm ox1}(C_{\rm si} + C_{\rm ox2})} = \frac{-\varepsilon_{\rm si}C_{\rm ox2}}{C_{\rm ox1}(t_{\rm si}C_{\rm ox2} + \varepsilon_{\rm si})} = \gamma'. \tag{15.74}$$

It is noted that γ' is dimensionless for the thin-film FD SOI MOSFETs, and the threshold voltage is linearly dependent on the back-gate bias. It is practical to linearize the body effect in bulk devices by introducing a body factor denoted by $n = 1 + C_D/C_{ox}$, where $C_D = \varepsilon_0\varepsilon_{si}/x_{d\,max}$ is the depletion capacitance. In the FD SOI MOSFET the body effect factor is defined by $n = 1 + \gamma'$. Typical values for the body factor are $n = 1.3$ to 1.5 for bulk MOSFETs and $n = 1.05$ to 1.1 for FD SOI MOSFETs.

(iii) Output characteristics and transconductance. The output characteristics of an FD SOI MOSFET are similar to those of a bulk MOSFET. In both devices the saturation drain current can be expressed by

$$I_{Dsat} \approx \frac{W\mu_n C_{ox1}}{2nL}(V_{g1} - V_{th})^2, \tag{15.75}$$

where W, L, and μ_n denote the channel width, length, and the electron surface mobility, respectively. Since the body effect is smaller in the FD SOI MOSFET than in the bulk MOSFET, a higher drain current can be obtained in this case.

The output characteristics of a PD SOI MOSFET show the kink effect due to impact ionization taking place near the drain region. This kink effect can be eliminated using contacts to the floating body of the MOSFET device underneath the gate.

In an FD SOI MOSFET, the transconductance g_m ($=\mathrm{d}I_{D\,sat}/\mathrm{d}V_g$) can be derived from (15.75), which yields

$$g_m = \frac{W\mu_n C_{ox1}}{nL}(V_{g1} - V_{th}). \tag{15.76}$$

As in the case of the drain current, a higher transconductance can be obtained in an FD SOI MOSFET than in a bulk MOSFET owing to the smaller body effect of SOI devices.

The maximum voltage gain of a bulk MOSFET is obtained when the value of g_m/I_D is maximized, and the voltage gain of a MOSFET is given by

$$\frac{\Delta V_{out}}{\Delta V_{in}} = \frac{g_m}{g_d} = \frac{g_m}{I_D}V_A, \tag{15.77}$$

where g_d is the output drain conductance and V_A is the Early voltage. It is noted that V_A for an FD SOI MOSFET is identical to that of the bulk MOSFETs discussed above. The maximum value of g_m/I_D occurs in the weak inversion regime for the MOSFET devices, and can be expressed by

$$\frac{g_m}{I_D} = \frac{\mathrm{d}I_D}{I_D\mathrm{d}V_g} = \frac{\ln(10)}{S} = \frac{q}{nk_BT}, \tag{15.78}$$

where S is the subthreshold slope and n is the body factor. In strong inversion g_m/I_D becomes

$$\frac{g_m}{I_D} = \sqrt{\frac{2\mu C_{ox}W/L}{nI_D}}. \tag{15.79}$$

Since the value of n (body factor) in an FD SOI MOSFET is lower than that in the bulk devices, values of g_m/I_D are significantly higher in FD SOI devices than in bulk devices. A typical value of g_m/I_D is 20–25 V^{-1} for bulk MOSFETs and 30–35 V^{-1} for FD SOI MOSFETs.

(iv) The subthreshold slope. The inverse subthreshold slope is defined as the inverse of the slope of the I_D versus V_g curve in the subthreshold regime, presented in a semilogarithmic plot:

$$S = \frac{dV_g}{d(\log I_D)} \approx \frac{k_B T}{q} \ln(10) \left(1 + \frac{C_D}{C_{ox}}\right) = n \left(\frac{k_B T}{q}\right) \ln(10). \quad (15.80)$$

Equation (15.80) is valid for both the bulk and FD SOI MOSFETs provided that the interface traps at the Si–SiO_2 interface are negligible. Note that the value of S given by (15.80) is $S \approx n \times 60$ mV/decade at room temperature. Since the body factor "n" for an FD SOI MOSFET is smaller than that of bulk devices, one can expect a smaller subthreshold swing for the FD SOI MOSFET than that of bulk devices.

15.6. Charge-Coupled Devices

As shown in Figure 15.19 a charge-coupled device (CCD) is referred to as an array of closely spaced (<2.5μm) MOS capacitors formed on a continuous oxide layer (100–200 nm thick) grown on a semiconductor substrate. An input gate and an output gate are added to both sides of the MOS capacitor array for the purpose of injecting and detecting the signal charges in the CCD array. The MOS capacitors are built closely enough to one another that minority carriers stored in the inversion layer associated with one MOS capacitor can be transferred to the surface channel region of an adjacent capacitor. The operation of a CCD is based on the storage and transfer of minority carriers (known as the charge packet) between the potential wells created by the voltage pulses applied to the gate electrode of the MOS capacitors. When a controlled sequence of clock voltage pulses is applied to the CCD array, the MOS capacitor is biased into deep depletion, and the charge packet from input signals can be stored and transferred from one potential well to another in a controlled manner across the semiconductor substrate. The basic types of CCDs include the surface-channel CCD (SCCD) and buried-channel CCD (BCCD). In an SCCD the charge packet is transferred along the surface channel, while in a BCCD doping of the semiconductor substrate is modified so that storage and transfer of the charge packet can take place in the buried channel of the semiconductor substrate.

In this section, the basic structure, operation principles, and characteristics of an SCCD are presented. The operation of an SCCD requires that the MOS capacitors in the CCD array be biased into deep depletion. Since the storage and transfer of charge packets in a CCD are achieved mainly by controlling the sequence of voltage pulses on the closely spaced MOS capacitors, it is important to understand

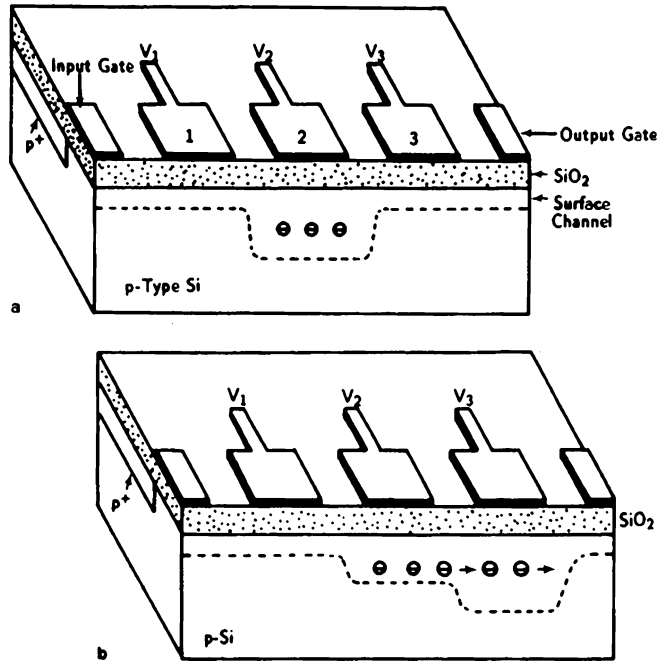


FIGURE 15.19. Schematic diagram of a three-phase CCD with input and output gates for signal charge injection and detection. Operating under (a) charge storage mode ($v_1 = v_3 > v_2$) and (b) charge transfer mode ($v_3 > v_2 > v_1$).

the transient behavior of an MOS capacitor operating in deep depletion under pulsed bias conditions. In a practical CCD, there are several different types of electrode configurations and clocking techniques that can be used to control its operation. Depending on the electrical performance, fabrication difficulty, and cell size, two-, three-, and four-phase CCDs have been made for digital and analog circuit applications. For example, a two-phase CCD has two MOS gates per cell, while a three-phase CCD has three MOS gates per cell.

15.6.1. Charge Storage and Transfer

Figures 15.19a and b show schematic drawings of a three-phase CCD operating in storage and transfer modes, respectively. The basic storage and transfer mechanisms for a CCD can be explained as follows. Figure 15.19a presents the storage mode for a three- phase SCCD. By applying gate voltage pulses with $v_2 > v_1 = v_3$, the charge packet is stored under the middle gate, which has a channel beneath it. If the applied gate pulses are such that $v_3 > v_2 > v_1$, then the right-hand gate causes the transfer of the charge packet from the channel region

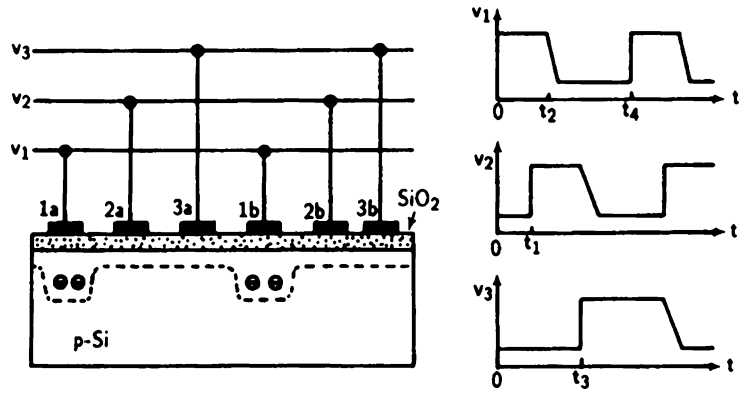


FIGURE 15.20. A three-phase SCCD along with its phase timing diagrams; the potential wells are shown for $t = 0$.

of the middle gate to the right. If one reduces the voltage pulse on the middle gate to v_1, then the bias voltage on the right-hand gate will be reduced to v_2. The net result is a shift of the charge packet one stage to the right. Figure 15.20 shows a three-phase SCCD with potential wells and their phase timing diagrams at $t = 0$.

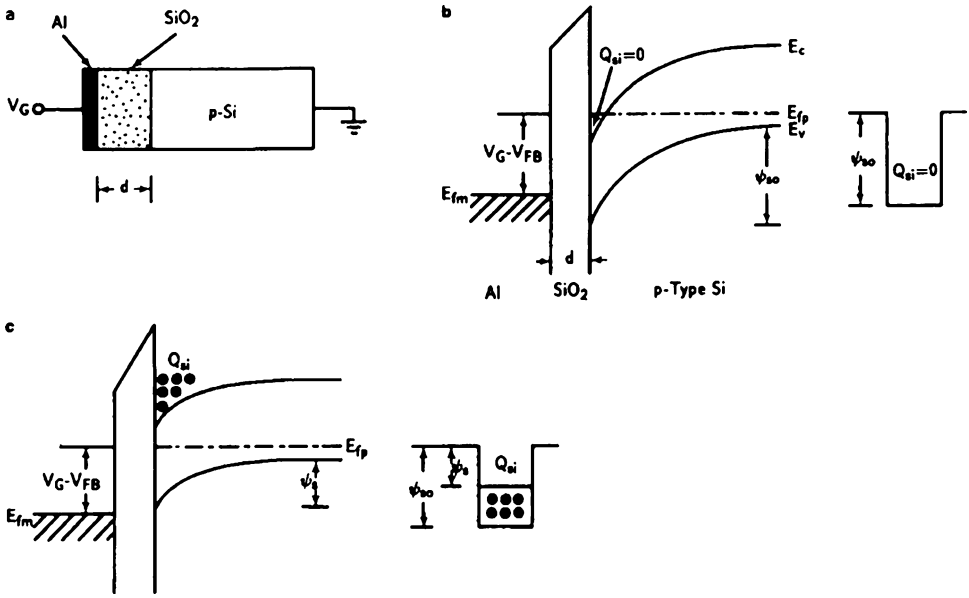


FIGURE 15.21. (a) Schematic drawing of an SCCD using an Al–SiO_2–p-type Si MOS capacitor structure. (b) Band bending into the deep depletion and the empty potential well under a large gate bias voltage. (c) Band bending at the Si–SiO_2 interface and the potential well partially filled with signal charges. After Barbe,[5] by permission, ©IEEE–1975.

Figure 15.21a shows a schematic drawing of a surface-channel MOS capacitor built on a p-type silicon substrate. If a large positive pulsed bias voltage is applied to the gate, then the condition of deep depletion is established under the gate. It is noted that no inversion layer is formed initially since no minority carriers are available in the depletion layer. Under deep-depletion conditions, a major portion of the applied voltage is across the depletion layer, and hence the surface potential φ_s is large. Figure 15.21b shows the energy band diagram for the MOS capacitor in deep depletion under zero signal charge conditions (i.e., $Q_{si} = 0$). In this case, the potential well is empty with a height equal to φ_{so}. The surface potential as a function of the gate voltage V_G for the MOS capacitor under deep depletion is given by

$$V_G - V_{FB} = -\frac{Q_s}{C_{ox}} + \varphi_s, \tag{15.81}$$

where V_G is the applied gate voltage and V_{FB} is the flat-band voltage shift. The total surface charge Q_s is equal to the sum of the depletion layer charge and the signal charge Q_{si}. This can be written as

$$Q_s = -qN_A W_d - Q_{si}, \tag{15.82}$$

where

$$W_d = \sqrt{2\varepsilon_s \varepsilon_0 \varphi_s / qN_A} \tag{15.83}$$

is the depletion layer width. Now substituting (15.82) and (15.83) into (15.81), one obtains

$$V_G - V_{FB} - \frac{Q_{si}}{C_{ox}} = \frac{\sqrt{2q\varepsilon_s \varepsilon_0 N_A \varphi_s}}{C_{ox}} \times \varphi_s. \tag{15.84}$$

The solution of the above equation for φ_s yields

$$\varphi_s = V'_G - B\left[\left(1 + \frac{2V'_G}{B}\right)^{\frac{1}{2}} - 1\right], \tag{15.85}$$

where

$$V'_G = V_G - V_{FB} - \frac{Q_{si}}{C_{ox}} \tag{15.86}$$

and

$$B = \frac{q\varepsilon_s \varepsilon_0 N_A}{C_{ox}^2}, \tag{15.87}$$

which shows that the surface potential is a function of the stored signal charge, the gate voltage, the substrate dopant density, and the oxide thickness. For a given gate voltage, φ_s decreases linearly with increasing stored signal charge. The linear relationship between φ_s and Q_{si} provides a simple explanation of the charge storage mechanism in the potential well of a CCD. The magnitude of φ_s specifies the depth of the potential well (W_d) as defined by (15.83). As shown in Figure 15.21c, filling

the potential well with signal charges will result in a linear reduction of the surface potential. Equation (15.85) is very important for CCD design, because the gradient of φ_s controls the movement of minority carriers in the potential well.

There are three basic charge transfer mechanisms in a CCD: thermal diffusion, self-induced drift, and the fringing field effect. When the signal charge packet is small, the predominant charge transfer mechanism is usually due to thermal diffusion. In this case, the total charge under the storage electrode decreases exponentially with time, and the decay time constant is given by

$$\tau_{th} = \frac{4L^2}{\pi^2 D_n}, \tag{15.88}$$

where D_n is the minority carrier diffusion constant and L is the length of the gate electrode. For example, if one assumes that $D_n = 10$ cm^2/s and $L = 10$ μm, then the decay time constant τ_{th} is 4 μs.

When the signal charge packet is very large (typically $> 10^{10}$ cm^{-2}), the charge transfer is dominated by the self-induced drift produced by electrostatic repulsion of the minority carriers in the potential well. In most cases the transfer of the first 99% of the charge is due to this mechanism. In some CCD operations, to improve transfer efficiency, the entire channel is filled with a large background charge known as fat zero. Self-induced drift is most important under this operation mode. The magnitude of the self-induced longitudinal electric field can be estimated by taking the gradient of the surface potential given by (15.85), which governs the transfer of charge carriers in a CCD. The decay of the initial charge packet Q_i because of self-induced drift can be calculated from

$$Q(t) = \left(\frac{t_0}{t + t_0}\right) Q_i \tag{15.89}$$

and

$$t_0 = \frac{\pi L^3 W C_{ox}}{2\mu_n Q}, \tag{15.90}$$

where L and W denote the gate electrode length and width, respectively, and μ_n is the electron mobility.

The surface potential under the storage electrode is influenced by the voltage applied to the adjacent electrodes because of the 2-D coupling of the electrostatic potential. The applied gate voltage results in a surface fringing field, which is present even when the signal charge is zero. The charge transfer process can be speeded up by using a fringing field established between the gate electrodes. This fringing field has a maximum at the boundaries between adjacent electrodes and minima at the centers of transfer gate electrodes. The magnitude of this fringing field increases with gate voltage and oxide thickness, and decreases with gate length and substrate doping density. For a surface channel CCD, with clock frequencies of several tens of megabits per second, charge transfer efficiencies greater than

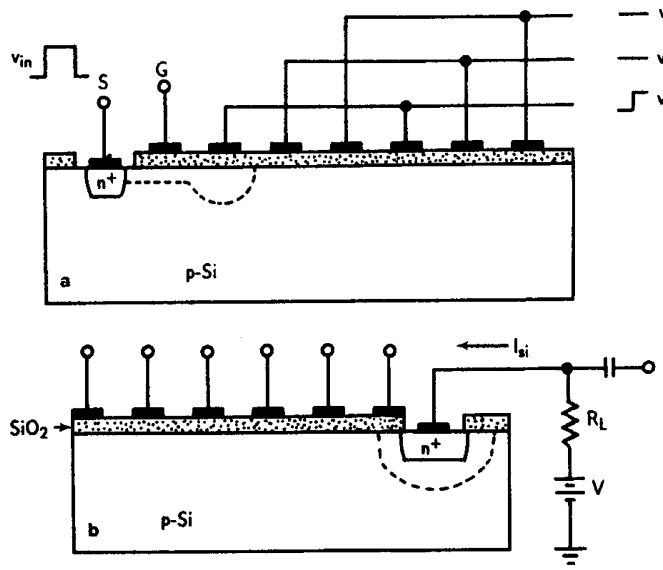


FIGURE 15.22. (a) Injection of minority carrier charges into the potential well of a CCD using a p-n junction diode. The n-type source is short-circuited to the substrate. (b) Detection of signal charge by the current-sensing technique.

99.99% (or transfer inefficiencies of less than 10^{-4}) can be obtained in the presence of a fringing field.

15.6.2. Charge Injection and Detection

In this section, charge injection and detection methods in a CCD are described. Charge packets can be injected by electrical or optical means. If a CCD is used as an image sensor, then the injection is carried out by optical means. On the other hand, in shift register or delay line applications, electrical injection is used instead. Figure 15.22a shows a p-n junction diode used to inject minority carriers into the potential well of a CCD. The n-type source is short-circuited to the substrate. When a positive pulse is applied to the input gate (V_{in}), the electrons injected from the source will flow into the potential well under the ϕ_1 gate electrode, and the current source keeps filling the potential well for the duration Δt of the input signal. Efficient injection can be achieved by biasing the source and input gates. Charge injection by optical means can be achieved by impinging light from the back side of the p-substrate. It should also be noted that optically generated minority carriers are attracted by the gate electrodes and accumulated in the potential wells.

The detection of signal charge packets in a CCD can be achieved by either current-sensing or charge-sensing methods. Figure 15.22b presents the current-sensing method. In this technique, a reverse-bias p-n junction diode is used as a

drain electrode at the end of the CCD array to collect the signal charge packet. When the signal charge packet reaches the drain junction, a current spike is detected at the output gate as a capacitive charging current. The charge-sensing detection method employs a floating diffusion region to periodically reset the voltage to a reference potential V_D. When the signal charge packet arrives in this region, the voltage there becomes a function of the signal charge, and the change of voltage at the floating diffusion region is detected using a MOSFET amplifier.

15.6.3. Buried-Channel CCDs

Section 15.6.2 presents the operation principles and general properties of SCCDs, in which charge storage and charge transfer occur in potential wells created in the surface inversion layer. In an SCCD, the interface traps play an important role in controlling charge transfer loss (i.e., transfer inefficiency) and noise in the CCD, particularly when the signal charge is small. To overcome problems associated with interface traps, it is common practice to move the channel away

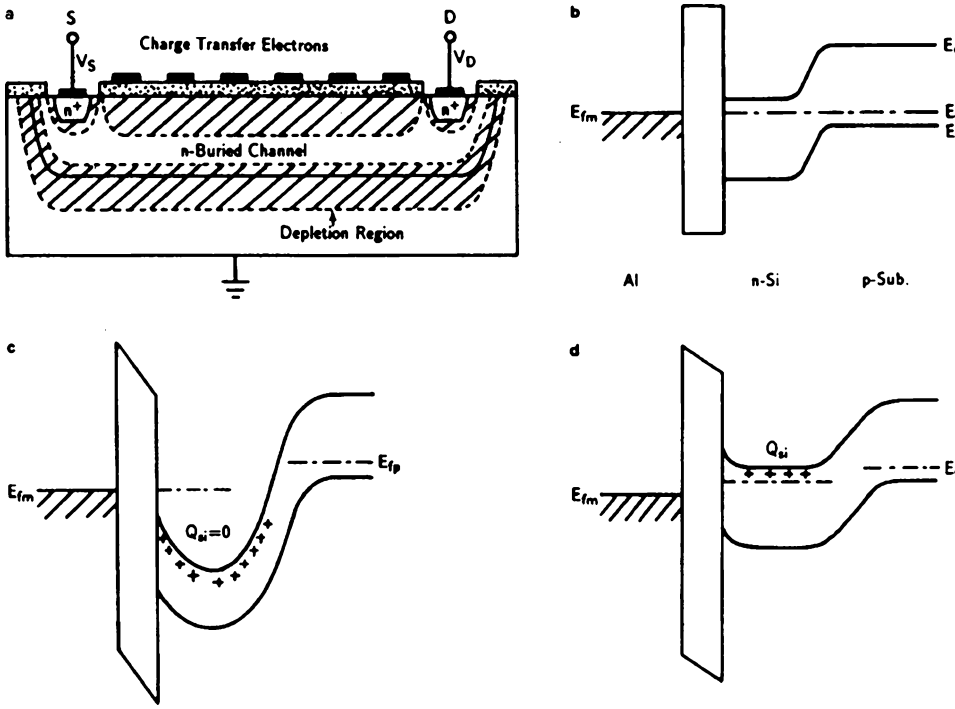


FIGURE 15.23. (a) Cross-sectional view of a buried-channel CD (BCCD). (b) Energy band diagram of a BCCD under equilibrium, (c) under reverse bias with an empty potential well and zero signal charge (i.e., $Q_{sig} = 0$), and (d) under reverse bias and in the presence of a signal charge packet (i.e., $Q_{sig} \neq 0$).

from the Si–SiO_2 interface. This results in a bulk- or buried-channel CCD (BCCD) as shown in Figure 15.23a. The buried channel is created using a thin n-type epitaxial or diffused layer on a p-type substrate. Figure 15.23b shows the energy band diagrams of a BCCD under equilibrium ($V_S = V_D = 0$), Figure 15.23c under reverse-bias ($V_S = V_D = V$) and zero signal charge conditions, and Figure 15.23d under reverse-bias conditions with signal charge. It is shown in Figure 15.23c that when a large positive bias voltage (reverse-bias) is applied to the source and drain gates, the majority carriers in the channel become completely depleted, and a potential well is formed in the n-buried channel. When a signal charge packet is injected from the source into the channel, a flat region in the mid-portion of the potential well results, as shown in Figure 15.23d. This potential well can store and transfer signal charge packets by using the controlled clock pulses applied to the MOS gate electrodes. This action is similar to that of the SCCD discussed earlier. The advantages of a BCCD over an SCCD include the elimination of problems associated with interface traps and the improvement of carrier mobility in the channel. Thus, transfer efficiency and channel mobility in a BCCD are expected to be higher than those of an SCCD. However, the drawbacks of a BCCD are additional process complications and a small capacitance, which reduces the signal-handling capability.

When CCDs are used in digital applications, ones and zeros are usually represented by the presence or absence of a charge packet in the channel under closely spaced MOS gates. The signals are clocked through the CCD at a rate set by the timing of the gate voltages. In analog applications, CCDs are frequently used as image sensors. For this application, the whole CCD is biased into deep depletion and exposed to a focused image for a time interval. The generation of charge carriers under the CCD gate is enhanced during the exposure time according to the brightness of the image. As a result, each channel will become charged to a level that represents the brightness at its location. This analog information can then be sent to the output gate and amplifier built on the edges of the CCD image array. This scheme has been demonstrated successfully in a portable TV camera. Finally, analog signals can also be stored in CCD arrays, since the amount of charge packet in the channel can be varied continuously. Therefore, CCDs are also used in analog delay line applications.

Recently, CMOS image sensors have gained popularity over CCD imagers because of the maturity of the fabrication technology of CMOS imagers. As a result, CMOS image sensors are now competing favorably against CCD imagers in most categories. Although the CMOS imager has not reached the imaging quality of CCDs, CMOS imager sensors are being used in a wide variety of applications owing to the advantages of CMOS imagers, which include (i) low-voltage operation and low power consumption, (ii) compatibility with integrating on-chip electronics, (iii) random access of image data, and (iv) lower cost as compared to CCDs. In spite of these advantages there are still problems in CMOS imagers such as sensitivity and operation under low-power conditions. However, because of the continued improvement in CMOS technology, it is anticipated that the

performance of CMOS imagers will eventually catch up with that of their CCD counterparts.

Problems

15.1. Draw the energy band diagrams and charge distributions for an Al–SiO_2 n-type Si MOS capacitor under (a) accumulation, (b) depletion, and (c) inversion conditions. Assume that the interface traps and work function difference are negligible.

15.2. Consider an Al-gate MOSFET fabricated on an n-type silicon substrate with dopant density $N_d = 2 \times 10^{15}$ cm^{-3}. If the thickness of the gate oxide is 100 nm and the interface trap density at the Si–SiO_2 interface is 2×10^{11} cm^{-2}, calculate:
   (a) The work function difference between aluminum and silicon (the modified work function $\phi'_m = 3.2$ eV for Al, and the modified electron affinity $\chi'_s = 3.5$ eV for Si);
   (b) The threshold voltage V_{TH}.

15.3. Using (15.38), (15.39), (15.45), and (15.46), show that the general expression for I_D is given by (15.48).

15.4. Consider a p-channel Al-gate MOSFET with dimensions and physical parameters given by $x_0 = 100$ nm, $L = 10$ μm, $W = 4$ μm, $N_d = 10^{15}$ cm^{-3}, $Q_{ss} = 2 \times 10^{11}$ cm^{-2}, and $\mu_p = 250$ cm^2/V·s.
   (a) Calculate I_D for $V_G = -2$, -4, -6, and -8 V and plot I_D versus V_D for these four different gate voltages.
   (b) If $V_G - V_{TH} = 1$ V, calculate the oxide capacitance and cutoff frequency for this transistor.
   (c) What are the transconductances of this MOSFET for $V_G = -2$, -4, and -6 V?

15.5. Calculate the drain saturation current $I_{D\,sat}$ and drain saturation voltage for an n-channel MOSFET with $x_0 = 100$ nm, $W/L = 15$, $\mu_n = 1100$ cm^2/V· s $V_{TH} = 0.5$ V, and for $V_G = 3$ and 5 V.

15.6. Consider a MOSFET built on a p-type silicon substrate with $N_A = 1 \times 10^{15}$ cm^{-3} and an aluminum metal gate with $\phi_{ms} = -0.27$ V and an oxide thickness of 100 nm. Calculate:
   (a) φ_{si} and Q_B at the onset of strong inversion,
   (b) The threshold voltages V_{TH} and V'_{TH} using (15.37) and (15.39),
   (c) V''_{TH} using (15.40) for $V_{sb} = -1$ and -3 V.

15.7. The oxide of a p-channel MOSFET contains mobile sodium ions that can move slowly toward the Si–SiO_2 interface under the influence of an applied electric field. Discuss the effect of these mobile sodium ions on the characteristics of such a MOSFET if a positive gate voltage is applied to the device.

15.8. The mobile charge in the channel per unit area, $Q_I(y)$, in an n-channel MOSFET is given by (15.45), and the potential $V(y)$ in the channel is given by (15.57) [with V_D replaced by $V(y)$]. Show that the small-signal capacitance

measured between the gate and the source in the saturation region is given by

$$C_{GS} = \frac{dQ_I}{dV_g} = \frac{2}{3} WLC_{ox},$$

where Q_I is the total charge in the channel and $V(y) = V_G - V_{TH}$ for $y = L$.

15.9. The potential distribution for the BCCD shown in Figure 15.23 can be obtained analytically using the depletion approximation for the case in which the impurity concentrations are constant in the n and p regions of the BCCD. Poisson equations for the potential are given by

$$\frac{d^2\varphi}{dx^2} = \begin{cases} 0, & -d_{ox} < x < 0, \\ \dfrac{-qN_d}{\varepsilon_0\varepsilon_s}, & 0 < x < W_n, \\ \dfrac{qN_A}{\varepsilon_0\varepsilon_s}, & W_n < x < W_n + W_p. \end{cases}$$

The boundary conditions are given by

(1) $\varphi = (V_G - V_{FB})$ at $x = -d_{ox}$,

(2) $\varphi = 0$ at $x = W_n + W_p$,

The potential and electric displacement are continuous at $x = 0$ and $x = W_n$. Show that the maximum potential well displayed in Figure 15.23b is given by

$$\varphi_{max} = \left(\frac{qN_A}{2\varepsilon_s\varepsilon_0}\right) x_p^2 \left(1 + \frac{N_A}{N_D}\right).$$

15.10. Using (15.67) show that the threshold voltage for the cases in which the back surface is (a) in accumulation and (b) inverted can be expressed by (15.68) and (15.69), respectively.

References

1. C. G. Garrett and W. H. Brattain, "Physical Theory of Semiconductor Surfaces," *Phys. Rev.*, **99**, 376 (1955).
2. D. K. Schroder, *Semiconductor Material and Device Characterization*, Chap. 6, Wiley Interscience, New York (1990).
3. M. H. White and J. R. Cricci, "Characterization of Thin-Oxide MNOS Memory Transistors," *IEEE Trans. Electron Devices* **ED-19**, 1280 (1972).
4. D. F. Barbe, "Imaging Devices Using the Charge-Coupled Concept," *Proc. IEEE* **63**, 38 (1975).
5. J. Brews, "Physics of the MOS Transistor," in *Silicon Integrated Circuits*, Part A (D. Kahng, ed.), Academic Press, New York (1981).

Bibliography

G. F. Amelio, W. J. Bertram, Jr., and M. F. Tompsett, "Charge-coupled Image Devices: Design Considerations," *IEEE Trans. Electron Devices* **ED-18**, 986 (1971).

A. Bar-Lev, *Semiconductors and Electronic Devices*, 2nd ed., Prentice-Hall, Englewood Cliffs (1984).

W. J. Betram, M. Mohsen, F. J. Morris, D. A. Sealer, C. H. Sequin, and M. F. Tompsett, "A Three-Level Metallization Three-Phase CCD," *IEEE Trans. Electron Devices* **ED-21**, 758 (1974).

W. S. Boyle and G. E. Smith, "Charge Couple Semiconductor Devices," *Bell Syst. Tech. J.*, **49**, 587 (1970).

C. Y. Chang and S. M. Sze, *ULSI Devices*, Wiley Interscience, New York (2000).

J. Y. Chen, "CMOS—The Emerging VLSI Technology," *IEEE Circuits & Device Magazine* **2**, 16 (1986).

J. P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI,* Kluwer, Boston, 1991.

S. Cristoloveanu and S. S. Li, *Electrical Characterization of Silicon-on-Insulator Materials and Devices,* Kluwer, Boston, 1995.

P. E. Gray and C. L. Searle, *Electronic Principles: Physics, Models and Circuits*, Wiley, New York (1969),

A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York (1967).

D. Kahng, *Historical Perspective on the Development of MOS Transistors and Related Devices*, *IEEE Trans. Electron Devices,* **ED-23**, 65 (1976).

C. K. Kim and M. Lenzlinger, "Charge Transfer in Charge Coupled Devices," *J. Appl. Phys.*, **42**, 3856 (1971).

R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed., Wiley, New York (1986).

D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*, 3rd edition, McGraw-Hill, New York, 2003.

E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.

L. C. Parrilo, "VLSI Process Integration," in: *VLSI Technology* (S. M. Sze, ed.), McGraw-Hill, New York (1983).

C. H. Sequin and M. F. Tomsett, *Charge Transfer Devices*, Chap. 5, Academic Press, New York (1975).

S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1981).

S. M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd edition, Wiley Interscience, New York (2002).

Y. Tau and T. K. Ning, *Physics of Modern VLSI Devices,* Cambridge University Press, London, 1998.

A. Theuwissen, *Solid State Imaging with Charge-Coupled Devices*, Kluwer Academic Publishing, Boston (1995).

M. F. Tomsett, "A Simple Regenerator for Use with Charge-Transfer Devices and the Design of Functional Logic-Arrays," *IEEE J. Solid-State Circuits* **SC-7**, 237 (1972).

R. R. Troutman, *Latchup in CMOS Technology*, Kluwer, Boston (1986).

Y. Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York (1987).

H. S. Wang, "CMOS Image Sensor-Recent Advances and Device Scaling Considerations," *IEEE IEDM Tech.*, pp. 201–204 (1997).

H-S. P. Wong, "MOSFET Fundamentals," in *ULSI Devices*, edited by C. Y. Chang and S. M. Sze, Wiley Interscience, New York, 2000.

C. Xu, W. H. Ki, and M. Chan, "A Low-Voltage Complementary Active Pixel Sensor (CAPS) Fabricated Using a 0.25 μm CMOS Technology," *IEEE Electron Device Letters*, **23** (7), 398 (2002).

E. S. Yang, *Microelectronic Devices,* McGraw-Hill, New York (1988).

D. N. Yang, S. G. Wu, "Nonsilicide Source/Drain Pixel for 0.25 μm CMOS Image Sensor," *IEEE Electron Device Letters*, **22** (2), 71 (2001).

16
High-Speed III-V Semiconductor Devices

16.1. Introduction

In this chapter the basic device physics, operational principles, and general characteristics of high-speed III-V compound semiconductor devices such as MESFETs and HEMTs are presented. The devices described here include GaAs- and InP-based metal–semiconductor field-effect transistors (MESFETs) and high electron mobility transistors (HEMTs). The GaAs-based high-speed devices are fabricated using the lattice-matched GaAs/AlGaAs material system grown on a semi-insulating GaAs substrate, while the InP-based devices utilize the lattice-matched InAlAs/InGaAs or InGaAs/InP material systems grown on a semi-insulating InP substrate. Although the GaAs/AlGaAs material technology is more mature than that of the InP/InGaAs material system, the InP-based devices can be operated at a much higher frequency and higher speed than those of the GaAs-based devices. This is due to the fact that the InGaAs/InP material system has a higher electron mobility and smaller electron effective mass than those of the AlGaAs/GaAs material system.

III-V compound semiconductors such as GaAs, InP, and InGaAs have been developed for high-speed device applications. The III-V semiconductor materials generally have higher electron mobilities and higher peak saturation velocities than silicon, as well as the availability of semi-insulating substrates. With the advent of MBE and MOCVD growth techniques, high-quality AlGaAs/GaAs, InGaAs/AlGaAs, InGaP/GaAs, and InGaAs/InAlAs heterostructure epitaxial layers and quantum-well/superlattice structures can be readily grown on either semi-insulating GaAs or InP substrates for the fabrication of a wide variety of high-speed devices. FETs grown on semi-insulating substrates can greatly reduce the leakage current and parasitic capacitances, and hence enable the integration of both lumped and distributed microwave components on the same substrate.

The successful development of GaAs monolithic microwave integrated circuits (MMICs) is a good example of using GaAs in high-speed digital IC applications. In fact, impressive results have been achieved in both GaAs- and InP-based FETs

and HBTs. This is due to the availability of high-quality epitaxial layers prepared using MBE and MOCVD growth techniques as well as new device processing techniques. For example, high-current (600 mA/mm) and high-cutoff-frequency ($f_T > 80$ GHz) HEMTs using delta-doped heterostructures have been fabricated on AlGaAs/GaAs material systems. Thin-channel and highly doped (2×10^{18} cm^{-3}) refractory gate self-aligned GaAs MESFETs fabricated by rapid thermal annealing (RTA) have produced transconductance with values of $g_m > 550$ mS/mm. InGaAs/InP HEMT with f_{max} greater than 100 GHz has been demonstrated. In fact, InP-based lattice-matched and pseudomorphic HEMTs have emerged as leading candidates for ultra-low-noise and high-frequency applications. Transconductance of $g_m = 1{,}000$ mS/mm has been realized in 0.1 μm gate HEMT devices, and a p-PHEMT with f_{max} of 600 GHz and a noise figure of 1.4 db at 94 GHz was reported recently. These high-speed devices are critical components of advanced satellite communications, radio astronomy, and wideband instrumentation.

Section 16.2 presents the device physics and structures, I-V characteristics, small-signal device parameters, and some second-order effects in a GaAs MESFET. Section 16.3 describes the equilibrium properties of two-dimensional electron gas (2-DEG) in the triangular potential well of an AlGaAs/GaAs heterostructure. The basic device theory, the operational principles, and I-V characteristics of an HEMT device are also discussed. The hot-electron transistor (HET) and its physical limitations are presented in Section 16.5. In Section 16.6, the device physics, general characteristics, and performance limitations of a resonant tunneling device (RTD) are presented. Finally, the basic physical principles and general properties of a GaAs Gunn-effect device are discussed in Section 16.7.

16.2. Metal–Semiconductor Field-Effect Transistors

The maturity of GaAs metal–semiconductor field-effect transistor (MESFET) technology has created a major impact on solid-state microwave technologies. The GaAs MESFET has made great strides in both microwave amplifier and digital applications for commercial and military systems. GaAs MESFET amplifiers operating up to 40 GHz are now commercially available. It is the most important microwave device in existence today with one of the highest unity gain cutoff frequencies (f_T) available in a transistor. For example, a GaAs MESFET with a 0.25 μm gate length exhibits a cutoff frequency of 80 GHz at 300 K. In fact, GaAs MESFETs, which form the basis of microwave monolithic integrated circuit (MMIC) technologies, are now emerging from research to commercial production, and 16 K static random access memory (SRAM) and ring oscillators have been built for different digital circuit applications. In this section, the device structures, *I–V* characteristics, and small-signal device parameters, as well as some second-order effects of a GaAs MESFET, are described.

16.2.1. Basic Device Structure and Characteristics

A GaAs MESFET is a rather simple field-effect transistor (FET). It is a three-terminal majority carrier device, which consists of two ohmic contacts (i.e., source and drain) separated by a Schottky barrier gate contact electrode. A conducting channel is formed between the source and the drain electrodes, and a depletion region in the channel is formed under the gate electrode with its width controlled by the gate voltage. Therefore, a MESFET device can be considered a voltage-controlled resistor. Figures 16.1a and b show cross-sectional views of two different GaAs MESFET device structures that are formed (a) by the epitaxial layer growth and (b) by the ion-implantation techniques. The basic device structure for a GaAs MESFET consists of an n-type GaAs active layer of 0.2–0.3 μm with a dopant density of 2×10^{17} cm^{-3} deposited on a high-resistivity buffer layer, which is grown on top of an undoped or Cr-doped semi-insulating (i.e., $\rho \geq 10^7$ Ω cm), (100)-oriented GaAs substrate. An undoped high-resistivity buffer layer of 3–5 μm thickness is grown between the active layer and the substrate to prevent out-diffusion of residual impurities from the substrate into the active layer. Two ohmic contact regions, separated by 5 μm or less, form the source and drain electrodes, and a Schottky barrier contact between the source and drain regions (i.e., the channel region) forms the control-gate electrode. To reduce the source- and drain-contact resistance, a heavily doped n^+ GaAs contact layer typically 0.1 μm thick and with a doping density of 2×10^{18} cm^{-3} is grown on top of the active layer to reduce the parasitic resistance. However, as shown in Figure 16.1b, most GaAs MESFETs available today are fabricated by direct ion implantation onto a semi-insulating GaAs substrate. Selective implantation enables an active channel region of dopant density about 10^{17} cm^{-3} to be realized. Heavily doped source and drain regions can be implanted immediately adjacent to the channel in order to minimize the source resistance using the self-aligned process. In this case, implantation is restricted to those areas necessary for MESFET fabrication, leaving the rest of the wafer in its semi-insulating condition.

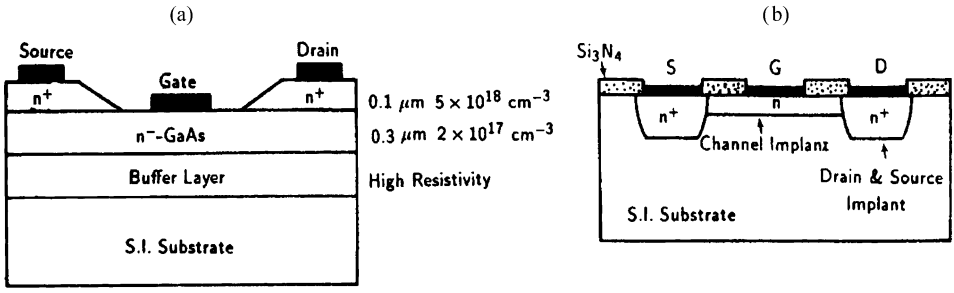


FIGURE 16.1. Cross-sectional views of a GaAs MESFET: (a) with a high-resistivity buffer layer, an n^--channel epilayer, and an n^+-cap layer; (b) with an n^+-source, drain, and n^--channel regions created by direct implantation on a semi-insulating GaAs substrate.

In general, a GaAs MESFET differs from a silicon JFET in several respects. The channel region consists of a very thin ($\approx 0.2\,\mu m$) n-type GaAs layer grown on a Cr-doped or undoped semi-insulating GaAs substrate. For an n-channel GaAs MESFET, the controlled gate is formed by a Schottky barrier metal contact (e.g., Ti/Pt/Au, or WSi_x by a self-aligned process) with gate lengths varying from a few microns to one tenth of a micron (0.1 μm). At a channel doping density of around 2×10^{17} cm^{-3}, the very thin channel under the gate is only partially depleted by the Schottky barrier built-in voltage (≈0.8 V). When a positive drain voltage V_{DS} is applied to the FET, the drain current I_{DS} will flow through the channel, producing current–voltage characteristics similar to those of a JFET. A negative gate voltage (V_{GS}) will further deplete the channel, reducing I_{DS} until the device cuts off when V_{GS} reaches the pinch-off voltage. On the other hand, a small positive gate voltage V_{GS} (<0.8 V) will turn the Schottky gate toward forward conduction, causing I_{DS} to increase dramatically. The carrier transit time τ through a channel of length L is approximately equal to L/v_s where v_s, is the saturation velocity of electrons in the channel. For a GaAs MESFET with $v_s = 2 \times 10^7$ cm/s and a channel length of 0.5 μm, the device unity current gain cutoff frequency f_T is about 70 GHz.

The GaAs MESFET structure shown in Figure 16.1a has been widely used as a low-noise, small-signal microwave amplifier. In this structure, the active layer is etched briefly before the Schottky gate contact is made. The recessed gate structure along with an n^+-ion implantation into the source and drain regions will reduce the source and drain parasitic series resistances and greatly improve the ohmic contacts in both regions. Reducing the parasitic resistances will also decrease the noise figure and increase the transconductance and cutoff frequency of the MESFET. For power amplification at microwave frequencies, high current is needed, which, in turn, requires a very wide gate. Therefore, a high-power MESFET usually consists of many small MESFETs connected in parallel to increase the total current and power output.

Although overlays are frequently used to reduce contact resistances and thicken up the bonding pads, the metallizations required to complete the basic device structure of a GaAs MESFET are the source and drain ohmic contacts and the Schottky gate contact. The gate length for a microwave FET usually varies between 0.25 and 1.5 μm. Such small dimensions can be achieved with a lift-off process using either the optical or electron beam lithographic technique. Typical Schottky barrier gate metallizations include using Al or a multilayer structure such as Ti/Pt/Au or WSi_4 contact. Au/Ge/Ni is commonly used for ohmic contacts on n-type GaAs, while Au/Zn is widely used for ohmic contacts on p-type GaAs.

Depending on the channel doping density and channel height, there are two types of MESFETs that are commonly used in GaAs digital IC applications. They are the normally on or depletion/mode (D-) MESFET and the normally off or enhancement mode (E-) MESFET. The D-MESFET requires a negative gate voltage (i.e., V_{GS}) to cutoff the drain current under all conditions, while the E-MESFET requires a positive gate voltage to open up the channel. By controlling the channel doping density N_D and channel height a, both E- and D-MESFETs can be fabricated on

a semi-insulating GaAs substrate. By solving the Poisson equation in the channel depletion region under the Schottky gate, it can be shown that a normally on MESFET can be obtained if

$$qN_{\mathrm{D}}a^2/2\varepsilon_0\varepsilon_{\mathrm{s}} > V_{\mathrm{bi}}, \tag{16.1}$$

and a normally off MESFET is obtained if

$$qN_{\mathrm{D}}a^2/2\varepsilon_0\varepsilon_{\mathrm{s}} < V_{\mathrm{bi}}, \tag{16.2}$$

where V_{bi} is the built-in potential of the MESFET, and other parameters have their usual meanings. It is seen that (16.1) and (16.2) are the basic conditions that may be used to guide the design of E/D MESFETs on a GaAs substrate.

The velocity–field relation and the general characteristics for a GaAs MESFET are discussed next. Figures 16.2a and b illustrate the velocity–field relation for an undoped GaAs at 300 K and 77 K, respectively, and show a negative differential mobility regime under steady-state conditions and at moderate field strength. This velocity–field relation may be represented by a two-piecewise linear approximation, as shown in curve 2, or a more accurate relation as shown in curve 3 of Figure 16.2a. The two-piecewise linear approximation is often used in the analysis of MESFET characteristics due to its simplicity. This is understandable if one notes that at low fields the electron drift velocity v_{d} is linearly related to the electric field with a proportionality constant equal to the low-field mobility μ_{n}, which satisfies

$$v_{\mathrm{d}} = \mu_{\mathrm{n}}\mathcal{E} \quad \text{for } \mathcal{E} < \mathcal{E}_{\mathrm{c}}, \tag{16.3}$$

and in the saturation regime,

$$v_{\mathrm{d}} = v_{\mathrm{s}} \quad \text{for } \mathcal{E} \geq \mathcal{E}_{\mathrm{c}}, \tag{16.4}$$

where $\mathcal{E}_{\mathrm{c}}$ is the critical field that divides the linear and saturation regions of drift velocity versus the electric field curve. A more precise equation describing the velocity–field relation is shown in curve 3 of Figure 16.2a, which is given by

$$v_{\mathrm{d}} = \frac{\mu_{\mathrm{n}}\mathcal{E}}{[1 + (\mu_{\mathrm{n}}\mathcal{E}/v_{\mathrm{d}})^2]^{1/2}}. \tag{16.5}$$

Since the transit time of electrons in the high-field regime is comparable to the time constant characterizing relaxation to steady-state velocity–field characteristics, the electron dynamics in a microwave FET are much more complex than the dynamics derived from the steady-state velocity–field characteristics. Figure 16.3a shows the device structure of a GaAs MESFET along with the origins of each circuit element, and Figure 16.3b presents its small-signal equivalent circuit. The dc device parameters are derived by examining the GaAs MESFET structure shown in Figure 16.3a, which has an n-type active layer (channel) of thickness a and doping density N_{D}, gate width Z, and gate length L. The saturation current I_{Dsat} that can be carried by the channel is given by

$$I_{\mathrm{sat}} = qN_{\mathrm{D}}Zav_{\mathrm{d}}, \tag{16.6}$$

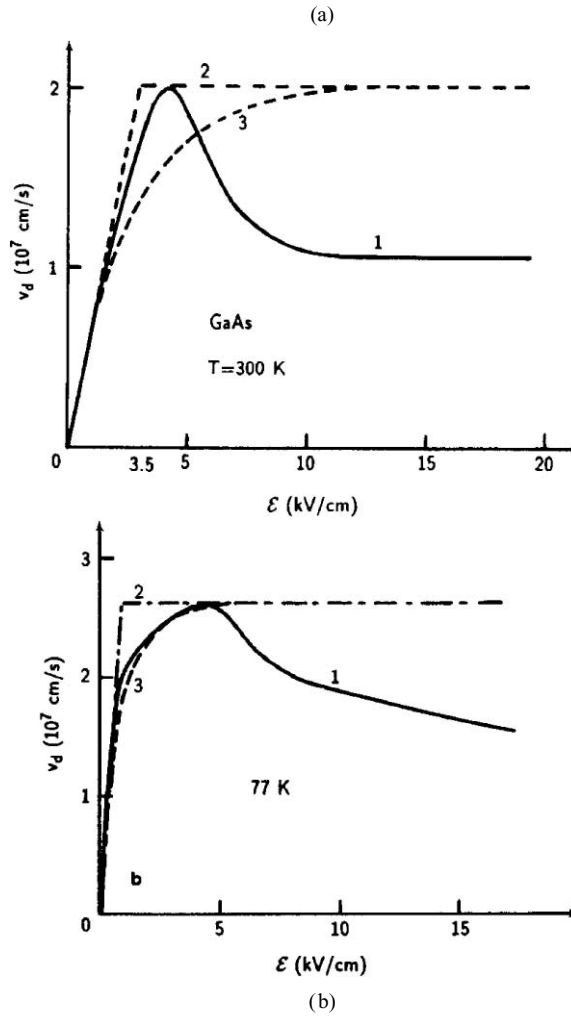


FIGURE 16.2. Drift velocity versus electric field (curve 1) in bulk GaAs at (a) 300 K and (b) 77 K, respectively. Curve 2 is calculated using a two-piece linear approximation, and curve 3 is calculated using (16.5).

where q is the electronic charge and v_s is the saturation velocity. If the metal gate forms a Schottky contact to this active layer and $V_{DS} = 0$, then using a one-sided abrupt junction approximation, the depletion layer width W_d of the Schottky gate to the active channel is given by

$$W_d = \sqrt{\frac{2\varepsilon_0\varepsilon_s(-V_{GS} + V_{bi})}{qN_D}}, \tag{16.7}$$

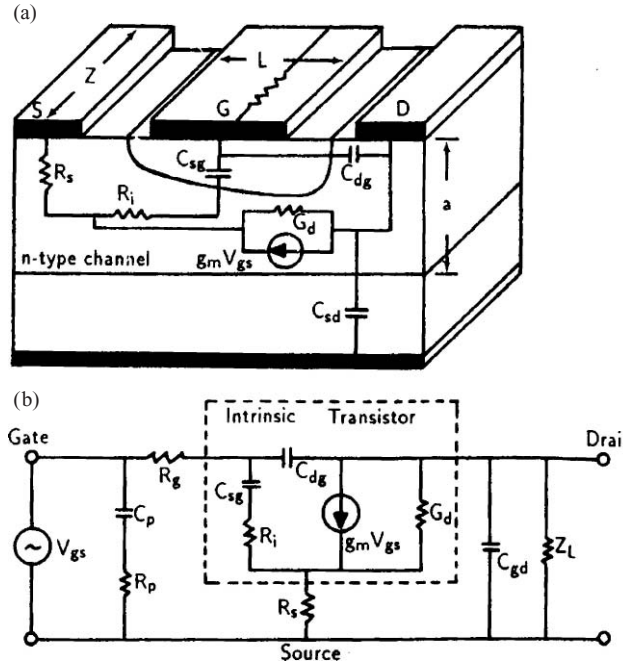


FIGURE 16.3. (a) Sketch of GaAs MESFET device structure along with the origins of each circuit element, shown in (b). (b) Small-signal equivalent circuit of a GaAs MESFET operating in the saturation region in a common-source configuration.

where V_{bi} is the built-in potential of the Schottky barrier contact gate and V_{GS} is the gate-to-source voltage. Thus, the channel height $b(x)[= a - W_d(x)]$ can be controlled by the gate voltage V_{GS}. When the channel is pinched off, the depletion layer depth extends from the surface ($y = 0$) completely through the n-type active layer ($y = a$), and V_{GS} equals the pinch-off (threshold) voltage V_{p0}. Equations (16.1) and (16.2) are the conditions for the pinch-off voltage required to deplete the channel at the drain side:

$$V_{p0} = \frac{qN_D a^2}{2\varepsilon_0 \varepsilon_s}. \tag{16.8}$$

It is seen that the relative importance of the roles of velocity saturation and pinch-off can be measured using the saturation index ($\mathcal{E}_s L / V_{p0}$), which is equal to the ratio of potential drop along the gate region at the saturation field to the pinch-off potential V_{p0} required to totally deplete the channel. A smaller saturation index represents a greater importance of the velocity saturation in limiting the source–drain current. We shall next derive the dc current–voltage relation and the small-signal device parameters for a GaAs MESFET.

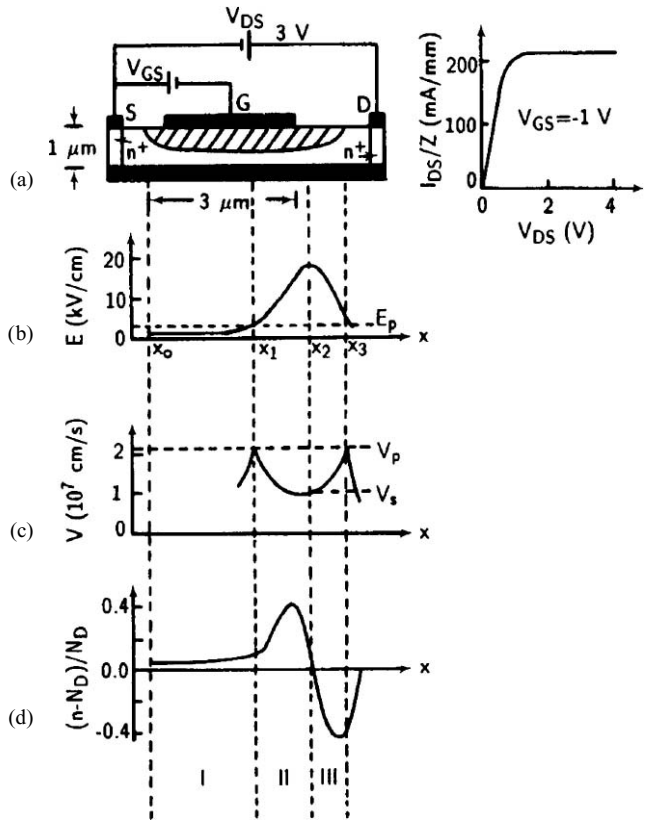


FIGURE 16.4. Distribution of electric field, drift velocity, and space charge in the channel of a GaAs MESFET. (a) Cross-section view showing the three regions of the channel: (I) the depletion region, (II) the electron accumulation region, and (III) formation of the Gunn domain. (b) Electric field versus distance in the channel. (c) Drift velocity versus distance in the channel. (d) Normalized space-charge density versus distance in the channel. Also shown in (a) is the normalized I_{DS}/Z versus V_{DS} plot. After Liechti,[1] by permission, © IEEE-1976.

16.2.2. Current–Voltage Characteristics

A MESFET operation can be described analytically if some simplified assumptions concerning the electric field distribution within the device, the carrier dynamics, and the channel profile are made. For example, if the drain voltage V_{DS} is below the knee voltage of the I_{DS} versus V_{DS} characteristic curves, then electrons travel with constant mobility in the channel. In this case, the current–voltage characteristics can be derived using the gradual channel approximation as in the case of a JFET (see Chapter 11). As shown in Figure 16.4a, the operation of a MESFET above the knee voltage may be considered by dividing the channel into three regions. In region I, carriers travel with constant mobility between the source end of the

gate and the velocity saturation point $x = x_1$, where the electric field rises to the value of the saturation field $\mathcal{E}_s$. In regions II and III, the carriers travel at their saturation velocity. These two regions meet at $x = L$, i.e., the drain end of the gate. Region III terminates when the electric field in the channel falls to the saturation field. The longitudinal electric field $\mathcal{E}_x$ as a function of position in the three regions of the channel is shown in Figure 16.4b. The electric field in regions I will increase initially as the carriers enter the channel beneath the gate. The electric field continues to increase until it reaches the saturation field at the boundary between region I and II. At the knee voltage, the velocity saturation point is located at the drain end of the gate. At higher drain–source potentials, the velocity saturation point moves toward the source end of the gate, and the electric field in regions II and III increases to accommodate the drain–source potential. In region II, carriers travel at their saturation velocity v_s. To satisfy Gauss's law and current continuity in the channel, the increase in $\mathcal{E}_x$ must be accompanied by a reduction in channel height and by carrier accumulation beyond the charge-neutral value of region I. Therefore, the electric field in the channel should reach its peak value near the drain end of the gate. In region III, the carrier density falls as the channel height increases beyond the drain end of the gate, and as the channel height rises above that at the velocity saturation point, the mobile carrier density drops below the charge-neutral value. The transition from carrier accumulation to carrier depletion is achieved by a sharp drop of the electric field in the channel region. Charge neutrality is restored when the electric field falls below the saturation field and electrons reenter the constant-mobility regime. The strong saturation in the drain–source current above the knee voltage is a direct result of the fact that most of the V_{DS} voltage drop is across the accumulation–depletion regions of the channel. Increasing V_{DS} will emphasize these regions and also increases the length of the velocity saturation region in the channel. Figure 16.4c shows the drift velocity versus the electric field in the three regions, where v_p is the peak velocity. Figure 16.4d illustrates the space-charge distribution in the channel.

As mentioned above, the current–voltage characteristics of a MESFET below the knee voltage, in which the electric field is low and the mobility is constant, can be analyzed using the gradual channel approximation. Figure 16.5 shows the symmetrical structure for the gradual channel analysis of a GaAs MESFET. It is shown in this figure that the channel current I_{ch} per half device is related to the channel potential $V(x)$ by

$$I_{ch} = G_0 L \left(1 - \frac{W_d(x)}{a}\right) \frac{dV(x)}{dx}, \tag{16.9}$$

where G_0 is the conductance of a fully opened channel, which can be expressed by

$$G_0 = \frac{qN_D\mu_n Za}{L}, \tag{16.10}$$

where Z, a, and L are the device dimensions defined in Figures 16.3a and 16.5, respectively; $W_d(x)$ is the depletion layer width in the channel under the gate,

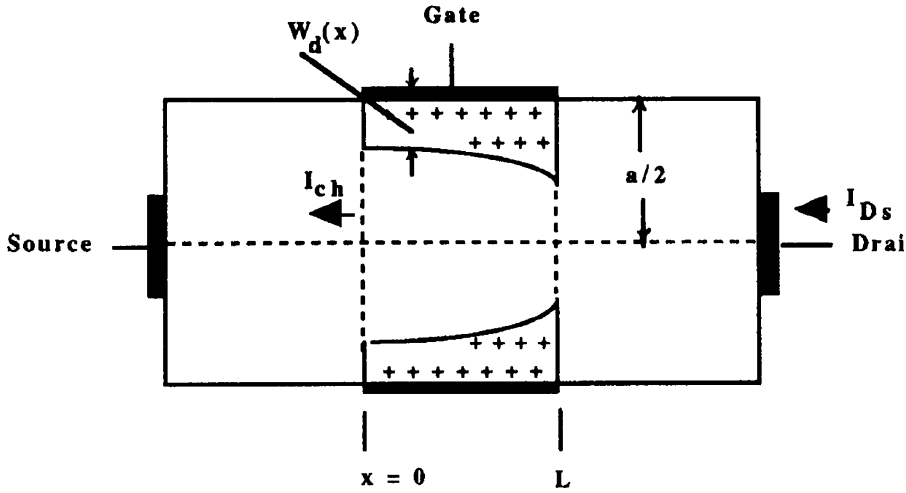


FIGURE 16.5. The symmetrical device structure used in the gradual channel approximation analysis of a GaAs MESFET.

while μ_n and N_D denote the low-field mobility and dopant density, respectively. A general expression for the depletion layer width $W_d(x)$ as a function of distance in the channel can be obtained by including $V(x)$ in (16.7), which yields

$$W_d(x) = \sqrt{\frac{2\varepsilon_0\varepsilon_s(V(x) - V_{GS} + V_{bi})}{qN_D}}. \tag{16.11}$$

If the source and drain series resistances are neglected in the analysis, then the channel current can be obtained by substituting (16.11) into (16.9), and integrating x from 0 to L and $V(x)$ from 0 to V_i, which yields

$$I_{ch} = G_0 \left\{ V_i - \frac{2}{3} \frac{[(V_i - V_{GS} + V_{bi})^{3/2} + (-V_{GS} + V_{bi})^{3/2}]}{V_{p0}^{1/2}} \right\}, \tag{16.12}$$

where I_{ch} is the channel current, and V_i is the voltage drop across the gate region. If one neglects the series resistance of the gate-to-source and gate-to-drain and contact resistance, then $V_i = V_{DS}$. Here V_{bi} is the built-in voltage, V_{GS} is the gate voltage, G_0 is defined by (16.10), and V_{p0} is the pinch-off voltage given by (16.8). It is noted that (16.12) is valid when $V_i \ll V_s = \mathcal{E}_s L$, where $\mathcal{E}_s = v_s/\mu$ is the average electric field under the gate to reach a sustaining domain in the drain region. For a typical GaAs MESFET with $L \approx 1$ μm and $V_s \ll (V_{bi} - V_{GS})$, the channel current varies almost linearly with channel voltage up to the saturation point, and is given by

$$I_{ch} \approx G_0 \left[1 - \left(\frac{V_{bi} - V_{GS}}{V_{p0}} \right)^{1/2} \right] V_{ch} = G_d V_{ch}, \tag{16.13}$$

where

$$G_d \approx G_0 \left[1 - \left(\frac{V_{bi} - V_{GS}}{V_{p0}} \right)^{1/2} \right] \tag{16.14}$$

is the drain conductance. From (16.13), the channel saturation current I_{sat} can be written as

$$I_{sat} = G_d V_s. \tag{16.15}$$

For a GaAs MESFET with low pinch-off voltage (e.g., $V_{p0} \leq 2V$), the current–voltage relation operating in the saturation region can be described accurately using the "square law" model, which is given by

$$I_{Dsat} = K(V_{GS} - V_T)^2, \tag{16.16}$$

where I_{Dsat} denotes the drain–source saturation current, $V_T (= k_B T/q)$ is the threshold voltage, and

$$K = \frac{2\varepsilon_0 \varepsilon_s \mu_n v_s Z}{a(\mu V_{p0} + 3 v_s L)}. \tag{16.17}$$

The source series resistance effects are neglected in (16.16).

Figure 16.6 shows plots of I_{DS} versus V_{DS} for (a) low pinch-off voltage GaAs MESFET ($V_{p0} = 1.8$ V, $N_D = 1.81 \times 10^{17}$ cm^{-3}, $L = 1.3$ μm, and $Z = 20$ μm), and (b) high pinch-off voltage GaAs MESFET ($V_{p0} = 5.3$ V, $N_D = 6.5 \times 10^{16}$ cm^{-3}, $L = 1.0$ μm, and $Z = 500$ μm).

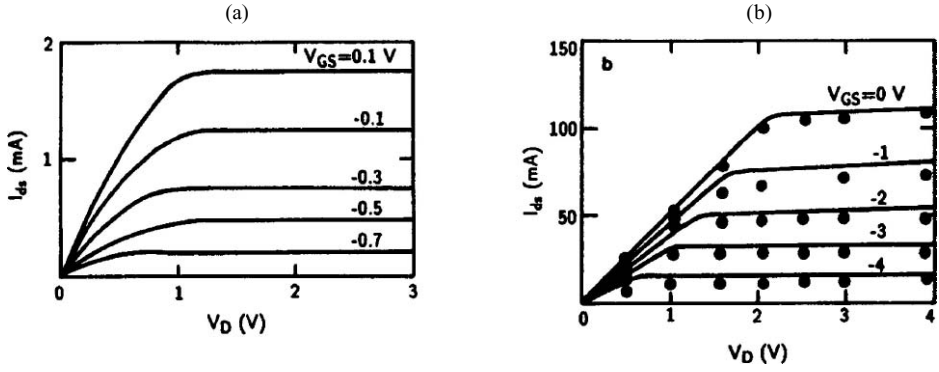


FIGURE 16.6. IDS versus V_{DS} for (a) a GaAs MESFET with low pinch-off voltage ($V_{p0} = 1.8V$, $N_D = 1.8 \times 10^{17}$ cm^{-3}, $L = 1.3$ μm, and $Z = 20$ μm), and (b) a GaAs MESFET with high pinch-off voltage ($V_{p0} = 5.3$ V, $N_D = 6.5 \times 10^{16}$ cm^{-3}, $L = 10$ μm, and $Z = 500$ μm). After Pucel et al. (2) by permission.

16.2.3. Small-Signal Device Parameters

The small-signal ac device parameters for the MESFET are presented in this section. A general expression for the transconductance g_m in the saturation region can be derived from (16.12) by taking the derivative of the channel current with respect to the gate voltage at $V_i = V_s$, which yields

$$g_m = \left.\frac{\partial I_{ch}}{\partial V_{GS}}\right|_{Vi=Vs} = G_0 \frac{(V_s + V_{bi} - V_{GS})^{1/2} - (V_{bi} - V_{GS})^{1/2}}{V_{p0}^{1/2}}. \quad (16.18)$$

If $V_s \ll (V_{bi} - V_{GS})$, then (16.18) reduces to

$$g_m \approx \frac{G_0 V_s}{2[V_{p0}(V_{bi} - V_{GS})]^{1/2}} = v_s Z \left[\frac{q N_D \varepsilon_0 \varepsilon_s}{2(V_{bi} - V_{GS})}\right]^{1/2}. \quad (16.19)$$

It is of interest to note that (16.19) is completely different from that predicted by Shockley theory, according to which the transconductance in the saturation region should equal the drain conductance in the linear region as given by (16.14). However, it has been shown that the theoretical prediction given by (16.19) agrees well with experimental data for a GaAs MESFET. If the effects due to the gate-to-source series resistance R_s and the source contact resistance R_{sc} are taken into account, then the intrinsic transconductance given by (16.19) should be modified to

$$g'_m = \frac{g_m}{1 + (R_s + R_{sc}) g_m}, \quad (16.20)$$

which is smaller than the value of g_m predicted by (16.19).

To derive the unity current gain cutoff frequency f_T for a MESFET, one should first derive expressions for the total charge under the gate and the gate–source capacitance, using the depletion layer width under the gate given by (16.11). If one assumes that $V(x)$ is a linear function of position in the channel under the gate (i.e., $V(x) = V_i(x/L)$), then (16.11) can be rewritten as

$$W_d(x) = \frac{a(V_i x/L + V_{bi} - V_{GS})^{1/2}}{2V_{p0}^{1/2}}. \quad (16.21)$$

Now using (16.21), the total charge under the gate in the linear region for $V_i \le V_s$ can be obtained with the aid of (16.21), yielding

$$\begin{aligned} Q_d &= q N_D Z \int_0^L W_d(x)\, dx \\ &= \frac{2ZL(2\varepsilon_0 \varepsilon_s N_D)^{1/2}}{3V_{ch}} [(V_{ch} + V_{bi} - V_{GS})^{3/2} - (V_{bi} - V_{GS})^{3/2}]. \end{aligned} \quad (16.22)$$

For $V_i \ll (V_{bi} - V_{GS})$, the total charge under the gate can be approximated to

$$Q_d \approx (q N_D Z L a/2) \left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)^{1/2}. \quad (16.23)$$

Equation (16.22) allows one to derive expressions for the gate-to-source and drain-to-gate capacitances of a GaAs MESFET shown in Figure 16.3a, with result

$$C_{gs} = \left.\frac{\partial Q_d}{\partial V_{GS}}\right|_{(V_{ch}-V_{GS}\,=\,\text{const})}$$
$$= \frac{2ZL(2\varepsilon_0\varepsilon_s N_D)^{1/2}}{3V_{ch}^2}\Big[(V_{ch}+V_{bi}-V_{GS})^{3/2}-(V_{bi}-V_{GS})^{3/2}$$
$$-\frac{3}{2}(V_{bi}-V_{GS})^{1/2}V_{ch}\Big]. \tag{16.24}$$

Similarly,

$$C_{dg} = \left.\frac{\partial Q_d}{\partial V_{ch}}\right|_{(V_{GS}\,=\,\text{const})}$$
$$= \frac{2ZL(2\varepsilon_0\varepsilon_s N_D)^{1/2}}{3V_{ch}^2}\Big[\frac{3}{2}V_{ch}(V_{ch}+V_{bi}-V_{GS})^{3/2}$$
$$-(V_{bi}-V_{GS})^{3/2}-(V_{bi}-V_{GS})^{3/2}V_{ch}\Big]. \tag{16.25}$$

If $V_i \ll (V_{bi}-V_{GS})$, then (16.24) and (16.25) can be simplified to

$$C_{gs} = C_{dg} = \frac{ZL}{2}\sqrt{\frac{\varepsilon_0\varepsilon_s q N_D}{2(V_{bi}-V_{GS})}}. \tag{16.26}$$

The current gain β in the common-source configuration is given by

$$\beta = \frac{i_{DS}}{i_{GS}} = \frac{g_m}{\omega C_{gs}}, \tag{16.27}$$

where i_{DS} and i_{GS} denote the small-signal ac drain-to-source current and gate-to-source current, respectively. The unity current gain cutoff frequency is obtained from (16.27) by setting β equal to one. Thus,

$$f_T \approx \frac{g_m}{2\pi C_{gs}}. \tag{16.28}$$

For $V_i \ll (V_{bi}-V_{GS})$, (16.28) reduces to

$$f_T \approx \frac{v_s}{\pi L} = \frac{1}{\pi\tau}, \tag{16.29}$$

where v_s is the saturation velocity of electrons, and $\tau = L/v_s$ is the transit time for electrons to travel the length of the gate at saturation velocity. For a GaAs MESFET with 1 μm gate length, the value of f_T was found to be 25.5 GHz, which is in good agreement with experimental data.

Equation (16.29) shows that in order to maintain high-frequency operation in a GaAs MESFET, it is imperative that the transit time be kept as short as possible. In practice, the unity current gain cutoff frequency is usually lower than that predicted

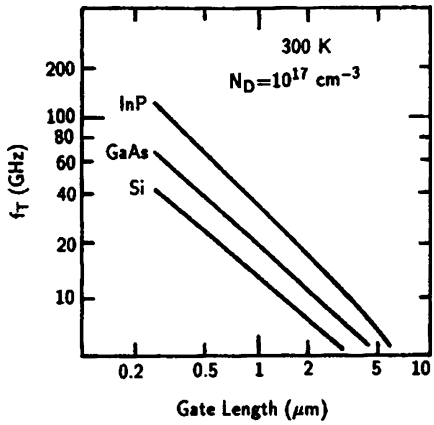


FIGURE 16.7. Calculated unity gain cutoff frequency f_T as a function of the gate length for silicon, GaAs, and InP FETs at 300 K. After Sze,[3] with permission by Wiley.

by (16.29) due to the gate fringing capacitance, interelectrode capacitances, and other parasitic effects. Figure 16.7 shows the calculated f_T as a function of gate length for silicon, GaAs, and InP FETs. It is noted that the InP FET has a higher f_T than that of a GaAs FET because of its higher peak velocity. For GaAs MESFETs with gate lengths less than 0.5 μm, one can expect f_T to be greater than 30 GHz.

Experimental results show that the ratio $g_m C_{gs}$ depends on both V_{GS} and V_{DS} even above the knee voltage. In practice, velocity saturation is not attained at the source end of the gate, and the properties of the constant-mobility region of the channel must be taken into account. This causes g_m to decrease more rapidly with V_{GS} than predicted by (16.16). The reduction of carrier density and carrier mobility near the active layer/substrate interface as well as the conduction in the substrate or buffer layer can also cause a reduction of g_m. The combined effect of these deviations from the simple model is to cause the current gain to diminish as the gate potential approaches pinch-off and the channel current approaches zero.

In order to maintain high transconductance, it is important to reduce the effective source resistance of the MESFET. Further degradation of g_m may be caused by the inductance of a bond wire between the source and the ground. The reduction of such an inductance is an important feature of GaAs power MESFETs where many small FET elements are connected in parallel.

The small-signal equivalent circuit of a MESFET operating in the saturation region in a common-source configuration is shown in Figure 16.3b. It is seen that in an intrinsic FET, the total gate-to-channel capacitance is equal to the sum of C_{gd} and C_{gs}, and the input resistances R_i and R_{ds} under the gate show the effects of channel resistance. The extrinsic (or parasitic) elements include the source resistance R_s, the drain resistance R_d, and the substrate capacitance C_{sd}.

The gate current flowing through the Schottky barrier gate-to-channel junction of the MESFET is given by

$$I_G = I_s[e^{qV_G/nk_BT} - 1], \tag{16.30}$$

where $I_s = A^*T^2A\exp(-q\phi_{Bn}/k_BT)$ is the saturation current of the Schottky barrier contact at the gate; n is the diode ideality factor. The input resistance R_i for the MESFET can be expressed by

$$R_i = \left(\frac{\partial I_G}{\partial V_G}\right)^{-1} = \left(\frac{nk_BT}{q}\right)(I_G + I_s)^{-1}. \tag{16.31}$$

As I_G approaches zero, the input resistance becomes very high (e.g., $R_i = 250\,\text{M}\Omega$ for $I_s = 10^{-10}$ A). The source and drain series resistances, which cannot be modulated by the gate voltage, will introduce a voltage drop between the gate–source and gate–drain electrodes. As a result, they will reduce the drain conductance as well as the transconductance in the linear region (i.e., the low-field constant-velocity regime) of operation. In the saturation region, however, the transconductance is affected only by the source resistance, because for $V_D > V_{DS}$, increasing V_D will have little or no effect on the drain current.

The characteristic switching time of a GaAs MESFET operating in the saturation region can be derived from (16.14), (16.15), and (16.23), yielding

$$\tau_s = \frac{Q_d(V_s)}{I_{sat}} \approx \frac{L}{v_s}\left(\frac{V_{bi} - V_{GS}}{V_{p0} + V_{bi} - V_{GS}}\right)^{1/2}. \tag{16.32}$$

Equation (16.32) shows that the switching time is proportional to the transit time under the gate and that the saturation velocity rather than the peak velocity determines the switching time. It is also shown that decreasing the gate length can reduce the switching time and increase the cutoff frequency of the MESFET. However, there are some physical limitations that can limit the switching speed of the MESFET, particularly the parasitic capacitances between the gate, drain, and source contacts. For a typical GaAs MESFET, a switching time in the picosecond range can be easily obtained.

Finally, we consider the power required by a MESFET at the saturation point. This is given by

$$\begin{aligned} P &= (qN_Dv_sLZa\mathcal{E}_s)\left[1 - \left(\frac{V_{bi} - V_{GS}}{V_p}\right)^{1/2}\right] \\ &= \left(\frac{qN_DZL^2a\mathcal{E}_s}{\tau}\right)\left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)^{1/2}. \end{aligned} \tag{16.33}$$

From (16.32) and (16.33), the power-delay product can be written as

$$P\tau = (qN_DZaL_2\mathcal{E}_s)\left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)^{1/2} = ZL^2\mathcal{E}_s\sqrt{2\varepsilon_0\varepsilon_s qN_D(V_{bi} - V_{GS})}, \tag{16.34}$$

where $\mathcal{E}_s = v_s/\mu_n$ is the average electric field under the gate that reaches the domain-sustaining field in the drain region. Using (16.34), the value of the power-delay product for a typical GaAs MESFET is estimated to be in the femtojoule range, which is in good agreement with experimental data.

16.2.4. Second-Order Effects

It is noted that the behavior of a practical GaAs MESFET does not always follow the theoretical predictions described in the previous section. There are several second-order effects observed in MESFETs attributable to the nonideality (e.g., high concentration of carbon acceptors and EL2 deep donor centers) of the semi-insulating GaAs substrate. The second-order effects include backgating (or sidegating), light sensitivity, low output resistance, low source–drain breakdown voltage, low output power gain at RF frequencies, drain current transient lag effects, temperature dependence, and the subthreshold current effect. Among these problems, backgating is the most significant for both digital and analog circuit applications.

The backgating effect refers to the reduction of drain current in a MESFET as a result of the presence of other nearby neighboring MESFETs that happen to be negatively biased with respect to the source of the device under consideration. In response to changes in voltage on the substrate or adjacent devices, the substrate conducts enough current to modulate the interface space-charge region. When this interfacial depletion region widens into the active channel, the drain–source current I_{DS} is reduced.

The degree of backgating effect in a MESFET can vary significantly from substrate to substrate, making the prediction of backgating threshold unreliable. One approach, which has often been used to alleviate this problem, is the use of proton (or oxygen) implantation between the MESFETs devices. The unannealed implantation produces a high concentration of defects, which act as electron traps at the surface down to a depth of 30–40 nm. The backgating threshold voltage is significantly increased through this process step, and the effect of backgating is reduced considerably. The proton bombardment is usually carried out after the alloying step of ohmic contacts.

The effect of backgating can also be reduced by growing a high-resistivity buffer layer on a semi-insulating GaAs substrate, as shown in Figure 16.1a. A number of possible buffer layers that have been suggested for this purpose include undoped GaAs, AlGaAs, and superlattices (GaAs/GaAlAs). Recently, a new buffer layer grown by the MBE technique at a low substrate temperature ($T = 150$ to 300°C) using Ga and As_4 beam fluxes has been developed. It is highly resistive, optically inactive, and crystalline. High-quality GaAs active layers have been grown on top of this LT (low temperature) buffer layer. GaAs MESFETs fabricated in the active layer grown on top of such an LT GaAs buffer layer have shown total elimination of backgating and sidegating effects with a significant improvement in output resistance and breakdown voltages, while other characteristics of MESFET

performance remain about the same as those of other MESFETs reported in the literature, using alternative approaches.

The drain lag effect refers to the drain current overshooting and recovering slowly when a positive step voltage V_{DS} is applied to the drain electrode under saturation operation. The effect is attributable to the presence of deep-level defects (e.g., EL2 centers) in the semi-insulating GaAs substrate below the channel of MESFETs. In saturation, there is an accumulation of electrons beyond the velocity saturation point, where the channel becomes very narrow and the electric field is very high at the drain end of the gate. Therefore, the drain current becomes very sensitive to a small variation in channel height. If a positive voltage step is applied to the drain electrode, the capacitance through the substrate between the drain electrode and the channel will cause a sudden widening of the channel, leading to an abrupt increase in drain current. Another manifestation of this effect can be observed in the frequency domain, which shows a considerable increase of small-signal output conductance g_{ds} with frequency in the saturation region for frequencies between 100 Hz and 1 MHz at room temperature. The effect can be explained in terms of trapping and capture mechanisms taking place at the channel/substrate interface. At high frequencies, the traps are too slow to capture and release electrons during one cycle of the ac signal, and hence they do not counteract the effect of drain capacitance on the channel–substrate interface, and thereby the drain conductance is large in the saturation region. On the other hand, the traps can follow the ac signal at low frequencies and effectively shield the channel from the drain capacitance through the substrate, and thus the drain conductance is decreased.

The subthreshold current flow in the channel from source to drain electrode beyond the pinch-off voltage is a well-known phenomenon in a MESFET device. The pinch-off is a transition between a region of normal conduction in which the current conduction in the channel is due to the drift of electrons and a region of subthreshold conduction in which the currents are due to both drift and diffusion. For small V_{DS}, electrons can be transported by diffusion (via a concentration gradient between the source and drain electrodes), and the current flow is characterized by an exponential dependence of I_{DS} on V_{DS} and V_{GS}.

Finally, the temperature effect should also be considered in MESFET device operation. The temperature dependence of the drain current of a MESFET is influenced by two related mechanisms, namely, the variation of the built-in voltage (V_{bi} of the channel–substrate interface) and the variation in the channel transconductance factor K. In fact, there are two built-in voltages of interest that exhibit temperature dependence. The built-in potentials for both the Schottky barrier gate and the channel–substrate interface are affected by the temperature dependence of $V_n\,[= (k_B T/q)\ln(n_0/N_c)]$. Any change in these built-in voltages will affect the threshold voltage of the MESFET. The channel transconductance parameter K, as defined in the square-law relationship $I_D = K(V_{GS} - V_T)^2$, can also vary with temperature. The channel transconductance factor $K(= Z\mathcal{E}\mu_n/2La)$ is found to decrease with increasing temperature because the mobility decreases with increasing

temperature and the effective channel thickness increases with temperature due to the temperature dependence of V_{bi} discussed above.

16.3. High Electron Mobility Transistors

16.3.1. Introduction

Besides GaAs MESFETs discussed in the previous section, several newly developed high-speed devices such as HEMTs, RTDs, and HETs using lattice-matched AlGaAs/GaAs and InGaAs/InAlAs material systems have been developed for high-speed and high-frequency applications. Furthermore, nonlattice-matched pseudomorphic quantum-well structures such as AlGaAs/InGaAs/GaAs have also been successfully grown in conventional AlGaAs/GaAs HEMTs without extensive crystal defects if the InGaAs layer is thin enough (i.e., less than 20 nm) so that lattice mismatch can be accommodated by elastic strain rather than by the formation of dislocations. In fact, significant improvement in low-noise microwave performance has been accomplished in pseudomorphic HEMTs relative to conventional HEMTs. In this section the basic device principles, structure, and characteristics of an AlGaAs/GaAs HEMT device are described.

The AlGaAs/GaAs modulation-doped (or high electron mobility) field-effect transistor (MODFET or HEMT) introduced in 1981 has offered high-speed and excellent gain, low noise, and power performance at microwave and millimeter-wave (30 to 300 GHz) frequencies. This device, using novel properties of the two-dimensional electron gas (2-DEG) at the interface between the GaAs and $Al_xGa_{1-x}As$ epitaxial layer and an evolutionary improvement over the GaAs MESFET, has been used extensively in both hybrid and monolithic integrated circuits. The concept evolves from the fact that high electron mobilities in the undoped 2-DEG GaAs layer can be achieved if electrons are transferred across the heterointerface from the heavily doped, wider-bandgap $Al_xGa_{1-x}As$ layer ($x \approx 0.3$) to the nearby undoped GaAs buffer layer. This process is now known as *modulation doping*, and FETs formed using such a structure are called MODFETs. In addition to the name MODFET, other acronyms such as HEMT (high electron mobility transistor), TEGFET (two-dimensional electron gas field-effect transistor), and SDFET (selectively doped field-effect transistor) have also been used in the literature. These acronyms are all descriptive of various aspects of the same device. The most commonly used name for this device, however, is the HEMT. It is worth noting that the HEMT device is comparable to Josephson junction devices for high-speed applications with very short switching times and low power dissipation. The conventional AlGaAs/GaAs HEMT is very similar to a GaAs MESFET.

In this section theoretical aspects of the 2-DEG in a modulation-doped AlGaAs/GaAs heterostructure are discussed. The basic characteristics and factors affecting the performance of a HEMT and the model for predicting current–voltage (I–V) and capacitance–voltage (C–V) behaviors will also be discussed in this section.

The switching speed of a FET can be improved by reducing the carrier transit time and increasing the values of drain current I_{DS} and transconductance g_m of

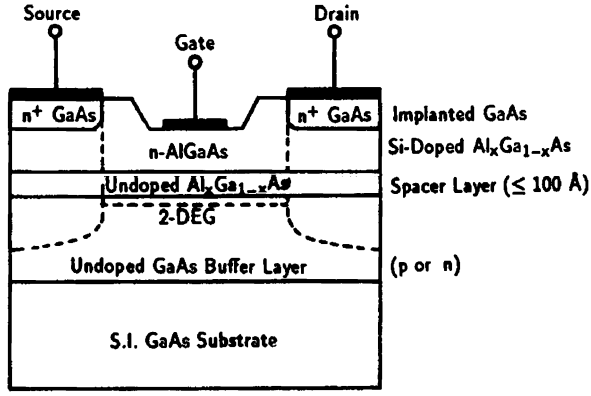


FIGURE 16.8. Cross-sectional view of a GaAs HEMT, showing the heavily doped n^+-GaAs implanted regions under source and drain ohmic contacts, Si-doped AlGaAs layer, 2-DEG region formed between the undoped AlGaAs spacer layer and the undoped GaAs buffer layer grown on a semi-insulating. GaAs substrate.

the device. In order to design a high-speed FET, one needs to optimize the device parameters including smaller gate length L, higher carrier concentration n_0, higher saturation velocity v_s, and larger gate width-to-length ratio (i.e., aspect ratio). Increasing the dopant density beyond 5×10^{18} cm^{-3} while maintaining high-saturation velocity v_s, however, cannot be achieved simultaneously in a MESFET. Increasing the dopant density will reduce the electron velocity due to the increase of ionized impurity scattering. Therefore, to meet the requirements of large n_0 and v_s, a HEMT structure is employed. Figure 16.8 shows a cross-sectional view of an AlGaAs/GaAs HEMT structure. The structure consists of an undoped GaAs buffer layer, an undoped $Al_xGa_{1-x}As$ spacer layer, and a Si-doped n^+-$Al_xGa_{1-x}As$ ($x = 0.33$) layer grown sequentially on a semi-insulating (S.I.) GaAs substrate using the MBE technique. The source and drain regions are formed by the ion implantation of an n^+-GaAs layer on the doped n^+-$Al_xGa_{1-x}As$ layer. A Schottky barrier contact is formed on the doped $Al_xGa_{1-x}As$ between the source and drain contacts to serve as the gate electrode. Since the bandgap energy for $Al_xGa_{1-x}As$ (e.g., $E_g = 1.8$ eV for $x = 0.3$) is larger than for GaAs, and the energy level of the GaAs conduction band is lower than that of $Al_xGa_{1-x}As$, electrons will diffuse from the doped $Al_xGa_{1-x}As$ top layer into the undoped GaAs buffer layer. These electrons cannot drift away from the AlGaAs/GaAs interface in the GaAs layer due to the Coulomb attractive force of the ionized donor impurities in the AlGaAs layer. Nor can they go back to the doped AlGaAs layer due to the potential barrier at the heterointerface of the AlGaAs/GaAs layer. A triangular potential well (quantum well) is formed in the undoped GaAs layer with well width less than the de Broglie wavelength (i.e., ≈ 25 nm for GaAs at 300 K) and quantization of energy levels results along the direction perpendicular to the heterointerface. As a result, a two-dimensional electron gas (2-DEG) sheet charge is accumulated inside the triangle potential well at the undoped GaAs layer near the GaAs/AlGaAs interface, which

forms a conducting channel for the HEMT device. The density of 2-DEG sheet charge can be modulated by the applied gate voltage in the device. The dopant densities and dimensions are chosen such that the AlGaAs layer is fully depleted of free electrons, and the channel conduction is due primarily to the 2-DEG sheet charge in the undoped GaAs buffer layer. In general, the density of the 2-DEG sheet charge will depend on the doping density and aluminum mole fraction x of the Si-doped $Al_xGa_{1-x}As$ layer, and the thickness of the undoped $Al_xGa_{1-x}As$ spacer layer.

Since the 2-DEG sheet charge in the channel of the undoped GaAs buffer layer is spatially separated from the ionized donor impurities by a thin undoped AlGaAs spacer layer (≤ 6 nm), they will experience no ionized impurity scattering inside the triangle potential well in the undoped GaAs buffer layer. Consequently, both the electron mobility and electron velocity in the channel are expected to be very high. For example, 2-DEG electron mobility values up to 8500 $cm^2/V{\cdot}s$ at 300 K and 50,000 $cm^2/V{\cdot}s$ at 77 K have been obtained in the channel region of the GaAs buffer layer. This, along with the small separation ($\approx$ 30 nm) between the gate and conducting channel, leads to extremely high transconductance (e.g., $g_m > 500$ mS/mm at 77 K), large current-carrying capabilities, small source resistances, and very low noise figures. For example, a HEMT amplifier operating at 77 K has a noise figure of 0.25 dB at 10 GHz and 0.35 dB at 18 GHz. Since the transconductance g_m is large, one expects that the unity current gain cutoff frequency f_T will be very high for a HEMT device. Since the maximum oscillation (for unity power gain) frequency f_{max} of a HEMT is strongly influenced by the parasitics (i.e., gate and source resistances, rf drain conductance, feedback and input capacitances), it is essential that these parasitic components be minimized in order to further improve the high-frequency performance. We shall next derive theoretical expressions for the Fermi level, density of 2-DEG sheet charge, electric field at the interface, and the current–voltage characteristics of a HEMT device.

16.3.2. Equilibrium Properties of 2-DEG in GaAs

In this section, the Fermi level, the density of 2-DEG sheet charge, and the electric field at the interface of a modulation-doped AlGaAs/GaAs heterostructure are discussed. Only the lowest (ground state) and first excited subbands (i.e., E_0 and E_1) in the triangular potential well of the undoped GaAs buffer layer will be considered. If the electric field in the triangular potential well is assumed quasi-constant, then the solution for the longitudinal quantized energy levels can be expressed as

$$E_n \approx \left(\frac{\hbar^2}{2m_l^*}\right)^{1/3} \left(\frac{3}{2}\pi q\mathcal{E}\right)^{2/3} \left(n + \frac{3}{4}\right)^{2/3}, \tag{16.35}$$

where m_l^* is the longitudinal effective mass, n is the quantum number, and $\mathcal{E}$ is the electric field. For GaAs, the two subbands E_0 and E_1 in which the 2-DEG sheet

charge resides are given by

$$E_0 \approx 1.83 \times 10^{-6}\mathcal{E}^{2/3}\ \text{eV} \quad \text{and} \quad E_1 \approx 3.23 \times 10^{-6}\mathcal{E}^{2/3}\ \text{eV}, \tag{16.36}$$

where $\mathcal{E}$ is the electric field in V/m.

In order to derive the current–voltage relation for the HEMT, it is important first to establish a relationship between the interface electric field ε_{i1} and the 2-DEG sheet charge concentration. For an AlGaAs/GaAs modulation-doped heterostructure, the electric field in the undoped GaAs buffer layer obeys the Poisson equation, which is given by

$$\frac{\mathrm{d}\mathcal{E}_1}{\mathrm{d}x} = -\frac{q}{\varepsilon_0\varepsilon_s}[n(x) + N_{al}], \tag{16.37}$$

where $n(x)$ is the bulk electron concentration and N_{a1} is the ionized acceptor density in the undoped p-GaAs buffer layer. Integration between the limit of the depletion region ($\mathcal{E}_1 = 0$) and the interface ($\mathcal{E}_1 = \mathcal{E}_{i1}$) yields

$$\varepsilon_0\varepsilon_1\mathcal{E}_{i1} = qn_s + qN_{a1}W_1, \tag{16.38}$$

where ε_1 is the dielectric permittivity of the undoped p-GaAs, n_s is the 2-DEG sheet charge density, and W_1 is the depletion layer width. In a HEMT, N_{a1} in the undoped p-GaAs layer is usually very small, and hence the second term in (16.38) is negligible. Thus, (16.38) becomes

$$\varepsilon_0\varepsilon_1\mathcal{E}_{i1} \approx qn_s. \tag{16.39}$$

Equation (16.39) can also be applied to the undoped n-type GaAs layer. The subband positions given in (16.35) can also be expressed in terms of the 2-DEG sheet charge concentration, and they are given by

$$E_0 = \gamma_0 n_s^{2/3} \quad \text{and} \quad E_1 = \gamma_1 n_s^{2/3}, \tag{16.40}$$

where γ_0 and γ_1 are parameters that can be adjusted to obtain the best fit with the experimental data. To deal with the 2-DEG sheet charge in the undoped GaAs buffer layer, the relation between n_s and the Fermi level position must be derived first. The density of states associated with a single quantized energy level for a 2-DEG electron system is a constant, which can be expressed by

$$D = \frac{qm^*}{\pi\hbar^2}, \tag{16.41}$$

where a spin degeneracy of 2 and a valley degeneracy of 1 have been used. Between the two subbands E_0 and E_1, the 2-dimensional density of states is given by D, and for energies between E_1 and E_2 it is equal to $2D$. Using the Fermi–Dirac statistics, the 2-DEG sheet charge concentration as a function of the Fermi level position and temperature can now be written as

$$\begin{aligned} n_s &= D\int_{E_0}^{E_1} \frac{\mathrm{d}E}{1 + \mathrm{e}^{(E-E_F)/k_BT}} + 2D\int_{E_1}^{\infty} \frac{\mathrm{d}E}{1 + \mathrm{e}^{(E-E_F)/k_BT}} \\ &= D\left(\frac{k_BT}{q}\right)\ln[(1 + \mathrm{e}^{(E_F-E_0)/k_BT})(1 + \mathrm{e}^{(E_F-E_1)/k_BT})], \end{aligned} \tag{16.42}$$

which at low temperatures reduces to

$$n_s = D(E_F - E_0) \tag{16.43}$$

if the second subband is empty, and

$$n_s = D(E_1 - E_0) + 2D(E_F - E_1) \tag{16.44}$$

if the second subband is occupied. From the published data taken at low temperatures using Shubnikov, de Hass, or cyclotron resonance experiments, values of γ_0 and γ_1 for p-type GaAs are found equal to 2.5×10^{-12} and 3.2×10^{-12}, respectively. From the measured cyclotron mass, D is found to be 3.24×10^{21} cm^{-2} V^{-1}. It is seen from (16.43) that the 2-DEG sheet charge density n_s is equal to the product of the density of states D and the energy difference between the Fermi level and the ground state when the second subband is empty.

Figures 16.9a and b show the energy band diagrams of a single-period modulation-doped n$^+$-$Al_xGa_{1-x}As$/p-GaAs heterostructure: (a) in equilibrium and isolated from the influence of any external contact, and (b) under an applied gate bias voltage. It is noted that the position for the two presumed subbands in the quasitriangular potential well shown is only for illustrative purposes. The structure consists of a Si-doped $Al_xGa_{1-x}As$ ($x \leq 0.3$) layer of thickness d_d and an undoped $Al_xGa_{1-x}As$ spacer layer of thickness d_i (i.e., 2 to 6 nm), which serves as a buffer layer to further reduce the scattering of 2-DEG in the undoped GaAs layer (thickness d_1) by the ionized impurities in the space-charge region of the Si-doped $Al_xGa_{1-x}As$ layer. The electric displacement vector at the interface of $Al_xGa_{1-x}As$/GaAs can be calculated using the depletion approximation in the space-charge layer. In this case, the potential $V_2(x)$ in the space-charge region of the $Al_xGa_{1-x}As$ can be derived from the Poisson equation, which is given by

$$\frac{d^2V_2(x)}{dx^2} = -\frac{q}{\varepsilon_0\varepsilon_2}N_{d2}(x). \tag{16.45}$$

If the heterojunction interface is chosen as origin, then the following boundary conditions prevail:

$$V_2(0) = 0, \left(\frac{dV_2}{dx}\right)_{x=-W_2} = 0, \left(\frac{dV_2}{dx}\right)_{x=0} = -\mathcal{E}_{i2}, \tag{16.46}$$

where W_2 is the space-charge layer width in the Si-doped $Al_xGa_{1-x}As$ layer, and $\mathcal{E}_{i2}$ and $V_2(-W_2)$ are given, respectively, by

$$\mathcal{E}_{i2} = -\left(\frac{q}{\varepsilon_0\varepsilon_2}\right)\int_0^{-W_2} N_{d2}(x)\,dx \tag{16.47}$$

and

$$V_2(-W_2) = v_{20} = \mathcal{E}_{i2}W_2 - \frac{q}{\varepsilon_0\varepsilon_2}\int_0^{-W_2} dx \int_0^x N_{d_d}(x')\,dx'. \tag{16.48}$$

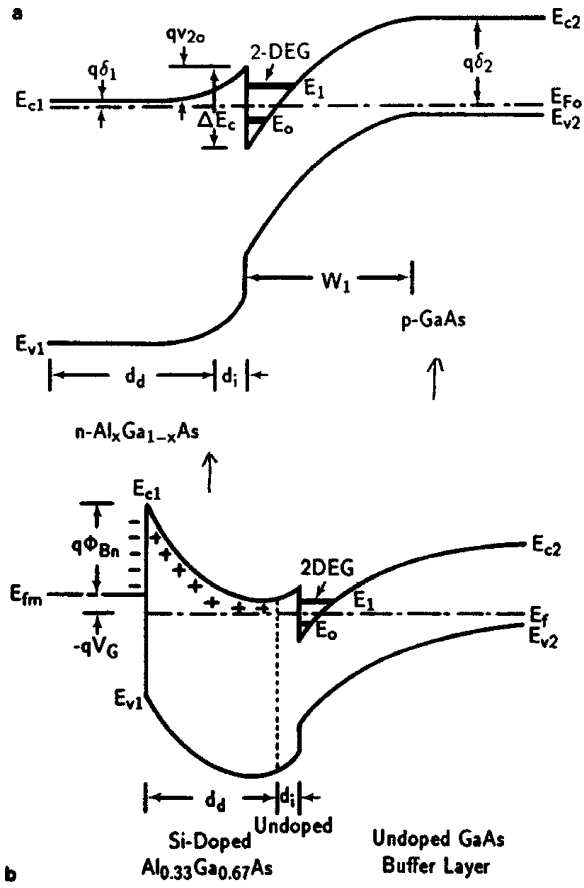


FIGURE 16.9. Energy band diagrams for an $Al_xGa_{1-x}As/GaAs$ HEMT shown in (a) in equilibrium and (b) under applied bias conditions ($V = -V_G$).

For the HEMT structure shown in Figure 16.9a, one can write

$$N_{d2}(x) = \begin{cases} 0 & \text{for } -d_i < x < 0, \\ N_{d2} & \text{for } x < -d_i, \end{cases} \tag{16.49}$$

where N_{d2} is the dopant density in the Si-doped AlGaAs layer. Solving (16.47) through (16.49) yields

$$\varepsilon_0 \varepsilon_2 \mathcal{E}_{i2} = qN_{d2}(W_2 - d_i) \tag{16.50}$$

and

$$v_{20} = \frac{qN_{d2}}{2\varepsilon_0 \varepsilon_2}(W_2^2 - d_i^2). \tag{16.51}$$

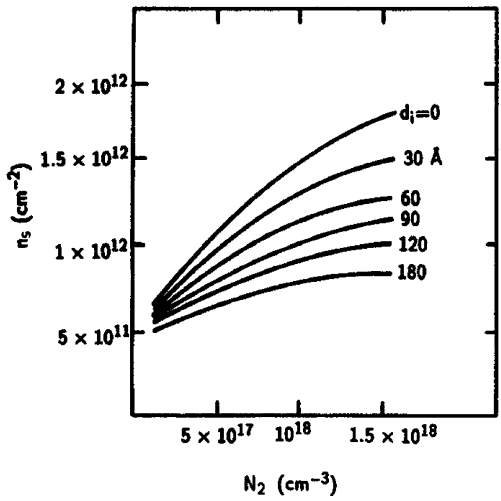


FIGURE 16.10. 2-DEG sheet charge density in the p^--GaAs layer (with $N_a = 10^{14}$ cm^{-3}) as a function of donor impurity density (N_2) in the Si-doped $Al_{0.3}Ga_{0.7}As$ layer for different undoped AlGaAs spacer layer thicknesses calculated at 300 K. Delagebeaudeuf and Ling,[4] by permission, © IEEE–1982.

Since the band bending of the AlGaAs layer at the heterointerface is denoted by v_{20}, one can easily show from (16.50) and (16.51) that

$$\varepsilon_0\varepsilon_2\mathcal{E}_{i2} = \sqrt{2q\varepsilon_0\varepsilon_2 N_{d2}v_{20} + q^2N_{d2}^2d_i^2} - qN_{d2}d_i. \tag{16.52}$$

From Figure 16.9a, one obtains

$$v_{20} = \Delta E_c - \delta_2 - E_{F0}. \tag{16.53}$$

Using Gauss's law and neglecting interface traps, the 2-DEG sheet charge density, which is related to the dielectric constant and the electric field, can be expressed by

$$qn_s = \varepsilon_0\varepsilon_1\mathcal{E}_{i1} = \varepsilon_0\varepsilon_2\mathcal{E}_{i2}. \tag{16.54}$$

Now, solving (16.50) through (16.54) yields a general expression for the 2-DEG sheet charge density, which reads

$$\begin{aligned} n_s &= \sqrt{2\varepsilon_0\varepsilon_2 N_{d2}v_{20}/q + N_{d2}^2d_i^2} - N_{d2}d_i \\ &= \frac{Dk_BT}{q}\ln\left(1 + e^{(E_{F0}-E_0)/k_BT}\right)\left(1 + e^{(E_{F0}-E_1)/k_BT}\right). \end{aligned} \tag{16.55}$$

The Fermi level E_{F0} can be solved numerically from (16.55) by iteration procedures. It is noted that the value of n_s calculated from (16.55) and (16.42) should be the same. Otherwise, values of the Fermi energy E_{F0} must increase until this condition is met. Figure 16.10 shows the 2-DEG sheet concentration in the p^--GaAs layer (with $N_{al} = 10^{14}$ cm^{-3}) as a function of donor impurity density (N_{d2}) in the Si-doped $Al_{0.3}Ga_{0.7}As$ layer for different undoped AlGaAs spacer layer thickness d_i at 300 K. The 2-DEG sheet concentration is seen to increase as the thickness

of the undoped AlGaAs spacer layer is decreased, and to increase as the doping concentration of the Si-doped AlGaAs layer is increased.

16.3.3. 2-DEG Charge Control Regime

The charge control regime, i.e., the region between the Schottky contact on the Si-doped AlGaAs layer and the AlGaAs/GaAs heterointerface, is totally depleted, and the electrostatic potential obeys the Poisson equation subject to the conditions

$$N_{d2}(x) = \begin{cases} 0 & \text{for } -d_i < x < 0, \\ N_{d2} & \text{for } -d_d < x < -d_i. \end{cases} \tag{16.56}$$

If one chooses the origin at the AlGaAs/GaAs interface and lets $V_2(0) = 0$ at $x = 0$, then $V_2(-d_d)$ can be written as

$$V_2(-d_d) = -v_2 = \mathcal{E}_{i2} d_d - \frac{q}{\varepsilon_2} \int_0^{-d_d} dx \int_0^x N_{d2}(x') dx'. \tag{16.57}$$

Equation (16.57) can be readily solved using the boundary conditions given by (16.56), yielding

$$v_2 = \frac{q N_{d2}}{2\varepsilon_0 \varepsilon_2} (d_d - d_i)^2 - \mathcal{E}_{i2} d_d. \tag{16.58}$$

From Figure 16.9b and (16.58) we can establish the relation

$$\varepsilon_0 \varepsilon_2 \mathcal{E}_{i2} = \frac{\varepsilon_0 \varepsilon_2}{d_d} (V_{p2} - v_2), \tag{16.59}$$

where

$$V_{p2} = \frac{q N_{d2}}{2\varepsilon_0 \varepsilon_2} (d_d - d_i)^2. \tag{16.60}$$

By examining Figure 16.9b, the potential v_2 is given by

$$v_2 = \phi_M - V_G + E_F - \Delta E_c. \tag{16.61}$$

Now substituting (16.61) into (16.59) yields

$$\varepsilon_0 \varepsilon_2 \mathcal{E}_{i2} = \left(\frac{\varepsilon_0 \varepsilon_2}{d_d} \right) (V_{p2} - \phi_M - E_F + \Delta E_c + V_G). \tag{16.62}$$

Therefore, in the absence of interface states, the total charge in the 2-DEG GaAs buffer layer can be obtained by solving (16.62) and (16.54), which yields

$$Q_s = q n_s = \frac{\varepsilon_0 \varepsilon_2}{d_d} (V_{p2} - \phi_M - E_F + \Delta E_c + V_G). \tag{16.63}$$

Since E_F, which is a function of V_G, is usually much smaller than the other terms given in (16.63), one can approximate (16.63) by

$$Q_s \approx \frac{\varepsilon_0 \varepsilon_2}{d_d} (V_G - V_{off}), \tag{16.64}$$

where

$$V_{\text{off}} = \phi_{\text{M}} - \Delta E_{\text{c}} - V_{\text{p2}} \tag{16.65}$$

is the "off voltage" that eliminates the 2-DEG. Equations (16.64) and (16.65) are obtained by neglecting the Fermi energy E_{f}, and hence they are insensitive to the exact positions of the two subbands. Therefore, (16.64) and (16.65) are applicable to both p^-- and n^--type GaAs buffer layers. If the interface state charge Q_{i} is included, then (16.65) becomes

$$V_{\text{off}} = \phi_{\text{M}} - \Delta E_{\text{c}} - V_{\text{p2}} - \frac{d_{\text{d}}}{\varepsilon_0 \varepsilon_2} Q_{\text{i}}. \tag{16.66}$$

For a given AlGaAs layer width, there exists a threshold voltage V_{Gth}, which separates the charge control regime from the equilibrium state. This can be obtained by equating the two expressions for $\varepsilon_0 \varepsilon_2 \mathcal{E}_{\text{i2}}$ given by (16.52) and (15.62); one thereby obtains

$$V_{\text{Gth}} = \phi_{\text{M}} - \delta_2 - \left(\sqrt{\frac{q N_{\text{d2}} d_{\text{d}}^2}{2\varepsilon_0 \varepsilon_2}} - \sqrt{(\Delta E_{\text{c}} - \delta_2 - E_{\text{F0}}) + \frac{q N_{\text{d2}} d_{\text{i}}^2}{2\varepsilon_0 \varepsilon_2}} \right)^2 . \tag{16.67}$$

16.3.4. Current–Voltage Characteristics

The current–voltage (I–V) characteristics of a HEMT device can be derived using the charge control model and the gradual channel approximation. If $V_{\text{c}}(x)$ is the channel potential under the gate at position x, and V_{GS} is the applied gate voltage, then the effective potential V_{eff} for charge control at x is given by

$$V_{\text{eff}}(x) = V_{\text{GS}} - V_{\text{c}}(x). \tag{16.68}$$

Using (16.64) and (16.68), the 2-DEG sheet charge in the channel can be rewritten as

$$Q_{\text{s}}(x) = q n_{\text{s}} = \frac{\varepsilon_0 \varepsilon_2}{d} \left[V'_{\text{GS}} - V_{\text{c}}(x) \right], \tag{16.69}$$

where $V_{\text{GS}} = (V_{GS} - V_{\text{off}})$ and $d = d_{\text{d}} + d_{\text{i}} + \Delta d$; $\Delta d = \varepsilon_0 \varepsilon_2 a / q \approx 8\,\text{nm}$ for $a = 1.25 \times 10^{-21}\ \text{V}\cdot\text{cm}^2$. The channel current at position x is given by

$$I_{\text{DS}} = Q_{\text{s}}(x) Z v(x), \tag{16.70}$$

where Z is the gate width and $v(x)$ is the electron velocity at position x. It is seen that $v(x) = \mu \mathcal{E}$ for a 2-DEG sheet charge in the channel is generally field-dependent and the electron mobility (μ) as a function of the electric field can be expressed by

$$\mu = \frac{\mu_0}{1 + \dfrac{1}{\mathcal{E}_{\text{c}}} \dfrac{\text{d}V}{\text{d}x}}, \tag{16.71}$$

where $\mathcal{E}_c$ is the critical field strength in which the velocity saturation occurs, and μ_0 is the low-field electron mobility. For field strengths less than $\mathcal{E}_c$, the drift velocity $v(x)$ is equal to $\mu_0\mathcal{E}$, and v(x) varies linearly with the electric field $\mathcal{E}$ with a constant mobility μ_0; $v(x)$ becomes saturated (i.e., $v(x) = v_s$) if $\mathcal{E} \geq \mathcal{E}_c$. The current–voltage (*I–V*) relation for a HEMT can be derived using a simple two-piecewise linear approximation for the v_d versus $\mathcal{E}$ relation, which may be written as

$$v_d = \begin{cases} \mu_0\mathcal{E} & \text{for} \quad \mathcal{E} < \mathcal{E}_c, \quad (16.72) \\ v_s & \text{for} \quad \mathcal{E} \geq \mathcal{E}_c. \quad (16.73) \end{cases}$$

This is the so-called gradual channel approximation, which is used widely in modeling the *I–V* relation of a FET. Therefore, using the two-piecewise linear approximation and assuming that the electric field in the channel is parallel to the heterointerface, we can derive an analytical expression for the *I–V* relation of a HEMT in the linear region ($V_{DS}/L < \mathcal{E}_c$) and in the saturation region ($V_{DS}/L \geq \mathcal{E}_c$). We note that this approximation does not include the diffusion current component in the channel.

(i) Linear Regime ($\mathcal{E} < \mathcal{E}_c$). In the ohmic regime, where the electric field is much smaller than the critical field strength $\mathcal{E}_c$, the drain current given by (16.70) can be expressed in the form

$$I_{DS} = Q_s(x)Zv(x) = Q_sZ\mu\mathcal{E}_x = Q_sZ\mu - dV_c(x). \quad (16.74)$$

Now substituting $Q_s(x)$ given by (16.69) into (16.74), and integrating both sides of (16.74) from $x = 0$ to $x = L$, we obtain

$$\int_0^L I_{DS}dx = \int_{V_c(0)}^{V_c(L)} \mu Z \frac{\varepsilon_0\varepsilon_2}{d}\left[V_G' - V_c(x)\right]\left(-\frac{dV_c(x)}{dx}\right). \quad (16.75)$$

If the source and drain resistances are neglected, then (16.75) reduces to

$$I_{DS} = \frac{\varepsilon_0\varepsilon_2\mu Z}{dL}\left(V_{GS}'V_{DS} - V_{DS}^2/2\right). \quad (16.76)$$

Equation (16.76) is obtained from the conditions that $V_c = 0$ at $x = 0$ and $V_c = V_{DS}$ at $x = L$. It is noted that I_{DS} is constant in the channel and $V_c(x)$ increases with distance from source to drain. The electric field reaches a maximum near the drain side of the channel, and velocity saturation will occur first at the drain side of the gate region. In the linear region, where the drain voltage V_{DS} is very small, (16.76) can be simplified to

$$I_{DS} = \frac{\varepsilon_0\varepsilon_2\mu Z}{dL}\left(V_{GS}'V_{DS}\right), \quad (16.77)$$

which shows that the drain current varies linearly with drain voltage, and the HEMT device acts like a pure voltage-controlled resistor (by V_{GS}'). If the source and drain resistances are not negligible, then the channel voltages at the source

and drain sides of the gate region, $V_c(0)$ and $V_c(L)$, are given, respectively, by

$$V_c(0) = R_s I_{DS} \tag{16.78}$$

and

$$V_c(L) = V_{DS} - R_D I_{DS}, \tag{16.79}$$

where R_s and R_D denote the source and drain series resistances, respectively. Now solving (16.75), (16.78), and (16.79), one obtains

$$I_{DS} = \frac{\varepsilon_0 \varepsilon_2 \mu Z}{dL} \left\{ V'_{GS} \left[(R_S + R_D) I_{DS} - V_{DS} \right] - \tfrac{1}{2} \left[(R_S + R_D) I_{DS} - V_{DS} \right]^2 \right\}. \tag{16.80}$$

For small V_{DS} and I_{DS}, the first-order approximation enables (16.80) to be reduced to

$$I_{DS} = V_{DS} \left[-(R_S + R_D) + \frac{Ld}{Z\mu\varepsilon_0\varepsilon_2 V'_{GS}} \right]^{-1}, \tag{16.81}$$

which again shows that a linear relation exists between the drain current and drain voltage.

(ii) Saturation regime ($\mathcal{E} \geq \mathcal{E}_c$). In the saturation regime, velocity saturation occurs first at the drain side of the gate region with $\mathcal{E}(L) = \mathcal{E}_c$. The drain current under the saturation condition is given by

$$I_{Dsat} = \frac{Z\varepsilon_0\varepsilon_2 v_2}{d} \left(\sqrt{\left(V'_{GS} - R_S I_{Dsat} \right)^2 + \mathcal{E}_c^2 L^2} - \mathcal{E}_c L \right). \tag{16.82}$$

For a long-gate HEMT, (15.82) can be approximated by

$$I_{Dsat} \approx \frac{Z\varepsilon_0\varepsilon_2\mu}{2dL} \left(V'_{GS} - R_s I_{Dsat} \right)^2, \tag{16.83}$$

and for a short-gate HEMT, (15.82) becomes

$$I_{Dsat} \approx \frac{Z\varepsilon_0\varepsilon_2 v_s}{d} \left(V'_{GS} - R_S I_{Dsat} - \mathcal{E}_c L \right), \tag{16.84}$$

which is valid for $(V'_{GS} - R_s I_{Dsat}) \gg \mathcal{E}_c L$. In fact, the experimental results for a short-gate GaAs HEMT device confirm the linear relation between I_{DS} and V_{GS} in the saturation region.

The two-piecewise linear model for the channel charge and current as functions of the gate and drain voltages presented in this section gives a first-order description of the *I–V* characteristics of a HEMT in the linear and saturation regions of operation. A more realistic three-piecewise linear model for predicting the *I–V* characteristics of a HEMT has also been developed. Such a model uses a three-piecewise linear approximation for the v_D versus $\mathcal{E}$ relation to derive the current–voltage characteristics of a HEMT device. A comparison of the two- and three-piecewise linear approximations with the actual velocity versus electric

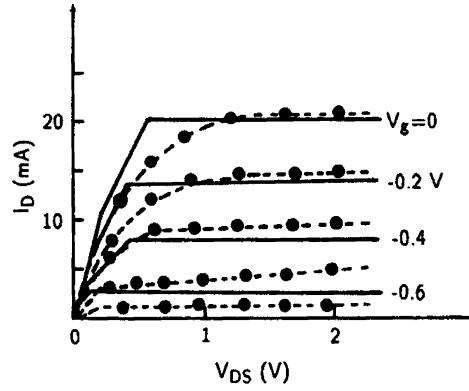


FIGURE 16.11. Comparison of the calculated and measured I_D versus V_{DS} curves for a normally on GaAs HEMT using a three-piece wise linear model. After Lee et al.,[5] by permission, © IEEE-1983.

field relation shows that the three-piecewise linear approximation yields roughly 10–20% improvement in accuracy over the two-piecewise linear approximation. Figure 16.11 shows the *I–V* characteristics of a normally on HEMT. The solid lines are calculated from the three-piecewise linear model and the dot and dashed lines are the experimental data. A more rigorous model using numerical simulation of the current–voltage characteristics of a GaAs HEMT has also been reported in the literature.

From the above description, one sees that some advantages associated with a HEMT device include the large 2-DEG sheet charge density ($\approx 10^{12}$ cm^{-2}), high electron mobility, and high saturation velocity. These unique features have resulted in significant improvement in device performance (i.e., high speed and low noise) compared to conventional GaAs MESFETs. In the past few years, successful development of 0.25-μm-gate-length AlGaAs/GaAs and InGaAs/InP HEMTs has offered new promise for low-noise applications at microwave frequencies. Further improvement in current gain and noise performance of HEMT devices can be achieved by further reducing the gate length to 0.1 μm or less. However, a 0.1 μm gate length generates undesirable short-channel effects. The effects are largely due to the space-charge injection of carriers (which is inversely proportional to the square of the effective gate length) into the buffer layer under the channel. This increases the HEMT output conductance and results in a shift of the pinch-off voltage and transconductance reduction near the pinch-off region. To overcome this problem, an AlGaAs/InGaAs/GaAs pseudomorphic HEMT structure has been developed recently using a 0.1 μm gate length. Using the MBE growth technique, the AlGaAs/InGaAs/GaAs pseudomorphic HEMT is grown at a lower temperature than that of the conventional AlGaAs/GaAs HEMT structure. In this structure a thin strained superlattice (TSSL) of an undoped $In_{0.35}Ga_{0.65}As$ (5 nm)/GaAs (1.5 nm)/$In_{0.35}Ga_{0.65}As$ (5 nm) pseudomorphic active layer structure is grown on an undoped GaAs buffer layer. Due to the InGaAs quantum-well channel structure, this pseudomorphic HEMT structure can greatly improve the carrier confinement and hence reduces the short-channel effects. In addition, due

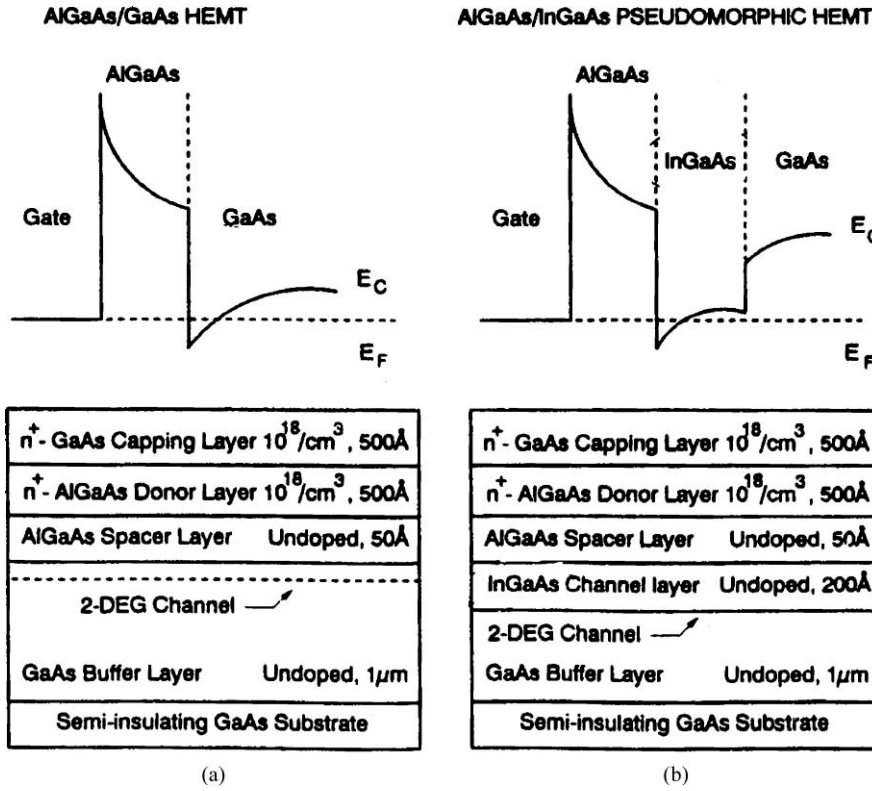


FIGURE 16.12. Energy band diagrams and cross-sectional views for (a) a conventional AlGaAs/GaAs HEMT, and (b) an AlGaAs/InGaAs pseudomorphic HEMT.

to the superior transport properties in the InGaAs channel and the large conduction band discontinuity with the AlGaAs spacer layer (5 nm), the pseudomorphic HEMT also provides a very high electron velocity and 2-DEG sheet charge density. A maximum extrinsic transconductance, g_m, of 930 mS/mm with excellent pinch-off characteristics at room temperature has been reported recently for a 0.1-μm gate-length planar-doped pseudomorphic HEMT. The device has a maximum stable gain of 19.3 dB measured at 18 GHz. At 60 GHz the device demonstrated a minimum noise figure of 2.5 dB with an associated gain of 8 dB. The unity current gain cutoff frequency f_T for this device is around 100 GHz.

Figure 16.12 shows a comparison of a conventional AlGaAs/GaAs HEMT with a GaAs-based pseudomorphic HEMT. The difference between these two structures is that in the latter, a thin (typically 5–20 nm) layer of $In_xGa_{1-x}As$ ($x = 0.15$–0.35) is inserted between the doped AlGaAs barrier layer and the GaAs buffer layer. It is clear that there will be a lattice constant mismatch between AlGaAs/InGaAs/GaAs layers introduced by the thin InGaAs channel layer. The strain effect created by

this lattice mismatch will be absorbed totally by the InGaAs quantum-well layer. If the thickness of the InGaAs layer is less than the critical thickness, then the lattice mismatch strain between the InGaAs and GaAs can be accommodated elastically, and the InGaAs layer is compressed to mirror the structure of the GaAs and AlGaAs layers (hence the name "pseudomorphic"). This critical thickness is dependent on the InAs molar fraction. For example, at 35% InAs molar fraction, the critical thickness is about 5 nm. Since ΔE_c increases and carrier confinement and transport properties improve with higher InAs molar fraction it is desirable to increase the InAs molar fraction to the highest possible level in the pseudomorphic HEMT. However, the lattice mismatch strain between the InGaAs and GaAs layers has precluded the use of a very high InAs molar fraction in the pseudomorphic channel for device applications.

16.3.5. Other III-V Semiconductor HEMTs

In addition to the AlGaAs/GaAs HEMTs discussed above, HEMTs have also been fabricated from InAlAs/InGaAs heterostructure on InP substrates. The advantages of using an $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.17}As$ lattice-matched material system for HEMT fabrication include (1) a large Γ–L valley separation ($\Delta E_{\Gamma\text{-}L} = 0.55$ eV) in $In_{0.53}Ga_{0.47}As$, which leads to low intervalley noise, (2) high density of 2-DEG with $n_s > 3 \times 10^{12}\text{cm}^{-2}$, (3) low sheet resistance of 2-DEG ($\rho \approx 150\ \Omega$/squre) for lower thermal noise, and (4) high electron velocity in short-channel GaInAs ($v_p \approx 6$–7×10^7 cm/s), which leads to high f_T and g_m. A transconductance g_m equal to 800 mS/mm and a cutoff frequency f_T of 130 GHz have been demonstrated for an AlInAs/InGaAs on InP HEMT with a gate length of $L = 0.2\ \mu$m and $W = 25\ \mu$m, at a pinchoff voltage $V_p = -1.3$ V. A strained-channel $In_{0.52}Al_{0.48}As/In_{0.65}Ga_{0.35}As$ HEMT grown by the MBE technique has been reported recently using 0.1 μm T-gate technology; this is shown in Figure 16.13. A strained channel was used to enhance the f_T performance due to its higher electron mobility, velocity, sheet carrier concentration, and conduction band discontinuity (ΔE_c). An f_T value as high as 220 GHz has been obtained for this device. Figure 16.14 shows a comparison of the unity current gain cutoff frequency f_T versus gate length L_g for several III-V HEMTs, GaAs MESFET, and Si NMOS at 300 K. The results show that a value of f_T greater than 300 GHz can be achieved from AlInAs/GaInAs HEMTs with a gate length of 0.1 μm.

Power capabilities of HEMTs at microwave and millimeter-wave frequencies have also made steady progress over the past few years. High-gain and high-efficiency performance HEMTs operating from 10 to 60 GHz have been reported in the literature. The advantages that a HEMT device has over a GaAs MESFET for power applications are higher power gain and higher efficiency. These advantages are due to higher current gain cutoff frequencies resulting from higher electron velocities, which lead to greater output power. For example, a double heterojunction (DH) type HEMT with a 1 μm gate length could generate 1.05 W output power at 20 GHz and 0.58 W at 30 GHz. Recently, millimeter-wave power HEMTs employing a single-chip multiple-channel AlGaAs/GaAs structure have generated output

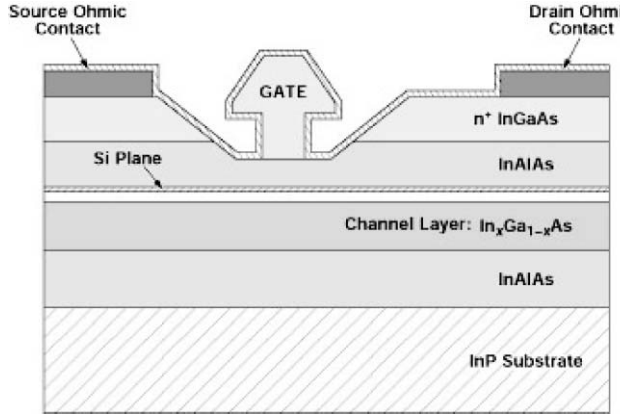


FIGURE 16.13. A typical InGaAs/InAlAs HEMT grown on InP substrates.The low-noise performance can be improved by two approaches: (1) optimizing the indium content and thickness of the channel layer and (2) shortening the gate length.

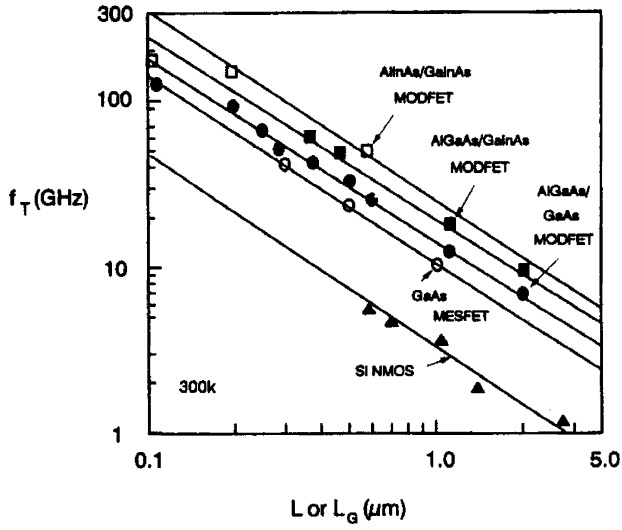


FIGURE 16.14. Comparison of f_T versus gate length for several III-V semiconductor HEMTs, GaAs MESFET, and Si NMOS devices. After Sze,[6] with permission by John Wiley.

power of 1.0 W with 3.1 dB gain and 15.6% efficiency at 30 GHz using a 0.5 μm gate length and 2.4 mm gate periphery.

The conventional AlGaAs/GaAs HEMT has demonstrated significantly better low-noise performance than the GaAs MESFET, at frequencies up to 60 GHz. It has replaced the GaAs MESFET for microwave low-noise applications in many cases at both room temperature and low temperatures. The conventional HEMT has also

established technology for high-speed digital applications. It is expected that the strong trend toward the integration of HEMTs into both hybrid and monolithic low-noise circuits will continue to blossom. As HEMT fabrication technology matures further, it is anticipated that GaAs MESFETs may be replaced entirely by HEMTs for low-noise applications. As for the pseudomorphic HEMT, this device offers even better noise performance than the conventional AlGaAs/GaAs HEMT, and its power performance is also superior to that of either the conventional HEMT or the GaAs MESFET. The GaAs-based pseudomorphic HEMTs, with higher carrier density, are very attractive for microwave power applications.

The InP-based InAlAs/InGaAs HEMTs with extremely short gate length have demonstrated excellent high-frequency performance. Since the unity current gain cutoff frequency f_T is almost inversely proportional to the gate length L_g, a further improvement of high-frequency performance can be achieved with shorter L_g. Using a *T*-shaped gate with gate length of 100 nm, a double-sided δ-doped InAlAs/InGaAs HEMT grown on the InP substrate has achieved a maximum oscillating frequency f_{max} of 241 GHz and a unity gain cutoff frequency f_T of 205 GHz. These InP-based InAlAs/GaInAs HEMTs will most likely make a great impact on ultrahigh-speed digital and millimeter-wave low-noise applications. Finally, wide-band-gap AlGaN/GaN HEMTs formed on SiC substrate have also been reported recently. The AlGaN/GaN modulation-doped heterostructure has some unique features: (1) It is the only heterostructure system in wide-band-gap semiconductors, and hence it can exploit the capabilities of wide-band-gap semiconductors for high-power handling capability and modulation-doped structures for high-speed devices, and (2) the channel sheet charge in an AlGaN/GaN HEMT could reach as high as 10^{13} electrons/cm^2, which is about five times as high as in an AlGaAs/GaAs HEMT; higher channel charge increases the device current handling capability. Combining the features given above with the higher breakdown voltage, it is clear that the AlGaN/GaN HEMT structure is well suited for high-power and high-frequency applications. A typical layer structure of an AlGaN/GaN HEMT consists of a Si-doped AlGaN donor layer with source, drain and gate contacts, a thin undoped AlGaN spacer layer, an undoped GaN layer, a transition layer (buffer), and a SiC substrate for the growth of the AlGaN/GaN HEMT structure.

Finally, it is worth noting that the GaAs MESFET is still the workhorse for MMIC technology and competes directly with advanced Si RF technologies. It is mostly based on ion implantation into GaAs semi-insulating substrates. This is the least expensive process concerning the raw material cost, since no epitaxial layers are required. Current technologies in the market are processed with gate length from 0.8 μm down to 0.25 μm. Values of f_T in the range of 25 GHz are available in production depending on the gate length used. These FETs can easily achieve noise figures below 1 dB in the 1–2 GHz frequency range. The power performance reaches into the 10 W class in *x*-band range for phased-array radar applications. Most of the GaAs MESFET devices are of depletion mode type, which requires negative bias to control the gate. Enhancement mode with shallow channel needs single-polarity supply only. The GaAs HEMT technology in principle is similar

to the GaAs MESFET structure, except that epitaxial wafers provide the active layers. The epilayers could be grown by the MBE or MOCVD technique. These GaAs HEMT structures incorporate single or double heterostructure transitions that deliver high electron mobility in the 2-DEG channel. A gate length down to 0.12 μm is required for achieving an $f_T = 100$ GHz value. For applications in the mobile communications market the gate length is increased to 0.5 μm with values of f_T still up to the 30–40 GHz range. HEMT devices deliver the lowest noise figure of the RF technologies together with high-gain performance.

16.4. Hot Electron Transistors

Hot electron transistors (HETs) are based on an old concept first proposed by C. A. Mead in 1960. The main objective for the HET is to reduce both the base resistance and transit time and to increase the current density of the BJTs for high-frequency performance.[7] To achieve these goals, various metal/insulator/semiconductor structures have been proposed, but due to difficulty in fabricating these structures success has so far been limited.

The capability of growing very thin epitaxial layers on semi-insulating substrates has been greatly enhanced by the availability of the MBE growth technique. For example, using GaAs substrate and lattice-matched large-band-gap $Al_xGa_{1-x}As$, it is possible to grow a high-resistivity undoped $Al_xGa_{1-x}As$ barrier layer on a GaAs base layer with barrier height adjusted by the Al molar fraction x, or using modulation doping to form a metal-like 2-DEG sheet charge on the GaAs base layer. As a result, several different HET structures have been reported. The main difference in these structures is the method by which the hot electrons (i.e., electrons with energies a few k_BT above the conduction band edge) are injected into the thin base region. The hot electron injection can be achieved by injection over the barrier or by tunneling. For base thickness smaller than the mean free path of hot electrons, the majority of hot electrons that inject into the thin base are collected by the collector, where they are thermalized to lattice temperature. The electrons lost in the base constitute the base current, which can usually be removed very rapidly from the base in a HET since the base resistance $r_{bb'}$ is very low, and the transit time of the majority electrons across the thin base is extremely small.

Figure 16.15a shows cross-sectional view and conduction band diagram of an AlGaAs/GaAs tunneling HET in equilibrium, and Figure 16.15b illustrates the conduction band diagram under bias conditions. Figure 16.16 shows the conduction band diagrams of a modified AlGaAs/GaAs HET with 2-DEG sheet charge in the GaAs base formed by (a) modulation doping and (b) applied collector–base bias voltage. The device structures shown in Figures 16.15 and 16.16 have a GaAs base-layer thickness of 0.1 μm or less. In both cases, electrons in the base, being hot, travel at such a high velocity ($\approx 5 \times 10^7$ cm/s) that the base transit time τ_B is negligible. One expects the speed of the HET to be limited by the emitter capacitance charging time through the emitter resistance. Its value depends on the current injection mechanism, and is larger for tunneling HETs. Although the base

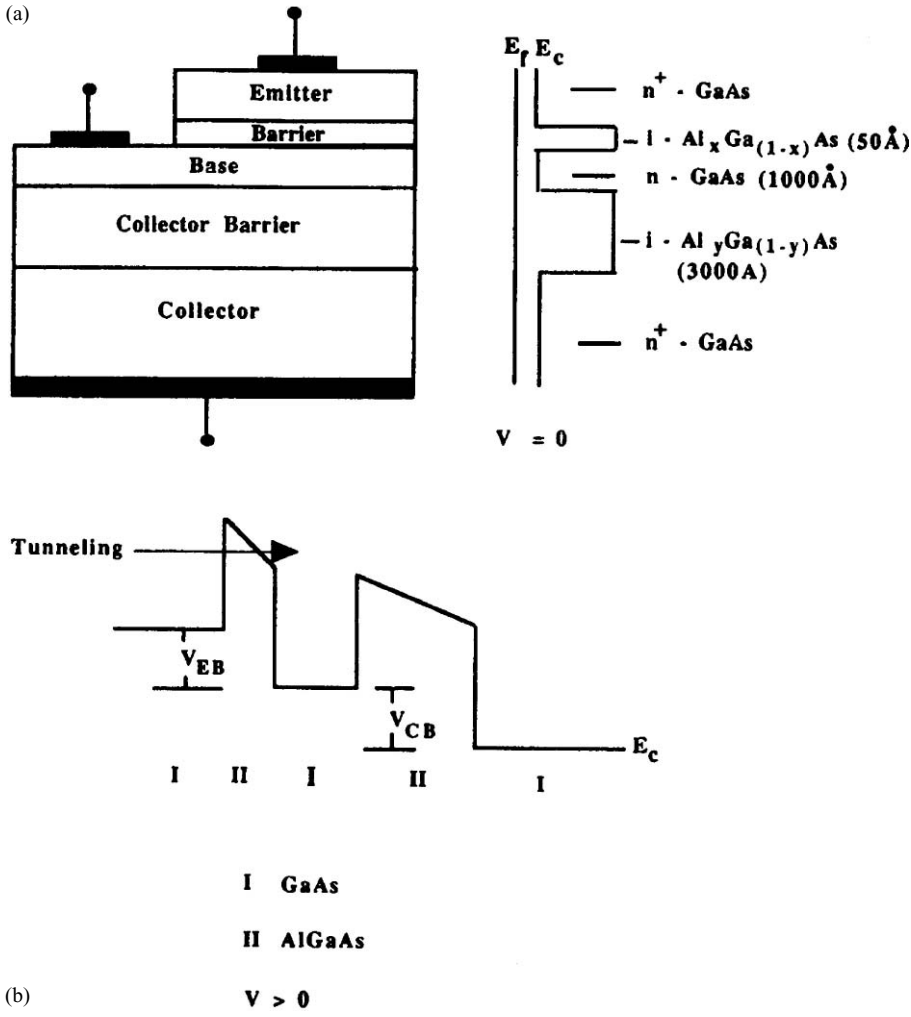


FIGURE 16.15. Conduction band diagrams of a tunneling hot electron transistor (HET) (a) in equilibrium and (b) under bias conditions. After Capasso,[8] by permission, © IEEE-1988.

conductivity in both structures can be made very high by either heavy doping or using 2-DEG, it is usually difficult to make good ohmic contacts to the base. HETs are usually unsuitable for operation at room temperature due to high leakage current caused by the thin injection barrier with low barrier height (i.e., $\phi_B \approx 0.25$ eV for HETs versus 1.3 eV for HBTs). Thus, the predicted subpicosecond performance has yet to be realized in a practical HET device.

The tunneling HET structure shown in Figure 16.15a consists of an n^+-GaAs emitter, an undoped thin (5 nm) $Al_xGa_{1-x}As$ barrier layer, an n^--GaAs base layer

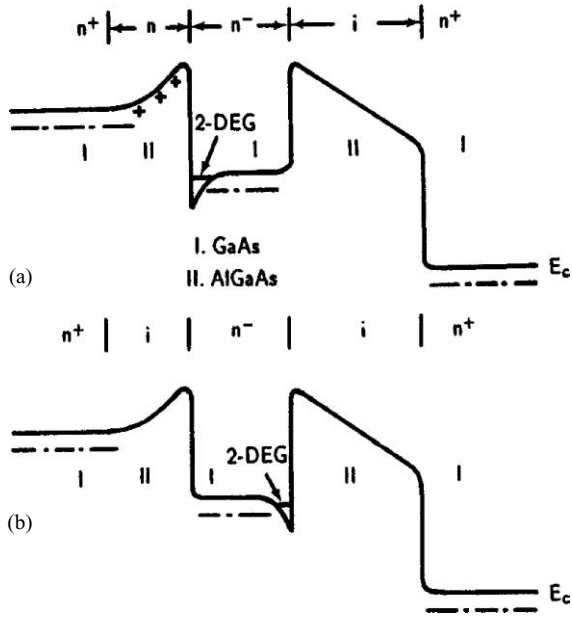


FIGURE 16.16. Conduction band diagrams of a modified HET with a 2-DEG in the base region formed by (a) modulation doping and (b) base collector voltage. The undoped GaAs base thickness is less than 0.1 μm.

(100 nm), an undoped thick (300 nm) $Al_xGa_{1-x}As$ barrier layer, and an n^+-GaAs collector. As shown in this figure, the carrier injection from the emitter to the base relies on tunneling through the AlGaAs barrier layer, which occurs when the base is biased positively with respect to the emitter. The effective barrier width of the AlGaAs barrier for electron tunneling in a HET structure can be controlled by the applied bias voltage. It is noted that the emitter barrier must be thin enough to allow tunneling, and the collector barrier thick enough to minimize the leakage current. The common-base current gain α can be equal to 1 if losses due to the spread of energy, scattering in the base, and reflection from the collector barrier are prevented. The first HET based on this structure with a common emitter current gain of 1.3 at 40 K was demonstrated recently. The low current gain may be attributed to significant loss of carriers in the base, which can be minimized by reducing the base width or collector barrier height. Reducing the base width will, however, increase the base resistance. A new approach to solving the problems associated with low current gain in such a tunneling HET has been reported recently. This involves the use of a resonant tunneling double-barrier quantum-well structure in such a HET. For example, a resonant tunneling HET can be obtained if the AlGaAs injection barrier shown in Figure 16.15a is replaced by an AlGaAs (5 nm)/GaAs (5.6 nm)/AlGaAs (5 nm) double-barrier quantum-well structure between the emitter and the thin base of this HET.

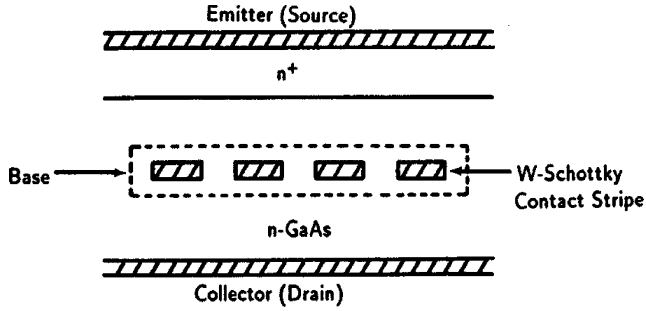


FIGURE 16.17. Cross-sectional view of a GaAs permeable base transistor (PBT). The grating tungsten/GaAs Schottky barrier contacts form the base region. After Bozler and Alley,[9,10] by permission.

The second HET structure shown in Figure 16.16a has the potential to produce a very small base resistance. In this structure, the GaAs base is undoped or lightly doped; the AlGaAs barrier between the emitter and base is triangular in shape and doped with donor impurities in Figure 16.16a and undoped in Figure 16.16b. A 2-DEG charge sheet is formed in the base by modulation doping in Figure 16.16a and by base collector voltage in Figure 16.16b, as in a HEMT. Since the base region is undoped, the carrier mobility of the 2-DEG is very high ($\approx$ 8000–9000 cm^2 (V s) at 300 K), and is much higher at 77 K due to the reduction in ionized impurity scattering. With such high electron mobility, the base will behave like a metal and the device will operate like a metal-base transistor (i.e., a permeable base transistor (PBT)). Since the base is undoped or lightly doped (to form 2-DEG), one expects the base resistance to be very high, which makes ohmic contact to the base very difficult. To overcome this problem, the base region has to be doped heavily so that the base resistance can be sharply reduced.

Another type of HET, which has been reported recently for high-speed and high-frequency applications, is called the permeable base transistor (PBT). Figure 16.17 shows a cross-sectional view of a GaAs PBT. The PBT is basically a vertical MESFET similar to a vacuum triode. The emitter and collector regions of the device are separated by a parallel array of metal stripes, which are connected to an external terminal of the base. The voltage applied to this terminal controls the current flow from the collector to the emitter terminal. In a GaAs PBT, the metal stripes are embedded in the GaAs by an epitaxial overgrowth process. The device structure consists of an n$^+$-GaAs substrate, an n-GaAs emitter, a thin tungsten grating Schottky contact on GaAs that forms the base, and an n-GaAs collector. The doping densities in the emitter and in the collector layers are adjusted so that the depletion region due to the tungsten-GaAs Schottky barrier extends across the openings in the grating. As an example, a tungsten grid consisting of a linewidth and spacing of 160 nm in a 30-nm-thick layer of tungsten have been used in such a structure. The flow of electrons from the emitter to the collector is only through the tungsten grating and is controlled by varying the tungsten base potential.

The advantages of a PBT for high-frequency and high-speed operation can be explained in terms of its transconductance g_m, output resistance R_0, base resistance R_B, base–emitter capacitance C_{BE}, and base–collector junction capacitance C_{BC}, which are due to the capacitive coupling across the depletion width surrounding the Schottky barrier base electrode. The unity current gain cutoff frequency f_T of the PBT is given by

$$f_T = \frac{g_m}{2\pi\,(C_{BE} + C_{BC})}. \tag{16.85}$$

The predicted value of f_T for a GaAs PBT is greater than 200 GHz. The maximum oscillation frequency f_{max} can be expressed by

$$f_{max} = \frac{f_T}{2\left(\dfrac{R_B + R_E}{R_0} + \dfrac{R_B g_m C_{BC}}{C_{BE} + C_{BC}}\right)^{1/2}}. \tag{16.86}$$

Based on (16.86), a value of f_{max} near 1000 GHz and a power-delay product of less than 1 fJ are predicted. A value of f_{max} around 100 GHz and a gain of 16 dB at 18 GHz have already been reported in the literature for a GaAs PBT.[9] Although HETs show promise for high-speed and high-frequency performance, many obstacles remain to be solved before practical HETs can be built for high-frequency and high-speed applications.

16.5. Resonant Tunneling Devices

Resonant tunneling through a double-barrier quantum-well structure (e.g., AlGaAs/GaAs/AlGaAs) was first reported by Chang et al.[11] in 1974. Subsequently, a variety of two- and three-terminal resonant tunneling devices (RTDs) have been reported. In general, RTDs can be implemented with few devices per function, and hence they have potential for high-speed applications with reduced circuit complexity because the intrinsic speed of a tunneling device is much faster than devices operating on a drift or diffusion process. Since carrier transport in a FET or HBT is limited either by the drift or diffusion process, devices such as RTDs operating on a tunneling process offer attractive advantages for high-speed applications. An RTD operates on the principle of quantum-mechanical tunneling through a multibarrier structure consisting of alternating layers of potential barriers (e.g., AlGaAs) and quantum wells (e.g., GaAs). In an RTD, the maximum tunneling current occurs when the injected carriers have certain resonant energies, which are determined by the Fermi energy in the doped cap region and the electron energy levels in the quantum wells. Energy band diagrams for a two-terminal double-barrier AlGaAs/GaAs/AlGaAs RTD structure under different bias conditions are shown in Figures 16.18a, b, and c, along with the resonance tunneling process. The current–voltage characteristic of a HET is shown in Figure 16.18d. Electrons originating in the conduction band of the doped GaAs are on the left-hand

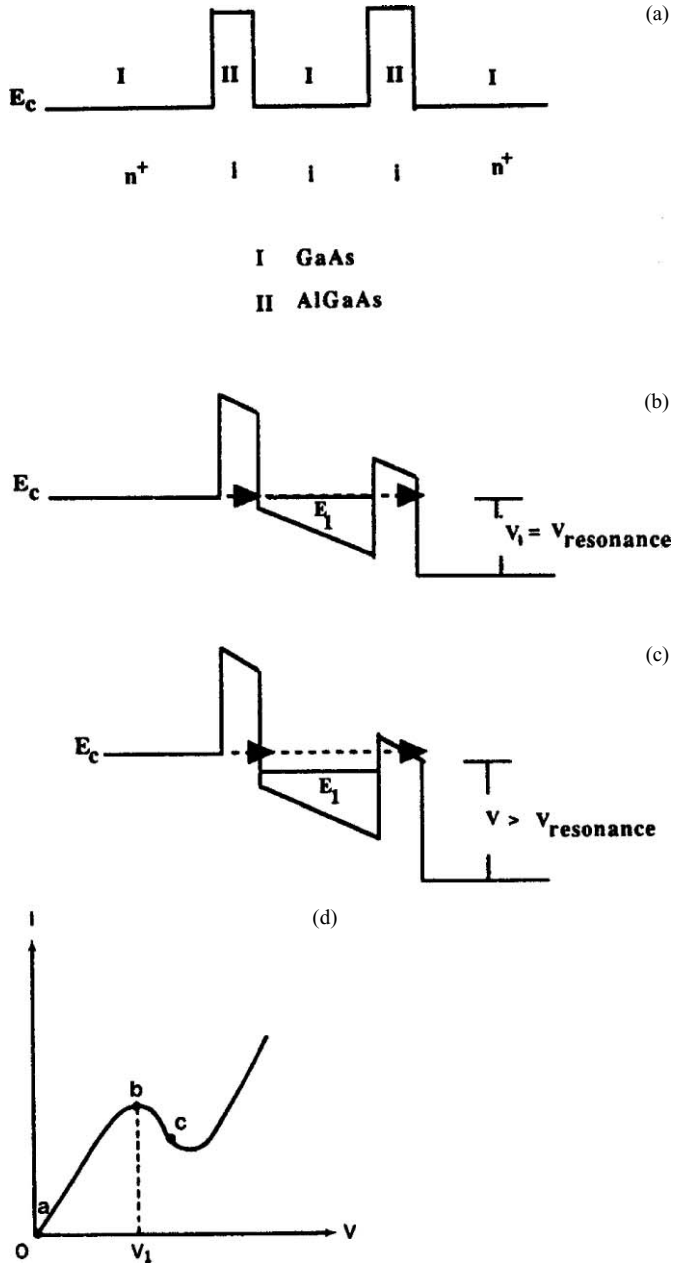


FIGURE 16.18. Energy band diagrams of a double-barrier (DB) AIGaAs/GaAs resonance tunneling device (RTD) structure along with the resonance tunneling process: (a) in equilibrium ($V = 0$), (b) under bias conditions with a peak current flow ($V_1 = E_1/q$), (c) with a reduced current flow ($V > E_1/q$), and (d) the current –voltage characteristic. After Capasso,[8] by permission, ©IEEE-1988.

side of the RTD, which is at the ground potential. These injected electrons tunnel through the AlGaAs barrier into the GaAs quantum well, and finally tunnel through the second barrier into the unoccupied states of the doped GaAs at the positive potential. Resonance occurs (i.e., the tunneling current reaches a maximum) when the energy of the electrons injected into the well becomes equal to the discrete quantum states in the well, as shown in Figure 16.18b. The tunneling current decreases rapidly when the discrete energy level in the well drops below the conduction band edge of the left-hand-side GaAs layer due to an increase in the applied bias voltage, as shown in Figure 16.18c. This leads to a negative differential resistance (NDR) in the *I–V* characteristics, as shown in Figure 16.18d. The solid circles shown in this figure correspond to different biases applied to the RTD. The NDR effect becomes more prominent at low temperatures, and hence it can be used for microwave generation and amplification. RTDs with a negative differential resistance and a peak-to-valley ratio exceeding 15 have been reported. Figures 16.19a, b, and c show the energy band diagrams of a three-terminal resonant tunneling transistor (RTT) with an emitter tunneling injection barrier and double-barrier quantum-well base under different bias conditions. Resonant tunneling occurs when the applied bias voltage is equal to the energy of the ground state in the quantum-well base layer, as shown in Figure 16.19b. The *I–V* characteristic curve similar to that of Figure 16.19d is expected for the RTT. A room temperature current gain of 7 (i.e., $\beta = \Delta I_C / \Delta I_B$) has been obtained for the AlGaAs/GaAs RTT shown in Figure 16.19. Other types of resonant tunneling transistors, including a graded emitter RTT with electrons ballistically launched into the base, an RTT with a parabolic quantum well in the base, and an RTT with a superlattice base, have also been reported with improved performance over the RTD.

Resonant tunneling devices (RTDs) are capable of achieving an intrinsic speed as high as 10^{-1} fs and an oscillation frequency of a few hundred GHz against a 100-GHz limit set by a Gunn device, Impatt diode, and Esaki diode. Recently, an AlGaAs/GaAs two-terminal negative differential resistance (NDR) RTD was successfully used to generate an oscillation frequency of 18 GHz with an output power of 5 μW at 200 K. Detecting and mixing studies at 2.5 THz demonstrated that the charge transport was faster than 0.1 ps. The RTD has also been used as an oscillator and in frequency-multiplier circuits.

Resonant tunneling transistors (RTTs) and circuit architectures with enhanced computational functionality are promising candidates for future nanoscale integration. A threshold logic full adder cell based on the RTT's multiple terminal linear threshold gates has been proposed recently for nanoscale integrated circuit applications. The threshold gate is composed of monolithically integrated resonant tunneling diodes and heterostructure field effect transistors. Together with a bit-level pipelining scheme, this leads to an efficient implementation with a minimal logic depth. The RTTs and linear threshold gates based on monostable–bistable logic transition elements (MOBILEs) are promising candidates for nanoscale integrated circuits. Recently, a design methodology of RTT logic gates and

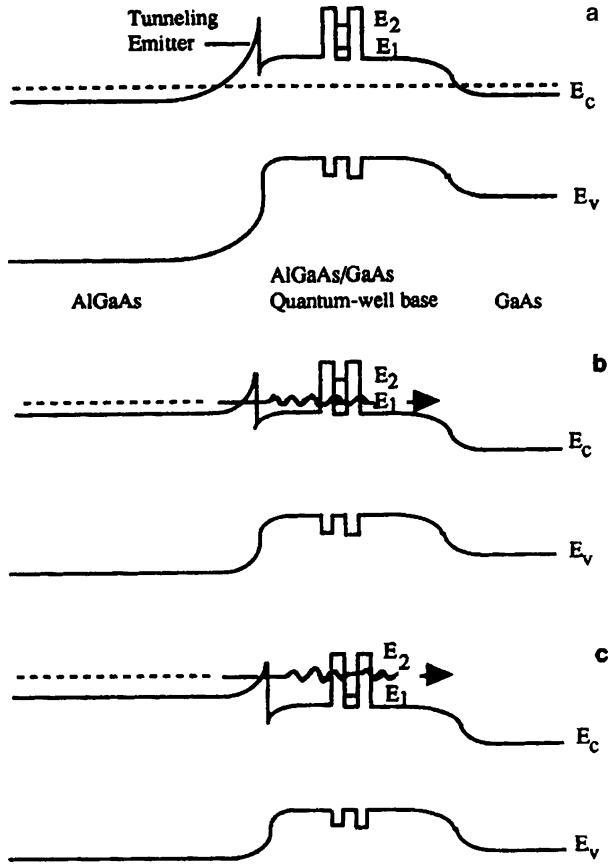


FIGURE 16.19. Energy band diagrams of a resonant tunneling transistor (RTT) with a tunneling emitter and double-barrier quantum-well base under different bias conditions: (a) $V = 0$, (b) $V = E_1/q$, and (c) $V = E_2/q$. After Capasso,[8] by permission, © IEEE-1988.

experimental results of a monolithically integrated NAND-NOR gate have been reported.

16.6. Transferred-Electron Devices

The transferred-electron device (TED) has been widely used as a local oscillator and power amplifier in the frequency range from 1 to 100 GHz. The TED, also known as the Gunn-effect diode, was first discovered by J. B. Gunn in 1963.[12] Gunn found that coherent microwave output was generated when a dc electric field was applied across a short n-type GaAs or InP sample with a critical field strength of a few thousand volts per cm. The oscillation frequency is approximately

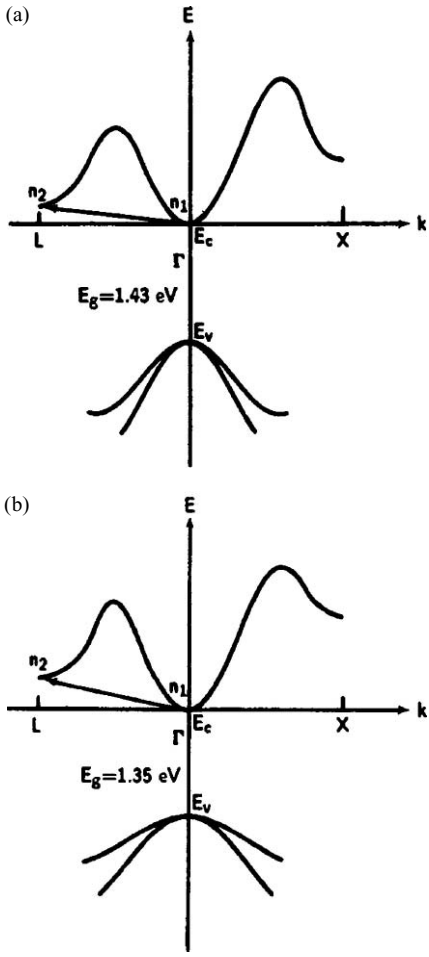


FIGURE 16.20. Energy band diagrams of (a) GaAs and (b) InP showing the transferred-electron (Gunn) effect from the central Γ valley to the L-satellite valleys along {111} axes. The energy separation Δ_{r-L} is equal to 0.31 eV for GaAs and 0.53 eV for InP. The populations in the Γ- and L-valleys are given by n_1 and n_2, respectively.

equal to the reciprocal of the carrier transit time across the length of the sample. The mechanism responsible for the negative differential resistivity (or mobility) is due to a field-induced transfer of electrons from the low-energy, high-mobility conduction band valley (i.e., the Γ -valley) to the higher-energy, lower-mobility satellite conduction valleys (L-conduction valleys), as first proposed by Ridley, Watkins, and Hilsum (RWH model).[13,14] Therefore, the transferred-electron effect has also been referred to as the Ridley–Watkins–Hilsum (RWH) effect. Several experiments performed on GaAs and $GaAs_{1-x}P_x$ samples revealed that the threshold field decreases with decreasing energy separation between the valley minima. The results provide convincing evidence that the transferred-electron effect is indeed responsible for the Gunn oscillation observed in GaAs and other III-V compound semiconductors.

To understand the physical mechanisms of the transferred-electron effect, which produces the negative differential resistance in a bulk semiconductor, let us consider the energy–momentum diagrams for the GaAs and InP crystals shown in Figure 16.20. The band structure consists of a low-energy, high-mobility central conduction band valley located at the Γ-point and several satellite valleys of higher energy and lower mobility located at L-points along the [111] axes of the first Brillouin zone. The energy separation ($\Delta E_{\Gamma-L}$) between the upper satellite valleys (L-valleys) and the lower conduction valley (Γ- valley) is $\Delta E_{\Gamma-L} = 0.31$ eV for GaAs and 0.53 eV for InP. If the densities of electrons in the upper and lower valleys are designated as n_2 and n_1, respectively, and the total carrier density is given by $n = n_1 + n_2$, then the steady-state current density of the bulk semiconductor is given by

$$J = q(\mu_1 n_1 + \mu_2 n_2)\mathcal{E} = qn v_{\mathrm{d}}, \tag{16.87}$$

where μ_1 and μ_2 denote the electron mobilities in the lower and upper conduction valleys, respectively, and v_{d} is the average drift velocity defined by

$$v_{\mathrm{d}} = \left(\frac{\mu_1 n_1 + \mu_2 n_2}{n_1 + n_2}\right)\mathcal{E} \approx \frac{\mu_1 \mathcal{E}}{1 + (n_2/n_1)}. \tag{16.88}$$

In (16.88) we have made use of the fact that $\mu_1 \gg \mu_2$. The population ratio n_2/n_1 between the upper and lower conduction valleys separated by the energy of ΔE_{21} is given by

$$\frac{n_2}{n_1} = R\, \exp\left(-\Delta E_{21}/k_{\mathrm{B}} T_{\mathrm{e}}\right), \tag{16.89}$$

where

$$R = \left(\frac{\nu_2}{\nu_1}\right)\left(\frac{m_2^*}{m_1^*}\right)^{3/2}$$

is the density-of-states ratio for the upper (L-band) and lower (Γ-band) conduction band valleys, ν_1 and ν_2 denote the number of lower and upper valleys, and m_1^* and m_2^* are the effective masses of electrons in the lower and upper valleys, respectively. For GaAs, $\nu_1 = 1$ and $\nu_2 = 4$, $m_1^* = 0.067 m_0$ and $m_2^* = 0.55\, m_0$, and $R = 94$.

The concept of energy relaxation time allows the electron temperature to be expressed in the form

$$T_{\mathrm{e}} = T + \frac{2q\mathcal{E} v_{\mathrm{d}} \tau_{\mathrm{e}}}{3k_{\mathrm{B}}}, \tag{16.90}$$

where τ_{e} is the energy relaxation time, which is on the order of 10^{-12} s. We note that the electron temperature T_{e} given by (16.90) is larger than the lattice temperature T, since the kinetic energy of an electron is increased by the accelerated electric field. Now substituting v_{d} from (16.88) and n_2/n_1 from (16.89) into

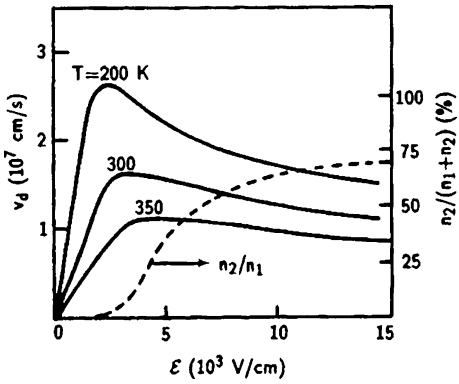


FIGURE 16.21. Drift velocity versus electric field curves for GaAs when $T =$ 200, 300, and 350 K; also shown is the population ratio n_2/n_1 versus electric field at 300 K. After Sze,[3] p. 647, with permission by John Wiley & Sons Co.

(16.90) yields

$$T_e = T + \left(\frac{2q\tau_e\mu_1}{3k_B}\right)\mathcal{E}^2\left[1 + R\ \exp\left(-\frac{\Delta E_{12}}{k_B T}\right)\right]^{-1}, \tag{16.91}$$

which shows that the electron temperature will increase with the square of the electric field above the critical field. For a given lattice temperature T, we can calculate T_e as a function of the electric field $\mathcal{E}$ using (16.91). The drift velocity versus electric field relation can also be derived from (16.88) and (16.89), which yields

$$v_d = \frac{\mu_1\mathcal{E}}{1 + R\ \exp(-\Delta E_{12}/k_B T)}. \tag{16.92}$$

Figure 16.21 shows the drift velocity versus electric field curves calculated from (16.92) for GaAs at three different lattice temperatures; also shown is the population ratio n_2/n_1 versus electric field at 300 K. It is of interest to note that at $\varepsilon = 15$ kV/cm approximately 70% of the total electron population is contributed by the upper satellite valleys. As shown in Figure 16.21, in the low-field regime, the velocity varies linearly with electric field (ohmic regime) and attains a peak value at critical field strength ($\mathcal{E}_c$). It then decreases with further increases in the electric field strength, corresponding to the negative differential resistance region (i.e., the NDR regime).

It is seen from the simple model presented above that (1) there is a well-defined threshold field $\mathcal{E}_c$ at the onset of negative differential resistivity or mobility, (2) the threshold field increases with increasing lattice temperature (see Figure 16.21), and (3) the NDR disappears if the lattice temperature is too high or the energy separation ΔE_{12} between the satellite and central valleys is too small. Therefore, in order to create an NDR effect via the electron transfer mechanism, the following conditions must be met: (1) the lattice temperature must be low enough such that in thermal equilibrium, most of the electrons reside in the lower conduction valley (i.e., the Γ-band), or $k_B T < \Delta E_{12}$; (2) the electron mobility μ_1 is much larger than μ_2 (i.e.,

$m_1 \ll m_2$), and the density of states for the upper valleys is much higher than for the lower valleys, and (3) $\Delta E_{12} \ll E_g$, so that avalanche breakdown does not occur before electrons are transferred into the upper satellite valleys by the applied field. Among semiconductors satisfying these conditions, n-type GaAs and InP crystals are the most widely studied materials for NDR devices. However, the transferred-electron effect has also been observed in many other compound semiconductors. Of particular interest are $Ga_xIn_{1-x}Sb$ ternary compounds, which have very low threshold fields and high electron velocities. For example, the critical field $\mathcal{E}_c$ for $Ga_{0.5}In_{0.5}Sb$ is only 600 V/cm and the peak velocity is $v_p = 2.5 \times 10^7$ cm/s.

Room-temperature experimental results show that the critical electric field, which defines the onset of NDR, is approximately 3.2 kV/cm for GaAs and 10.5 kV/cm for InP. The peak velocities are about 2.2×10^7 and 2.5×10^7 cm/s for high-purity GaAs and InP, respectively, while the maximum negative differential mobilities are found equal to -2400 cm^2/V·s for GaAs and -2000 cm^2/V·s for InP.

Fabrication of TEDs requires extremely pure and uniform materials with very low defect densities. Early TEDs were fabricated using bulk GaAs and InP materials with alloyed ohmic contacts. Modern TEDs are usually fabricated on epitaxial films grown on n^+ substrates using the VPE, MOCVD, or MBE technique. Typical donor densities are in the range of 10^{14} to 10^{16} cm^{-3}, and device lengths are in the range of a few microns to several hundred microns. Some high-power TEDs are made using selective metallization and mesa etching. To improve device performance, injection-limited cathode contacts have been used instead of the n^+ ohmic contacts. By using injection-limited contacts (e.g., Schottky barrier contact with low barrier height), the threshold field for the cathode current can be adjusted to a value approximately equal to the threshold field at the onset of NDR, resulting in uniform electric fields. For ohmic contacts, the accumulation or dipole layer grows some distance from the cathode, due to finite heating of the lower valley (Γ) electrons. This dead zone can be as large as 1 μm, which may limit the minimum device length and hence the maximum operating frequency. In an injection-limited contact, hot electrons are injected from the cathode, and hence the dead zone is reduced. Since transit time effects can be minimized, the device can exhibit a frequency-independent negative conductance shunted by its parallel-plate capacitance. If an inductance and a sufficiently large conductance are connected to the device, it can be expected to oscillate in a uniform-field mode at the resonance frequency.

Figures 16.22a, b, and c show the cross-section, dopant density profile, energy band diagram, and electric field distributions of three different cathode contacts of a Gunn device: (a) ohmic, (b) Schottky barrier, and (c) two-zone Schottky-barrier contacts. For the ohmic contact, there is always a low-field region near the cathode, and the field is nonuniform across the length of the device, as is clearly shown in Figure 16.22a. The Schottky barrier contact shown in Figure 16.22b consists of a low-barrier-height (0.15–0.3 eV) contact, which is generally very difficult to make in GaAs. The device in this case can be operated only in a very narrow temperature range due to its exponential temperature-dependent injection current.

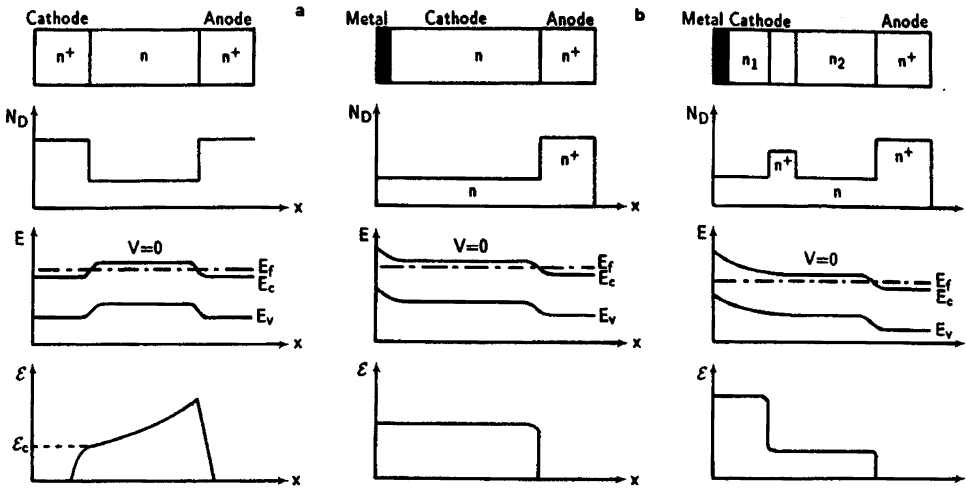


FIGURE 16.22. Doping profiles (N), energy band diagrams (E), and electric field distributions ($\mathcal{E}$) of a Gunn-effect device for three different cathode contacts: (a) ohmic, (b) Schottky barrier (low barrier), and (c) two-zone Schottky-barrier contact. After Sze,[3] p. 699, with permission by John Wiley & Sons.

The two-zone cathode contact shown in Figure 16.22c consists of a high-field zone and an n^+-zone. In this case, electrons are heated up in the high-field zone and subsequently injected into the active region with a uniform field. The structure in Figure 16.22c has been successfully used over a wide range of temperatures. The maximum efficiency obtained from an InP TED with a two-zone cathode contact is 24%. However, a GaAs TED with an injection-limited cathode contact has yet to be realized because of the Fermi level pinning effect.

Although GaAs- and InP-based Gunn diodes, based on the transferred-electron effect, have been successfully employed for microwave and millimeter-wave signal generation, the use of wide-band-gap material such as GaN with increased electrical strength offers the possibility to increase frequency and power capability of semiconductor devices. For example, GaN-based HEMTs with a record power density of 7 W/mm have been demonstrated recently. The high-frequency capability offered by GaN Gunn diodes is due to its higher electron velocity and reduced relaxation times in this material. Increasing electric field strength, which allows operation with higher doping densities and at higher bias, contributes to the high-power capability ($> 10^5$ W/cm^2) of the devices. Theoretical analysis reveals that GaN Gunn devices are expected to have twice the frequency and a hundred times the power capability of GaAs Gunn diodes. This makes the GaN Gunn diodes suitable for THz signal generation by means of multiplication of signal generated in nitride semiconductor diodes or harmonic generation at much higher levels than any other currently available semiconductor devices.

Problems

16.1. Calculate the necessary thickness of an active n-channel layer with dopant density of 2×10^{17} cm^{-3} grown on a semi-insulating GaAs substrate for a GaAs MESFET with a threshold voltage of -2 V. Assume that the surface states pin the Fermi level at $2/3E_g$ below the conduction band edge at the metal-GaAs interface under the Schottky contact gate.

16.2. The lateral electric field in the channel of a GaAs MESFET under saturation conditions is usually high enough for electrons to drift with their scattering limited velocity v_s.

(a) Show that the drain saturation current in this case is given by

$$I_{Dsat} = \frac{1}{3} G_0 V_p = \frac{2}{3} Z_q b N_D v_s,$$

where b is the undepleted active layer thickness at $V_{GS} = 0$.

(b) Calculate b for $N_D = 1.5 \times 10^{17}$ cm^{-3} and $I_{Dsat} = 250$ mA/mm gate width (Z), assuming $v_s = 1.1 \times 10^7$ cm/s.

16.3. (a) Derive an expression for the threshold voltage V_T in terms of channel dopant density N_D, channel height a, and built-in potential V_{bi}, for a GaAs MESFET under pinch-off conditions.

(b) If the channel of a GaAs MESFET is uniformly doped to 2×10^{17} cm^{-3}, and the built-in potential of the Ti-Pt-Au Schottky barrier gate contact on n-GaAs is 0.8 V, find the channel thickness required to obtain a threshold voltage of -1 V.

16.4. Using a three-piecewise linear relation between drift velocity and electric field, derive an expression for the current–voltage characteristics of a HEMT device (see the paper[5] by K. Lee et al.)

16.5. For a short-channel FET (i.e., gate length $L \leq 1\mu$m), a semi-empirical formula may be used to define the effective saturation velocity v_s in the channel, which is given by

$$v_s = 59 L^{-0.56} \text{ m/s},$$

where L is in meters.

(a) Calculate v_s for $L = 0.25$, 0.50, and 1.0 μm.

(b) Calculate the unity current gain cutoff frequency f_T ($= v_s/2\pi L$) for the gate lengths and saturation velocities given in (a).

16.6. (a) Show that the transconductance g_m of a uniformly doped n-channel MESFET is given by

$$g_m = \frac{\varepsilon_0 \varepsilon_s v_{sat} Z}{W_d}.$$

(b) Show that the expression for g_m for a nonuniform doping profile given by $N_D(y) = Ky$ for $y < W_d$ is the same as for the uniform doping profile given by (a).

(c) Derive an expression for the threshold voltage V_T.

(d) Calculate g_m from (a) if $\varepsilon_0\varepsilon_s = 1 \times 1^{-12}$ F/cm, $v_{sat} = 1.2 \times 10^7$ cm/s, $Z = 50\,\mu$m, and $W_d = 0.1\,\mu$m.

16.7. (a) Draw a small-signal equivalent circuit diagram for a MESFET showing both the extrinsic and intrinsic circuit elements.

(b) Show that f_T (cutoff frequency for unity current gain with output of FET shorted) of a MESFET can be expressed by

$$f_T = \frac{v_{sat}}{2\pi L},$$

where L is the gate length, and v_{sat} is the saturation velocity of electrons.

(c) Calculate f_T for $L = 1.0, 0.5, 0.1\,\mu$m and $v_{sat} = 1.2 \times 10^7$ cm/sec for a GaAs MESFET.

(d) Derive an expression for the extrinsic transconductance g_{em} and drain conductance g_{ds} of a MESFET showing the effect of the source resistance R_s.

16.8. (a) Discuss the second-order effects on the performance of a MESFET (e.g., backgating, drain current lag, temperature, subthreshold current, etc.).

(b) Draw a typical self-aligned process sequence of the GaAs MESFET fabrication steps.

(c) Explain why p-channel MESFETs cannot be achieved in GaAs.

16.9. (a) Construct the energy band diagram for a depletion mode (D-) AlGaAs/GaAs modulation doped FET (HEMT) including the spacer layer, and explain how two-dimensional electron gas (2-DEG) is formed in the undoped GaAs layer.

(b) Explain how high electron mobility is achieved in the 2-DEG GaAs layer.

(c) Explain why the transconductance g_m and cutoff frequency f_T for a HEMT can be much higher than those of MESFETs.

(d) Note that both the depletion (D-) and enhancement (E-) mode GaAs HEMTs can be made, but D-HEMTs are more common. Plot the energy band diagram for an AlGaAs/GaAs E-HEMT when $V_{GS} = 0$ and $V_{GS} > V_T$.

16.10. Consider a GaAs MESFET with device parameters given by channel length $L = 3\,\mu$m, channel height $a = 1\,\mu$m, $N_D = 2.6 \times 10^{15}$ cm^{-3}, and with $V_{GS} = -1$ V and $V_{DS} = 3$ V applied to the device.

(a) Draw the cross-sectional view of this MESFET showing the depletion region, the channel, and the Gunn domain region.

(b) Plot (i) electric field versus x, (ii) drift velocity versus x, and (iii) space-charge versus x in the channel.

References

1. C. A. Liechti, *IEEE Trans. Microwave Theory Tech.* **MTT-24**, 286 (1976).
2. R. Pucel, H. Haus, and H. Statz, *Advances in Electronics and Electron Physics*, Vol. 38, p. 195, Academic Press, New York (1975).

3. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1981).
4. D. Delagebeaudeuf and N. T. Ling, "Metal-n-AlGaAs/GaAs Two-Dimensional Electron Gas FET," *IEEE Trans. Electron Devices* **ED-29**, 955 (1982).
5. K. Lee, M. S. Shur, T. J. Drummond, and H. Morkoc, "Current–Voltage and Capacitance–Voltage Characteristics of Modulation-doped Field Effect Transistors," *IEEE Trans. Electron Devices* **ED-30**, 207 (1983).
6. S. M. Sze, *High Speed Devices*, Wiley, New York (1991).
7. C. A. Mead, *Proc. IRE* **48**, 359 (1960).
8. F. Capasso, in: *High-Speed Electronics* (B. Kallback and H. Beneking, eds.), Vol. 22, pp. 50–61, Springer-Verlag, Berlin (1986).
9. C. O. Bozler and G. D. Alley, "Fabrication and Numerical Simulation of the Permeable Base Transistor," *IEEE Trans. Electron Devices* **ED-27**, 1128 (1980).
10. C. O. Bozler and G. D. Alley, "The Permeable Base Tansistor and Its Application to Logic Circuits," *Proc. IEEE* **70**, 46 (1982).
11. L. L. Chang, L. Esaki, and R. Tsu, *J. Appl. Phys.* **24**, 593 (1974).
12. J. B. Gunn, "Microwave Oscillations of Current in III-V Semiconductors," *Solid State Commun.* **1**, 88 (1963).
13. C. Hilsum, "Transferred Electron Amplifiers and Oscillators," *Proc. IRE* **50**, 185 (1962).

Bibliography

E. Alekseev and D. Pavlidis, " DC and High-Frequency Performance of AlGaN/GaN HBTs," *Solid State Electronics*, vol. 44 (2), pp. 245–252 (2000).
J. S. Blackmore, "Electron and Hole Traps in GaAs," *J. Appl. Phys.* **53**, R123 (1982).
I. B. Bott and W. Fawcett, "The Gunn Effect in GaAs," in *Advances in Microwaves* (L. Yung, ed.), Vol. 3, pp. 223–300, Academic Press, New York (1968).
P. J. Bulman, G. S. Hobson, B. C. Taylor, "Transferred Electron Devices," Academic Press, London and New York (1972).
F. Capasso, in: *Semiconductors and Semimetals* (R. K. Willardson and A. C. Beer, eds.) Vol. 22, Part D, p. 2, Academic Press, New York (1985).
F. Capasso, J. Allam, A. Y. Cho, K. Mohammed, R. J. Malik, A. L. Hutchinson, and D. Sivco, *Appl. Phys. Lett.* **48**, 1294 (1986).
C. Y. Chang and Francis Kai, *GaAs High Speed Devices*, Wiley Interscience, New York (1994).
C. Y. Chang and S. M. Sze, *ULSI Devices*, Wiley Interscience, New York (2000).
T. H. Chen and M. S. Shur, "Capacitance Model of GaAs MESFETs," *IEEE Trans. Electron Devices* **ED-32**, 883 (1985).
R. Dingle, H. L. Stormer, A. C. Gossard, and W. Wiegmann, *Appl. Phys. Lett.* **37**, 805 (1978).
T. J. Drummond, H. Morkoc, K. Lee, and M. S. Shur, "Model for Modulation Doped Field Effect Transistor," *IEEE Electron Device Lett.* **EDL-3**, 338 (1981).
T. J. Drummond, W. Kopp, M. Keever, H. Morkoc, and A. Y. Cho, "Electron Mobility in Single and Multiple Period Modulation-Doped AlGaAs/GaAs Heterostructures," *J. Appl. Phys.* **23**, 230 (1984).
W. P. Dumke, J. M. Woodall, and V. L. Rideout, "GaAs–GaAlAs Heterojunction Transistor for High Frequency Operation," *Solid-State Electron.* **15**, 1339 (1972).

L. F. Eastman, "Very High Electron Velocity in Short GaAs Structures," in: *Advances in Solid State Physics* 12 (J. Treush, ed.), p. 173, Vieweg, Braunschweig (1982).
A. A. Grinberg and M. S. Shur, "Density of Two-dimensional Electron Gas in Modulation-doped Structure with Graded Interface," *Appl. Phys. Lett.* **45**, 573 (1984).
M. Hirano, K. Oe, and F. Yanagawa, "High-Transconductance p-Channel Modulation-Doped AlGaAs/GaAs Heterostructure FETs," *IEEE Trans. Electron Devices* **ED-33**, 620 (1986).
M. A. Hollis, S. C. Palmateer, L. F. Eastman, N. V. Dandekar, and P. M. Smith, *IEEE Electron Device Lett.* **EDL-4**, 440 (1983).
S Hutchinson, J Stephens, M Carr and M J Kelly, "Implant isolation scheme for current confinement in graded-gap Gunn diodes" *Electronics Letters* **32** 851–2 (1996)
H. Kroemer, "Theory of Wide-gap Emitter Transistors," *Proc. IRE* **45**, 1535 (1957).
H. Kroemer, "Theory of the Gunn Effect," *Proc. IEEE* **52**, 1736 (1964).
H. Kroemer, "The Gunn Effect Under Imperfect Cathode Boundary Conditions," *IEEE Trans. Electron Devices* **ED-15**, 819 (1968).
H. Kroemer, "Heterostructure Bipolar Transistors and Integrated Circuits," *Proc. IEEE* **70**, 13 (1982).
K. Lehovec and R. Zuleeg, "Voltage–Current Characteristics of GaAs-JFETs in the Hot Electron Range," *Solid-State Electron.* **13**, 1415 (1970).
S. Luryi, *Appl. Phys. Lett.* **47**, 490 (1985).
A. G. Milnes and D. L. Feucht, *Heterojunctions and Metal–Semiconductor Junctions*, Academic Press, New York (1972).
T. Mimura, S. Hiyamizu, T. Fijii, and K. Nanbu, "A New Field Effect Transistor with Selectively Doped GaAs/n-AlGaAs Heterojunctions," *Jpn. J. Appl. Phys.* **19**, L225 (1980).
H. Morkoc, J. Chen, U. K. Reddy, T. Henderson, P. D. Coleman, and S. Luryi, *Appl. Phys. Lett.* **42**, 70 (1986).
D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*, 3rd edition, McGrew Hill, New York, (2003).
C. Pacha, P. Glösekötter, K. Goser, U. Auer, W. Prost, F.-J. Tegude, "Resonant Tunneling Transistors for Threshold Logic Circuit Applications," 9*th Great Lakes Symposium on VLSI*, p. 344 (1999).
C. Pacha and K. Goser. "Design of Arithmetic Circuits Using Resonant Tunneling Diodes and Threshold Logic," *Proceedings of the 2nd Workshop on Innovative Circuits and Systems for Nanoelectronics*, pages 83–93. TU Delft, NL, September 1997.
K. Park and K. D. Kwack, "A Model for the Current–Voltage Characteristics of MODFETs," *IEEE Trans. Electron Devices* **ED-33**, 673 (1986).
B. K. Ridley, "Specific Negative Resistance in Solids," *Proc. Phys. Soc.* **82**, 954 (1963).
B. K. Ridley and T. B. Watkins, "The Possibility of Negative Resistance," *Proc. Phys. Soc.* **78**, 291 (1961).
T. G. Ruttan, "High Frequency Gunn Oscillators," *IEEE Trans. on MTT*, pp. 142–144 (1974).
L. P. Sadwick and K. L. Wang, "A Treatise on the Capacitance–Voltage Relation of High Electron Mobility Transistors," *IEEE Trans. Electron Devices* **ED-33**, 651 (1986).
E. F. Schubert and A. Fischer, "The Delta-Doped Field-Effect Tansistor (δFET)," *IEEE Trans. Electron Devices* **ED-33**, 625 (1986).
B. L. Sharma and R. K. Purohit, *Semiconductor Heterojunctions*, Pergamon, London (1974).
W. Shockley, *Bell Syst. Tech. J.* **30**, 990 (1951).
W. Shockley, "A Unipolar Field-effect Transistor," *Proc. IRE* **40**, 1365 (1952).
M. S. Shur, "Analytical Model of GaAs MESFETs," *IEEE Trans. Electron Devices* **ED-25**, 612 (1978).

M. S. Shur, "Analytical Models of GaAs MESFETs," *IEEE Trans. Electron Devices* **ED-32**, 18 (1985).

M. S. Shur, *GaAs Devices and Circuits*, Plenum Press, New York (1987).

M. S. Shur and L. F. Eastman, "Current–Voltage Characteristics, Small-Signal Parameters and Switching Times of GaAs FETs," *IEEE Trans. Electron Devices* **ED-25**, 606 (1978).

M. S. Shur and L. F. Eastman, "A Near Ballistic Electron Transport in GaAs Devices at 77 K," *Solid-State Electron.* **24**, 11 (1981).

P. M. Solomon and H. Morkoc, *IEEE Trans. Electron Devices* **ED-31**, 1015 (1984).

H Spooner and N R Crouch "Advances in hot electron injector Gunn diodes" *GEC Journal of Research* **7**, 34–45 (1990).

S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York (1981).

S. M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd edition, Wiley Interscience, New York (2002).

G. W. Taylor, H. M. Darley, R. C. Frye, and P. K. Chatterjee, "A Device Model for an Ion Implanted MESFET," *IEEE Trans. Electron Devices* **ED-26**, 172 (1979).

R. Tsu and L. Esaki, *Appl. Phys. Lett.* **22**, 562 (1973).

H. Unlu and A. Nussbaum, "Band Discontinuities at Heterojunction Device Design Parameters," *IEEE Trans. Electron Devices* **ED-33**, 616 (1986).

T. Wada and S. Frey, "Physical Basis of Short-Channel MESFET Operator," *IEEE Trans. Electron Devices* **ED-26**, 476 (1979).

G. W. Wang and W. H. Ku, "An Analytical and Computer-aided Model of the AlGaAs/GaAs High Electron Mobility Transistor," *IEEE Trans. Electron Devices* **ED-33**, 657 (1986).

Solutions to Selected Problems

Chapter 1

1.1. (a) Simple cubic; 1 atom per unit cell: $\frac{4}{3}\pi\frac{a^2/2}{a^3} = \frac{\pi}{6}$.

(b) BCC: 2 atom per unit cell; $2 \times \frac{4}{3}\pi\frac{\sqrt{3}a^2/4}{a^3} = \frac{\sqrt{3}\pi}{8}$.

(c) FCC: 4 atoms per unit cell; $4 \times \frac{4}{3}\pi\frac{\sqrt{2}a^3/4}{a^3} = \frac{\sqrt{2}\pi}{6}$.

(d) HCP: 6 atoms per unit cell; $6 \times \frac{4}{3}\pi\frac{a^3/2}{6\frac{\sqrt{3}}{2}a\frac{a}{2}\sqrt{\frac{8}{3}}a} = \frac{\sqrt{2}\pi}{6}$.

(e) Diamond: 8 atom per unit cell; $8 \times \frac{4}{3}\pi\frac{\sqrt{3}a^3/8}{a^3} = \frac{\sqrt{3}\pi}{16}$.

1.3. A fivefold axis of symmetry cannot exist in a lattice because it is impossible to fill all space with a connected array of pentagons.

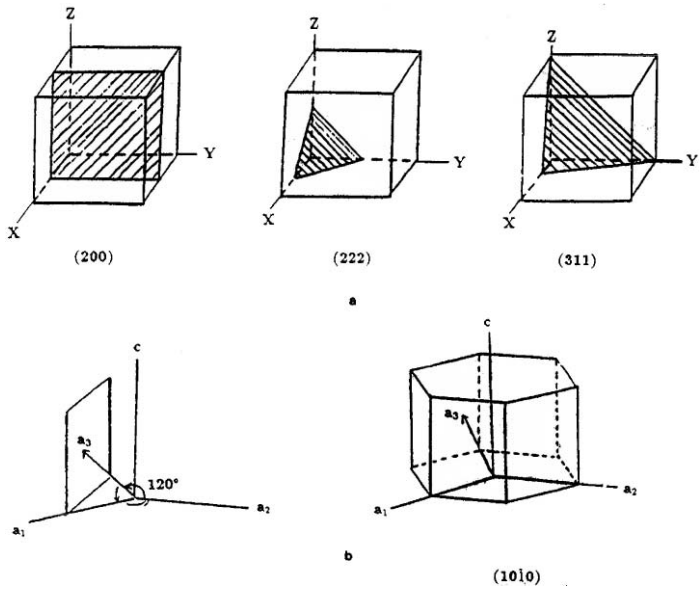
1.5. (a) A unit cell of the diamond lattice is constructed by placing atoms $\frac{a}{4}, \frac{a}{4}, \frac{a}{4}$ from each atom in an fcc.

(b) Total number of atoms per unit cell = 8; the distance is 1.54 Å; and 2.43 Å and 2.35 Å for Ge and Si, respectively.

1.7. (a) 4.

(b) The basis of the diamond structure consists of two atoms at $\left(000, \frac{a}{4}, \frac{a}{4}, \frac{a}{4}\right)$. Therefore, the primitive vectors of the diamond structure are the same as those of the fcc.

1.9.

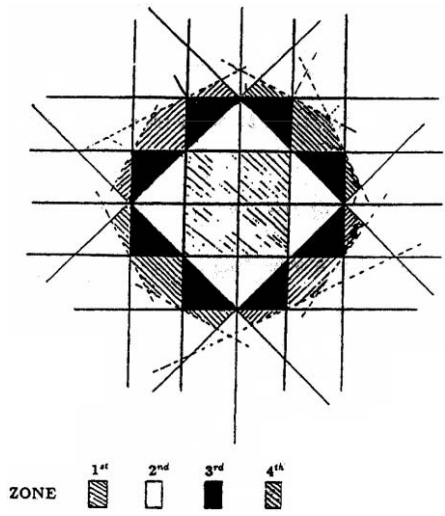


1.11. (a) {100}, {200}: both have six planes.
{110}, {220}: both have twelve planes.
{111}: eight planes.

(b) The normal distances are

$$(100) : a; (110) : \sqrt{2}a/2; (111) : \sqrt{3}a/3; (200) : 0.5a; (220) : \sqrt{2a}/4.$$

1.13.



Chapter 2

2.1. (a) Let M be the mass of the atom and C the force constant. The equations for this case are given as

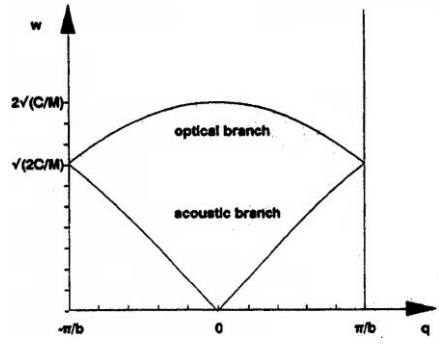
$$M\frac{\mathrm{d}^2X_{n,1}}{\mathrm{d}t^2} = -C(X_{n,1} - X_{n,2}) - C(X_{n,1} - X_{n-1,2}) \tag{1}$$

and

$$M\frac{\mathrm{d}^2X_{n,2}}{\mathrm{d}t^2} = -C(X_{n,2} - X_{n,1}) - C(X_{n,2} - X_{n+1,1}). \tag{2}$$

The harmonic solutions are $X_{n,1} = A\exp(iqna - i\omega t)$ and $X_{n,2} = B\exp[iq(n + \frac{1}{4})a - i\omega t]$. Now substituting $X_{n,1}$, $X_{n,2}$ given above into (1) and (2) and solving the two simultaneous linear equations for ω yields $\omega^2 = 2(C/M)[1 \pm \cos(qa/2)]$.

(b) At the zone boundary, i.e., $q = \pm\pi/b$, $\omega = \sqrt{2C/M}$ and $q = 0$, $\omega_0 = 2\sqrt{C/M}$ (optical phonon frequency). The dispersion curves are shown in the figure below.



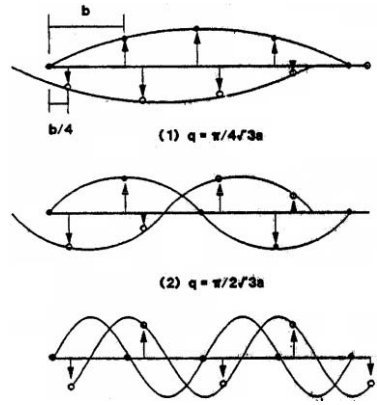
(c) From (a), one can find that the ratio of the amplitudes, B/A, is given by

$$\frac{B}{A} = -\exp\left(\frac{iqb}{4}\right).$$

Thus, for $t = 0$, the displacements for the $(n, 1)$ and $(n, 2)$ atoms are given by

$$X_{n,1} = A\ \exp(inqb) \quad \text{and} \quad X_{n,2} = -A\ \exp\left[i\left(n + \frac{1}{2}\right)qb\right].$$

The plot of the atomic displacements for the TO modes are shown in the figures below.



2.3. (a) According to (2.5), $\omega = \omega_m \sin(qa/2)$ and $\mathrm{d}\omega = \omega_m(a/2) \times \cos(qa/2)\mathrm{d}q$. Therefore,

$$D(\omega) = w(q)\left(\frac{\mathrm{d}q}{\mathrm{d}\omega}\right) = \frac{L}{\pi}\left(\frac{1}{\mathrm{d}\omega/\mathrm{d}q}\right), \quad D(\omega) = \frac{2L}{\pi a}(\omega_m^2 - \omega^2)^{-1/2},$$

and $v_\mathrm{g} = \dfrac{a}{2}\sqrt{\omega_m^2 - \omega^2}$.

(b) $D(\omega)\,\mathrm{d}\omega = \dfrac{L}{\pi}\dfrac{\mathrm{d}\omega}{\mathrm{d}\omega/\mathrm{d}q} = \dfrac{L\mathrm{d}\omega}{\pi v_g}$. Therefore, $D(\omega) = \dfrac{L}{\pi v_\mathrm{s}}$

2.5. (a) Similar to Problem 2.1 except for different spring constants,

$$M\frac{\mathrm{d}^2X_{2n}}{\mathrm{d}T^2} = K_1(X_{2n-1} - X_{2n}) + K_2(X_{2n+1} - X_{2n})$$

and

$$M\frac{\mathrm{d}^2X_{2n+1}}{\mathrm{d}t^2} = K_2(X_{2n} - X_{2n+1}) + K_1(X_{2n+2} - X_{2n+1}).$$

Solving these two equations, one obtains

$$\omega^2 = \left(\frac{K_1 + K_2}{M}\right) \pm \frac{\sqrt{K_1^2 + K_2^2 + 2K_1K_2\cos qa}}{M}.$$

2.7. (a) The fixed-boundary condition gives a standing wave solution $U_n = A\exp(-j\omega t)\sin(naq)$. At $n = 0, U_0 = 0$ and at $n = N, U_N = 0$. Then, $Naq = l\pi, l = 1, 2, \ldots, (N-1), q = \pi/Na, 2\pi/Na, \ldots, (N-1)\pi/Na$.
There are $(N-1)$ allowed independent values of q; the density of states in q-space equals L/π for $q \le \pi/a$ and 0 for $q > \pi/a$.

(b) The periodic boundary condition gives a running wave solution $U_{na} = U_{na+L}$ with $\exp(iLq) = 1$, $Lq = 2s\pi$,where $s = 0, \pm1, \pm2, \ldots$ and

$q = 0, \pm 2\pi/L, \pm 4\pi/L, \ldots, \pm N\pi/L$; the density of states in q-space is equal to $L/2\pi$ for $-\pi/a \leq q \leq \pi/a$, zero otherwise.

(c) See (a) and (b).

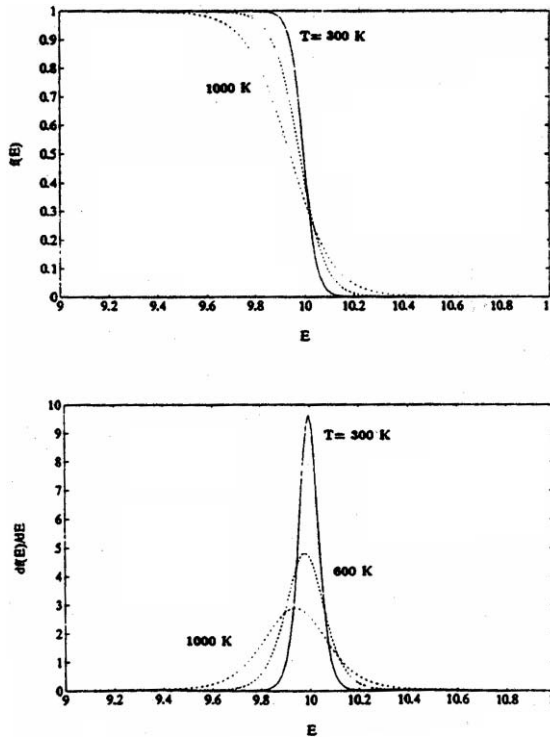
Chapter 3

3.1.
$$\langle v \rangle = \frac{\int_0^\infty vN(v)\mathrm{d}v}{\int_0^\infty N(v)\mathrm{d}v} = \frac{\int_0^\infty v^3 \exp\left(-\frac{mv^2}{2k_\mathrm{B}T}\right)\mathrm{d}v}{\int_0^\infty v^2 \exp\left(-\frac{mv^2}{2k_\mathrm{B}T}\right)\mathrm{d}v} = \sqrt{\frac{8k_\mathrm{B}T}{\pi m}}.$$

3.3. (a) According to Fermi–Dirac statistics,
$$\langle E \rangle = \frac{\int_0^\infty E\mathrm{g}(E)f(E)\mathrm{d}E}{\int_0^\infty \mathrm{g}(E)f(E)\,\mathrm{d}E} = \frac{\int_0^\infty CE^{3/2}f(E)\mathrm{d}E}{\int_0^\infty CE^{1/2}f(E)\,\mathrm{d}E}$$
assuming $g(E) = CE^{1/2}$.
Since $f(E) = 1$ at $T = 0$ K, therefore $\langle E \rangle = (\frac{2}{5}E_\mathrm{f}^2(0))/(\frac{2}{3}E_f(0)) = \frac{3}{5}E_\mathrm{f}(0)$.
For B-M statistics: $\langle E \rangle = \frac{1}{2}m\langle v \rangle^2 = \frac{3}{2}k_\mathrm{B}T$.

(b) The difference between Problems 3.2 and 3.3 at $T = 0$ K is that F-D statistics consider the Pauli exclusion principle and M-B statstics do not.

3.5.



3.7. (a) $g(E_{\rm f}(0)) = \dfrac{4\pi}{\hbar^3}(2m_0)^{3/2}\sqrt{E_{\rm f}(0)} = \dfrac{4\pi}{\hbar^3}(2m_0)^{3/2}\sqrt{\dfrac{\hbar^2}{8m_0}}\left(\dfrac{3n_0}{\pi}\right)^{1/3}$

$$= \frac{3n_0}{2E_{\rm f}(0)} = \frac{3n_0}{2k_{\rm B}T_{\rm f}},$$

since $U = U_0 + \dfrac{\pi^2}{6}(k_{\rm B}T)^2\dfrac{3n_0}{2k_{\rm B}T_{\rm f}'}\dfrac{\partial U}{\partial T} = \dfrac{\pi^2 k_{\rm B}Tn_0}{2T_f}$.

Chapter 4

4.1. According to (4.56),

$$E_k = E_k^0 + \Sigma_{k_j}\frac{H_{kk'}^2}{\left(E_k^0 - E_j^0\right)}, \quad \text{and} \quad H_{kk'} = v\left(\frac{\pi}{a}\right) + v\left(-\frac{\pi}{a}\right).$$

Therefore,

$$\begin{aligned}
E_k &= E_k^0 + \frac{[v(\pi/a)]^2}{\left(E_k^0 - E_{k'}^0\right)} + \frac{[v(-\pi/a)]^2}{\left(E_k^0 - E_k^0\right)} \\
&= E_k^0 + \frac{2m_0[v(\pi/a)]^2}{\hbar^2}\left[\frac{1}{k^2 - (k-\pi/a)^2} + \frac{1}{k^2 - (k+\pi/a)^2)}\right] \\
&\approx E_k^0 - \frac{4m_0[v(\pi/a)]^2a^2}{(\hbar\pi)^2} - \frac{16m_0[v(\pi/a)]^2a^4k^2}{\hbar^2\pi^4} \\
&= \frac{(\hbar k)^2}{2m_0}\left[1 - \frac{32m_0^2[v(\pi/a)]^2a^4}{\hbar^4\pi^4}\right] - \frac{4m_0[v(\pi/a)]^2a^2}{(\hbar\pi)^2} \\
&= E_0 + \frac{(\hbar k)^2}{2m^*},
\end{aligned}$$

where

$$E_0 = -\frac{4m_0[v(\pi/a)]^2a^2}{(\hbar\pi)^2} \quad \text{and} \quad m^* = \frac{m_0}{1 - \dfrac{32m_0^2[v(\pi/a)]^2a^4}{\hbar^4\pi^4}}.$$

4.3. Si : $m_{\rm cn}^* = 0.26m_0; m_{\rm dn}^* = 1.08m_0$.
Ge: $m_{\rm cn}^* = 0.12m_0; m_{\rm dn}^* = 0.56m_0$.

4.5. $E(k) = E_0 + B\cos\left(\dfrac{k_xa}{2}\right)\cos\left(\dfrac{k_ya}{2}\right)$; for $k_x, k_y \to 0$, $E(k) = E_0 + B \times \left[1 - \left(\dfrac{k_xa}{2}\right)^2\right]\left[1 - \left(\dfrac{k_ya}{2}\right)^2\right]$,

where $\left[1 - \left(\dfrac{k_xa}{2}\right)^2\right]\left[1 - \left(\dfrac{k_ya}{2}\right)^2\right] = \text{const}$. Therefore it is the equation of a circle.

4.7. Similar to Problem 4.6.

(a) $v_{\mathrm{g}} = \dfrac{2a\beta_n}{\hbar}[\sin(k_x a)\bar{x} + \sin(k_y a)\bar{y} + \sin(k_z a)\bar{z}]$.

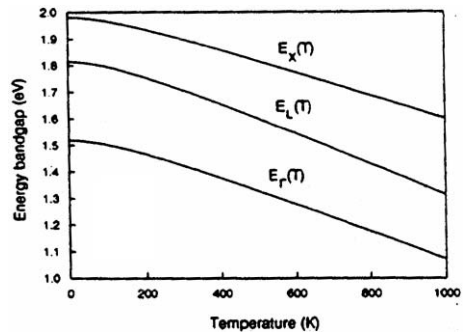
(b) $a = \mathrm{d}v_{\mathrm{g}}/\mathrm{d}t = F/m^*$, where m^* is in (c).

(c)
$$m_y^* = \begin{cases} 0, & \text{where } i \neq j, \\ \dfrac{\hbar}{2a^2\beta_n}\dfrac{1}{\cos(k_1 a)}, & i = j = 1, \\ \dfrac{\hbar}{2a^2\beta_n}\dfrac{1}{\cos(k_2 a)}, & i = j = 2, \\ \dfrac{\hbar}{2a^2\beta_n}\dfrac{1}{\cos(k_3 a)}, & i = j = 3. \end{cases}$$

4.9. For an fcc lattice, there will be twelve nearest neighbors, i.e., $R_{ij} = \{\pm a/2, \pm a/2, \pm a/2\}$. Therefore,

$$E_k = E_{\mathrm{n0}} - A_{\mathrm{n}} - 4\beta_n \left[\cos\left(\frac{k_x a}{2}\right)\cos\left(\frac{k_y a}{2}\right) + \cos\left(\frac{k_y a}{2}\right)\cos\left(\frac{k_z a}{2}\right) + \cos\left(\frac{k_x a}{2}\right)\cos\left(\frac{k_z a}{2}\right)\right].$$

4.11.



4.13. (a) A deep core potential can be neglected and a simple plane wave basis will yield rapid convergence in a pseudopotential calculation.

(b) A pseudopotential is dependent not only on the energy eigenvalues, but also on the angular-momentum components present in the core.

(c) A nonlocal correction to the local atomic potential term is adopted in a nonlocal pseudopotential calculation.

(d) Once the nonlocal pseudopotential is determined, the eigenvalues and eigenvectors can be found and calculation of the energy band spectrum is straightforward.

Chapter 5

5.1. (a) $T = 300$ K, exhaustion regime $n_0 = N_D = \sigma/\mu_n q = 4.6 \times 10^{14}$ cm^{-3}.

(b) $n_0 = N_D - N_A = N_c \exp\left(-\dfrac{E_c - E_F}{k_B T}\right)$; $E_F - E_c = -0.285$ eV.

(c) $E_F - E_c = -0.293$ eV.

(d) $T = 20$ K, according to (5.39). $n_0 = 3 \times 10^{10}$ cm^{-3}.

(e) $T = 77$ K, $N_A = 0$ and $p_0 - n_0 + N_D - n_D = 0$,
n-type $P_0 \ll n_0 \rightarrow n_0 = N_D - n_D$ and

$$n_0 = \frac{N_D}{1 + g \exp\left(\dfrac{E_D - E_f}{k_B T}\right)}, \rightarrow n_0 = 3.96 \times 10^{14} \text{ cm}^{-3} < N_D.$$

5.3. According to (5.49) and (5.50), $E_i = -0.0081$ eV and $r_1 = 74$ Å.

5.5. Equation (5.10); $n_0 = N_c \exp\left(-\dfrac{E_c - E_F}{k_B T}\right) = N_c \exp\left(-\dfrac{E_c - E_I}{k_B T}\right)$

$\times \exp\left(\dfrac{E_f - E_I}{k_B T}\right) = n_i \exp\left(\dfrac{E_f - E_I}{k_B T}\right)$.

Similarly, $p_0 = n_i \exp\left(\dfrac{E_I - E_F}{k_B T}\right)$.

5.7. Since $R_H = E_y/B_z J_x$ and $J_x = e(n_0\mu_n + p_0\mu_p)\mathcal{E}_x J_y = e(n_0\mu_n + p_0\mu_p)$,
$\mathcal{E}_y = en_0\mu_n B_z V_{xn} + ep_0\mu_p B_z V_{xp} = eB_z\mathcal{E}_x(-n_0\mu_n^2 + p_0\mu_p^2)$,

$\mathcal{E}_y = \dfrac{\mathcal{E}_x B_z(-n_0\mu_n^2 + p_0\mu_p^2)}{n_0\mu_n + p_0\mu_p}$. Thus $R_H = \dfrac{p_0\mu_p^2 - n_0\mu_n^2}{e(n_0\mu_n + p_0\mu_p)^2}$.

If $R_H = 0$, then $p_0\mu_p^2 = n_0\mu_n^2$.

5.9. According to charge neutrality, $0 = q(p_0 - n_0 + N_D - n_D - N_A + p_A)$;
for p-type, $N_A \gg N_D \gg n_D$, $p_0 = n_0 + N_A - p_A - N_D$ and

$$p_A = \frac{N_A}{1 + g^{-1} \exp\left(\dfrac{E_a - E_F}{k_B T}\right)}.$$

At low T, $E_A - E_f \gg k_B T$, $N_A^0 = p_A = N_A$.
As for the kinetic equation, one obtains

$$K_A(T) = N_v g_A^{-1} \exp\left(\frac{E_v - E_A}{k_B T}\right), \text{ for } N_A - N_D \gg p_0,\ p_0 \gg n_0,\ k_A(T)$$
$$= \frac{p_0(p_0 + N_D)}{(N_A - N_D)}.$$

(a) Lightly compensated case, $N_D \ll p_0$,

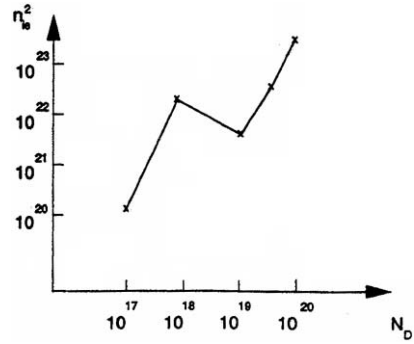
$$p_0 = \sqrt{(N_A - N_D)N_v g_A^{-1}}\ \exp\left(\frac{E_v - E_A}{2k_B T}\right).$$

(b) Heavily compensated case, $N_D \gg p_0$,

$$p_0 = \frac{N_A - N_D}{N_D} N_v g_A^{-1} \exp\left(\frac{E_v - E_A}{2k_B T}\right).$$

(c) Similar to Problem 5.4.

5.11.



Chapter 6

6.1. According to (6.27) and $n_0 \gg p_0 \gg n_1, p_1$, $\tau_0 = \tau_{p0} + \tau_{n0}\dfrac{\Delta p}{n_0 + \Delta p}$, where $n_0 = N_D = 2 \times 10^{15}$ cm^{-3}, $\Delta p = \Delta n$, $\tau_{p0} = \tau_{n0} = 10^{-8}$ s are all known.

6.3. Since the density of the trap is not small compared with the carrier density, Δn and Δp are not necessarily equal. Consider the case in which the disturbance in carrier density is small enough that only first-order terms in Δn and Δp need be considered. Therefore, the recombination rates are a linear function of Δn and Δp:

$$U_{cn} = A_{nn}\Delta n + A_{np}\Delta p \quad \text{and} \quad U_{cp} = A_{pn}\Delta n + A_{pp}\Delta p,$$

where

$$A_{nn} = C_n\left[\frac{n_1}{n_0 + n_t} + \frac{n_0 + n_1}{N_t}\right],\ A_{np} = -C_n\frac{n_0 + n_1}{N_t},$$
$$A_{pn} = -C_p\frac{p_0 + p_1}{N_t},\ A_{pp} = C_p\left[\frac{p_1}{p_0 + p_1} + \frac{p_0 + p_1}{N_t}\right].$$

Steady state, $U_{cn} = U_{cp} = U$,

$$\tau_p = \frac{\Delta p}{U} = \frac{A_{pp} - A_{np}}{A_{nn}A_{pp} - A_{np}A_{pn}}, \quad \text{and} \quad \tau_n = \frac{\Delta n}{U} = \frac{A_{nn} - A_{pn}}{A_{nn}A_{pp} - A_{np}A_{pn}}.$$

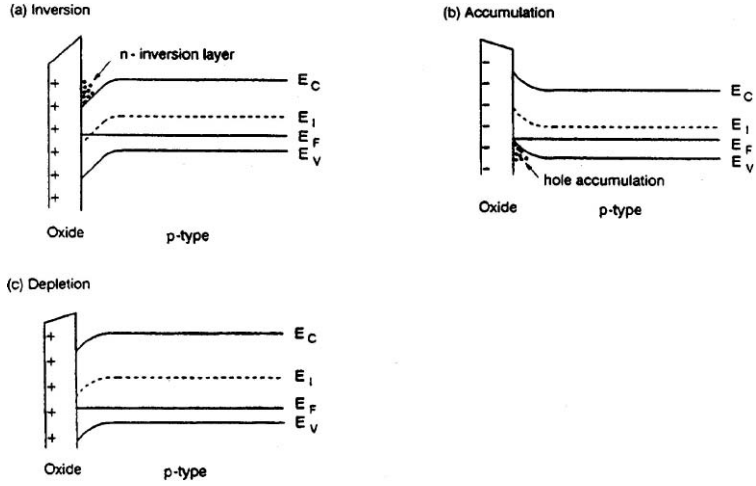
To obtain A's: $U_{cn} = C_n[(1 - f_t)\Delta n - (n_0 + n_1)\Delta f_t]$, $U_{cp} = C_p[f_t\Delta p + (p_0 + p_1)\Delta f_t]$, $\Delta p - \Delta n = N_t\Delta f_t$, $f_t = \dfrac{1}{1 + n_1/n_0} = 1 - \dfrac{1}{1 + p_1/p_0}$.

6.5. $J_n = 0 = q\mu_n n\mathcal{E} + qD_n \dfrac{dn}{dx}$ and $\dfrac{D_n}{\mu_n} = -\mathcal{E}\dfrac{n}{dn/dx}$.

$\dfrac{dn}{dx} = N_c F_{-1/2}(\eta)\dfrac{-1}{k_B T}(q\mathrm{E})$. Therefore $\dfrac{D_n}{\mu_n} = \dfrac{k_B T F_{1/2}(\eta)}{q F_{-1/2}(\eta)}$, where

$\eta = \dfrac{E_f - E_c}{k_B T}$.

6.7.



6.9. The theory is invalid if it is not exponential transients. (see the original paper by Lang).

Chapter 7

7.1. (a) Since E is applied only in the x-direction, $f(k_x, k_y, k_z) = f_0(k_x, \Delta k_x, k_y, k_z)$ and

$$F = -q(E + v \times B) = \hbar\frac{dk}{dt}, \quad dk_x = -\frac{qE_x\,dt}{\hbar} \text{ if } B = 0.$$

For small perturbation, $dt = \tau$ and $dk_x = \Delta k_x$. Therefore,

$$f(k_x, k_y, k_z) = f_0\left(k_x - \frac{qE_x\,dt}{\hbar}, k_y, k_z\right).$$

(b) According to (7.30) (BTE) we obtain

$$-v\nabla f = \frac{f - f_0}{\tau} \quad \text{and} \quad -v\tau\frac{(E - E_f)}{T}\cdot\frac{\partial T}{\partial r}\cdot\frac{\partial f_0}{\partial E} = f - f_0$$
$$= f_0\left(\frac{f}{f_0} - 1\right).$$

For the nondegenerate case: $\dfrac{\hbar^3 k_\mathrm{f}\tau}{2m^{*2}T}(k+k_\mathrm{f})(k-k_\mathrm{f})\dfrac{\partial T}{\partial r}\cdot\dfrac{f_0}{k_\mathrm{B}T}=f_0\left(\dfrac{f}{f_0}-1\right)$.

In addition, $\frac{f}{f_0}=\exp\left(\dfrac{E_\mathrm{eq}-E_\mathrm{noneq}}{k_\mathrm{B}T}\right)=\exp\left(\dfrac{\Delta E}{k_\mathrm{B}T}\right)\approx 1+\dfrac{\Delta E}{k_\mathrm{B}T}$ if $\dfrac{\Delta E}{k_\mathrm{B}T}$ is small.

Therefore, $\dfrac{\hbar^3 k_\mathrm{f}\tau}{2m^{*2}T}(k+k_\mathrm{f})(k-k_\mathrm{f})\dfrac{\partial T}{\partial r k_\mathrm{B}T}=\dfrac{\Delta E}{k_\mathrm{B}T}$ and

$\dfrac{\hbar k_\mathrm{f}\tau}{m^*T}(k-k_\mathrm{f})\dfrac{\partial T}{\partial r}=k-k_\mathrm{noneq}=\Delta k.$

Thus, $\Delta k_x=\dfrac{\hbar k_\mathrm{f}\tau}{m^*T}(k-k_\mathrm{f})\dfrac{\mathrm{d}T}{\mathrm{d}x}$.

(a) For $s=-\frac{3}{2}$, $\sigma_\mathrm{n}=\dfrac{2e^2\tau_0 k_\mathrm{B}T^{3/2}N_\mathrm{c}F_2(\eta)}{m_\mathrm{n}^*}$ and

$S_\mathrm{n}=\dfrac{-k}{e}\left[\dfrac{4F_3(\eta)}{F_2(\eta)}-\dfrac{E_\mathrm{f}}{k_\mathrm{B}T}\right]$.

7.3. According to (7.53),

$$\begin{aligned}\sigma_\mathrm{n}&=\frac{ne^2\langle\tau\rangle}{m_\mathrm{n}^*}=\frac{ne^2\tau_0\int_0^\infty E^{s+3/2}\partial f_0/\partial E\,\partial E}{m_\mathrm{n}^*\int_0^\infty E^{3/2}\partial f_0/\partial E\,\partial E}\\&=\frac{ne^2\tau_0(k_\mathrm{B}T)^s(S+\frac{3}{2})\int_0^\infty \varepsilon^{s+1/2}f_0\,\mathrm{d}\varepsilon}{m_\mathrm{n}^*\frac{3}{2}\int_0^\infty e^{1/2}f_0\,\mathrm{d}\varepsilon}\\&=\frac{2ne^2\tau_0(k_\mathrm{B}T)^s(S+\frac{3}{2})F_{\mathrm{s}+1/2}(\eta)}{3m_\mathrm{n}^*F_{1/2}(\eta)}.\end{aligned}$$

7.5. For the n-type semiconductor,

$$\begin{aligned}J_x=-nev_x&=-\int_0^\infty ev_x f(E)g(E)\,\mathrm{d}E=-e\int_0^\infty v_x\left(-v_x P_x\frac{\partial f_0}{\partial E}\right)g(E)\,\mathrm{d}E\\&=-e\int_0^\infty v_x^2\left[e\tau\mathrm{E}_x-\tau\left(\frac{E_\mathrm{f}-E}{T}\right)\frac{\partial T}{\partial x}\right]\frac{\partial f_0}{\partial E}g(E)\partial E\\&=\frac{-2e}{3m_\mathrm{n}^{*2}}\int_0^\infty \tau E g(E)\frac{\partial f_0}{\partial E}\left[eE_x-\frac{(E_\mathrm{f}-E)\,\partial T}{T\,\partial x}\right]\mathrm{d}E.\end{aligned}$$

As for Q_x, the derivation is similar to J_x except that $Q_\mathrm{x}=n v_\mathrm{x}E$.

7.7. For the longitudinal magnetic field B_x, $P_\mathrm{x}(E)=-e\tau\varepsilon_\mathrm{x}$, where $\partial T/\partial x=\partial T/\partial y=0$. Since

$$\begin{aligned}\sigma_\mathrm{n}&=\frac{J_x}{E_x}=\frac{-nev_x}{E_x}=\frac{e\int_0^\infty v_x^2P_x(E)g(E)\,\partial f_0/\partial E\,\mathrm{d}E}{E_x}\\&=\frac{e^2\int_0^\infty \tau E g(E)f_0\,\mathrm{d}E}{m_\mathrm{n}^*\int_0^\infty Eg(E)f_0\,\mathrm{d}E}=\frac{ne^2\langle\tau\rangle}{m_\mathrm{n}^*}=\sigma_0.\end{aligned}$$

Thus, there is no longitudinal magnetoresistance effect.

7.9. Let $s = -\frac{1}{2}$ for lattice scattering, $\tau_L = aT^{-1}E^{-1/2}$, $\mu_L = a_1 T^{-3/2}$.
For $s = \frac{3}{2}$, ionized impurity scattering, $\tau_i = bE^{3/2}$, $\mu_i = a_2 T^{3/2}$,

$$\mu_n^{-1} = \mu_L^{-1} + \mu_i^{-1} = a_1^{-1}T^{3/2} + a_2^{-1}T^{-3/2}.$$

Chapter 8

8.1. Using (8.18)–(8.20), $\sigma_{k'}(\theta', \phi')$ can be reduced to $\int_0^\infty \frac{(N\Omega)^2(H_{kk'})^2 m^* \delta(E_{k'} - E_k)K' \, dk'}{(2\pi\hbar)^2}$.
In addition, $k = k'$ and $H_{kk'}$ for an isotropic elastic scattering process satisfies

$$\sigma_{k'}(\theta', \phi') = \frac{(N\Omega)^2(H_{kk'})^2(m^*)^2}{(2\pi\hbar)^2\hbar^2} \int_0^\infty \delta(E_{k'} - E_k)\, dE_{k'}.$$

Because $\int_0^\infty \delta(E_{k'} - E_k)\, dE_{k'} = 1$, it follows that

$$\sigma_{k'}(\theta', \phi') = \frac{(N\Omega)^2(H_{kk'})^2(m^*)^2}{(2\pi\hbar^2)^2} = \frac{(N\Omega)^2(H_{kk'})^2 k^2}{(2\pi\hbar v_{k'})^2}.$$

8.3. (a) For an acoustical phonon, $\omega/q = \mu_s$.
$\Delta E = E_{k'} - E_k = \pm\hbar\omega_q = \pm\hbar\mu, 2k\sin(\theta'/2)$ due to conservation of energy and momentum. The maximum energy change occurs at $\theta' = \pi$. Thus $\Delta E_{max} = \hbar\mu_s 2k = 2\mu_s m^* v =$ equation (8.43).

(b) For $T = 100$ K, $v = 5.5 \times 10^6$ cm $\cdot$ sec^{-1}, $\frac{\Delta E}{E_e} = 0.218$.
ΔE is still small compared with E_e. Therefore the assumption of elastic scattering may be justified for T around 100 K.

8.5. Since $\tau_{LI}^{-1} = \tau_L^{-1} + \tau_I^{-1}$ where $\tau_L = l_L/\nu_T\sqrt{x'}$ and $\tau_I = B(x)\nu_T^3 x'^{3/2}$.
l_L is the mean free path for the lattice scattering, $B(x)$ is a slowly varying function, $\nu_T = \sqrt{3k_B T/m^*}$, and $x' = \nu^2/\nu_T^2$. Therefore,

$$\tau_{LI} = \frac{\tau_L x'^2}{\frac{l_L}{B(x')\nu_T^4} + x'^2} \quad \text{and} \quad \mu_{LI} = \frac{e\langle\tau_{LI}\rangle}{m^*},$$

and let $X^2 = 6\mu_L/\mu_i = l_L/B(3)\nu_T^4$,

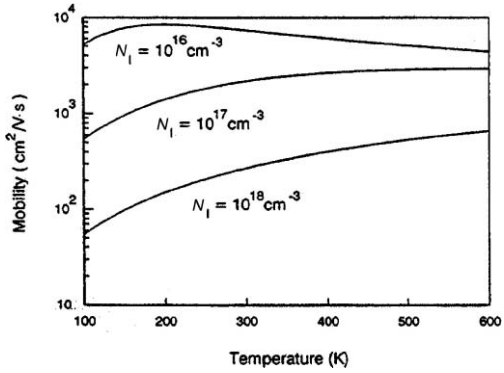
$$\mu_{LI} = \frac{e\tau_L}{m^*}\int_0^\infty \frac{x'^2\exp(-x')}{\frac{l_L}{B(3)\nu_T^4} + x'^2}\, dx'$$
$$= \mu_L\left\{1 + X^2\left[\text{Ci}\, X\cos(X) + \sin(X)\left(\text{Si}\, X - \frac{\pi}{2}\right)\right]\right\}.$$

8.7. Let $\tau_i = \tau_0 E^{3/2}$ and

$$\mu_I = \frac{e\langle\tau_i\rangle}{m^*} = \frac{e\tau_0(k_B T)^{3/2}\Gamma(4)}{m^*\Gamma(\frac{5}{2})} = \frac{64\sqrt{\pi}\varepsilon_0^2\varepsilon_r^2(2k_B T)^{3/2}}{\sqrt{m^*}e^3 N_I} L(2k\lambda_D)^{-1}$$
$$= \text{equation(8.36)}.$$

Since $L(2k\lambda_D) \approx \ln(4k^2\lambda_D^2) = \ln\left(\dfrac{8m^* E\varepsilon_0\varepsilon_r k_B T}{e^2 N_i \hbar^2}\right)$ for $k\lambda_D \gg 1$ and $E = 3k_B T$.

8.9



Chapter 9

9.1. (a) $A\phi_0 = AI_0/h\nu = 4.52 \times 10^{15}\ \text{s}^{-1}$.
(b) For $ad = 64 \gg 1$ and $\exp(-ax) = 0.1$. Thus $x = 0.00719$ cm.
(c) Assume no reflection, $G_E = 3.616 \times 10^{15}\ \text{s}^{-1}$.
(d) Assume $L_n = L_p$, $\Delta G = 2.79 \times 10^{-7}$ ℧.

9.3. (a) Equation (9.91), $I_{ph} = \left[\dfrac{eI_0 WL\mu_p(1+b)\tau_n(1-R)V}{l(L+s\tau_n)h\nu}\right]\left[1 + \dfrac{s\tau_n}{L(1+\alpha L)}\right]$.

Assume $\alpha L \gg 1, \alpha L \gg s\tau_n/L$ and $\Delta n_0 = I_0(1-R)\tau_n/(L+s\tau_n)h\nu$. Thus,

$$I_{ph} = \frac{eWL\mu_p(1+b)V\Delta n_0}{l} = \Delta G V, \quad \Delta G = \frac{eWL\mu_p(1+b)\Delta n_0}{l},$$

where $L = \sqrt{D_a\tau_n}$ and $D_a = \dfrac{2D_n}{1+b}$.

(b) According to (9.125),

$$I_{PME} = eW(1+b)\mu_p B D_a(\Delta n_0 - \Delta n_d) = eW(1+b)\mu_p B D_a \Delta n_0\ ,$$

for $\alpha d \gg 1, d \gg L, \Delta n_d = 0$ and $D_a = 2D_n/(1+b)$.

(c) $V_{PME} = \dfrac{I_{PME}}{\Delta G} = \sqrt{\dfrac{D_a}{\tau_n}}Bl$.

9.5. According to (9.91), $I_{ph} = 1.057 \times 10^{-4}$ A for $s = 0\ \text{cm} \cdot \text{s}^{-1}$.
$I_{ph} = 6.47 \times 10^{-5}$ A for $s = 100$.
$I_{ph} = 2.47 \times 10^{-4}$ A for $s = 10,000$.

9.7. Since $\Delta p \gg n_0$ for the high-injection case, and $B\Delta n^2 = g_E$. Therefore $\Delta n = \sqrt{\alpha I_0(1-R)/Bh\nu}$. In addition, at high injection $I_{ph} \propto \Delta n$. Thus, $I_{ph} \propto \sqrt{I_0}$.

Chapter 10

10.1. (a) For $T = 1873$ K, $J_s = 1$ A $\cdot$ cm^{-2} and $J_s' = \exp(4.39\sqrt{E}T^{-1})$ and $\ln(J_s') = 4.39\sqrt{E}/1873$.

(b) At 873 K, $\ln(J_s) = -24.40$, $\ln(J_s') = -24.40 + \dfrac{4.39\sqrt{E}}{873}$.

At 1500 K, $\ln(J_s) = -5.32$, $\ln(J_s') = -5.32 + \dfrac{4.39\sqrt{E}}{1500}$.

10.3. According to (10.25),

$$W = \sqrt{\frac{2(8.85 \times 10^{-14})(11.7)[0.81 - 0.026 \ln\left(\dfrac{N_D}{2.88 \times 10^{19}}\right) + V_R]}{1.609 \times 10^{-19} N_D}}.$$

10.5.
$$J_{sm} = \int q v_x \, dn, \, dn$$
$$= \frac{4\pi(2m^*)^{3/2}}{h^3} \int (E - E_c)^{1/2} \exp[-(E - E_f)/k_B T] \, dE,$$
$$(E - E_c)^{1/2} = \left(\frac{m^*}{2}\right)^{1/2} v.dE = mv \, dv, 4\pi v^2 \, dv = dv_x \, dv_y \, dv_z,$$
$$v^2 = v_x^2 + v_y^2 + v_z^2,$$
$$J_{sm} = 2q \left(\frac{m^*}{h}\right)^3 \exp(-qV_n/k_B T) \int dv_z \int dv_y \int_{v_{min}}^{\infty} v_x \exp[-m \times (v_x^2 + v_y^2 + v_z^2)/2k_B T] \, dv_x$$
$$= 2q \left(\frac{m^*}{h}\right)^3 \exp(-qV_n/k_B T) \left[\left(\frac{2k_B T\pi}{m^*}\right)^{1/2}\right]^2 \frac{k_B T}{m^*} \exp[-q \times (V_D - V_a)/k_B T]$$
$$= J_0 \exp(qV_a/k_B T) .$$

Comparing with (10.42), we know that
$V_D \gg V_R \Rightarrow J$ is the thermionic emission model,
$V_D \ll V_R \Rightarrow J$ is the diffusion model.

10.7. Since TiW–P–Si: $\phi_{Bp} = 0.55$ eV, therefore for TiW–n–P–Si $\Delta\phi_{Bp} = 0.90 - 0.55 = 0.35$ eV and

$$\Delta\phi_{Bp} = \frac{q}{2\varepsilon_0 \varepsilon_s N_D}(N_A W_p - N_D W_n)^2, \quad W_p = -W_n + \sqrt{W_n^2 + C},$$
$$C = \frac{N_D}{N_A} W_n^2 + \frac{2\varepsilon_0 \varepsilon_s (\phi_m - \phi_p)}{q N_A}$$

for $N_A = 10^{16}$ cm^{-3}, $(\phi_m - \phi_p) \approx 0.365$ eV. Since only two equations are available, we should assume one of the three W_p, W_n, and N_D unknowns to solve for the other two.

10.9. (a) For $D_s \to \infty \Rightarrow C_2 \to 0$.

The Fermi level at the interface is pinned down by the surface states at the values $q\phi_0$ above the valence band. The barrier height is independent of the metal work function.

(b) For $D_s \to 0 \Rightarrow C_2 = 1$,

$$q\phi_{Bn} = q(\phi_m - \chi_s) - q\Delta\phi.$$

The surface state is negligible.

(c) The information given is not enough.

Chapter 11

11.1. $\dfrac{\partial^2 V}{\partial x} = \dfrac{\partial \mathcal{E}}{\partial x} \approx \dfrac{q}{\varepsilon_s \varepsilon_0} ax$. Therefore, $\mathcal{E} = \displaystyle\int_{-W_d/2}^{W_d/2} \frac{q}{\varepsilon_s \varepsilon_0} ax$

$$= \frac{qa}{2\varepsilon_s \varepsilon_0}(x^2 - (W_d/2)^2),$$

$$V(x) = \int \mathcal{E}(x)\,dx = \frac{qa}{2\varepsilon_s \varepsilon_0}\left(x^3/3 - W_d^2 x/4 + W_d^3/12\right)$$

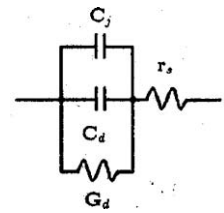
$$V_{bi} = \int_{-W_d/2}^{W_d/2} = \frac{qaW_d^3}{12\varepsilon_s \varepsilon_0}.$$

Therefore, W_d = equation (11.22),

For impurity concentration at depletion $W_d a/2$.

$V_{bi} \approx \dfrac{k_B T}{q} \ln\left(\dfrac{(aW_d)(aW_d)}{n_i^2}\right)$ = equation (11.23).

11.3.



11.5. (a) According to (11.76); let $Q_s(t_{off}) = 0$.

$$\tau_p I_p = \tau_p(I_f + I_r)\exp(-t/\tau_p)$$

Therefore $t_{off} = \tau_p \ln[(I_f + I_r)/I_r] = \tau_p \ln(1 + I_f/I_r)$.

(b) Since $\dfrac{\partial p_n(x,t)}{\partial t} = D_p \dfrac{\partial^2 p_n(x,t)}{\partial x^2} - \dfrac{p_n(x,t) - p_{n0}}{\tau_p}$.

Boundary condition $t = 0$, the initial distribution of holes in a steady-state solution to the equation $V_j = V_T \ln(p_n(0,t)/p_{n0})$.

Second boundary condition at t_{off} or near t_{off}, $p_n(0, t_{off}) \to 0$, $V_j \to -\infty$.

$p(0,t) = p_{n0}$ = constant.

The solution is given by Kingston, *IRE*, 1954(829) as $\text{erf}\sqrt{\left(\frac{t_{\text{off}}}{\tau_{\text{p}}}\right)} = \frac{I_{\text{f}}}{I_{\text{f}} + I_{\text{r}}}$.

11.7. (a) $J_{\text{p}} = 5.8 \times 10^{-10}$, 1.34×10^{-6}, 3.03×10^{-3}, and 6.85 A $\cdot$ cm^{-2} for $V_{\text{f}} = 0.1$, 0.3, 0.5, and 0.7 V, respectively.
(b) $J_{\text{n}} = 6.5 \times 10^{-13}$, 1.50×10^{-9}, 3.40×10^{-6}, and 0.0076 A $\cdot$ cm^{-2} for $V_{\text{f}} = 0.1$, 0.3, 0.5, and 0.7 V, respectively.
(c) $J = J_{\text{n}} + J_{\text{p}} = 0$, 5.3×10^{-6}, and 0.0121 A $\cdot$ cm^{-2} for $V_{\text{a}} = 0$, 0.3, and 0.5 V, respectively.

11.9. (a) $J_0 = 7.45 \times 10^{-3}$, 2.3×10^{-9}, and 1.2×10^{-17} A $\cdot$ cm^{-2} for Ge, Si, and GaAs, respectively.
(b) Because the E_{g} of GaAs is higher than the other two, and the J_0 of GaAs is the least of all.

11.11. If $N_{\text{a}} \gg N_{\text{d}}$, then $W \approx x_{\text{n}}$.

$$|\mathcal{E}_{\text{m}}| = \frac{V}{X_{\text{n}}} = \left(\frac{qN_{\text{d}}(\phi_0 + V_{\text{R}})}{2\varepsilon_0\varepsilon_{\text{s}}}\right)^{1/2} = 10^6 \text{ V} \cdot \text{cm}^{-1}$$

Therefore, for $N_{\text{A}} = 5 \times 10^{19}$ cm^{-3} and $V_{\text{R}} = 2$V, $N_{\text{d}} = 8 \times 10^{18}$ cm^{-3}; and for $N_{\text{A}} = 1 \times 10^{20}$ cm^{-3} and $V_{\text{R}} = 3$V, $N_{\text{d}} = 4 \times 10^{18}$ cm^{-3}.

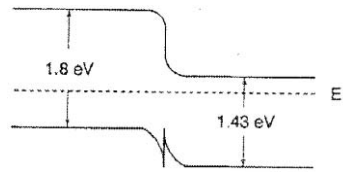
Chapter 12

12.1.

$\lambda(\mu\text{m})$	$\eta(x_0 = 0.4\,\mu\text{m})$	$\eta(x_0 = 0.8\,\mu\text{m})$	$\eta(x_0 = 1.2\,\mu\text{m})$
0.4	0.325	0.189	0.129
0.5	0.577	0.481	0.408
0.7	0.673	0.646	0.622
0.9	0.686	0.681	0.676
1.1	0.005	0.0019	0

12.3. Assuming 16% conversion efficiency at AM1 in our design.
Let $N_{\text{D}} = 5 \times 10^{17}$ cm^{-3}, $N_{\text{A}} = 1.5 \times 10^{16}$ cm^{-3}, $\tau_{\text{n}} = 10 \times 10^{-6}$ s, $D_{\text{n}} = 27\,\text{cm}^2 \cdot \text{s}^{-1}$, $A = 1.07$ cm^2. Then $J_{\text{sc}} = 31.6$ mA $\cdot$ cm^{-2}, $V_{\text{oc}} = 0.592$ V, FF $= 0.861$.

12.5.



12.7 (a) $I_D = 18.35 \times 10^{-6}[\exp(qV/nk_BT) - 1]$ for a Si Schottky barrier solar cell.
$I_D = 1.59 \times 10^{-11}[\exp(qV/k_BT) - 1] + 6.02 \times 10^{-8}[\exp(qV/2k_BT) - 1]$ for a Si-p-n junction solar cell.
(b) $V_{oc} = 0.231$ V for the Schottky barrier solar cell in (a).

12.9. AR coating thickness: $\dfrac{\lambda}{4 \times n} = \dfrac{0.84\,\mu m}{4 \times 1.5} = 1.43\,\mu m$.
Cutoff frequency: $20\,GHz = 1/2\pi RC$, therefore $RC = 7.96 \times 10^{-12}$ s, where $R = (1/qN_D\mu_n) \times (d/\pi r^2)$, $C_d = \pi r^2(qN_D\varepsilon_0\varepsilon_s/2V_d)^{1/2}$, $V_d = \phi_{Bn} - V_T \ln(N_c/N_d)$.
Let $d = 1\,\mu m$, $N_d = 1.6 \times 10^{17}\,cm^{-3}$. To obtain 100 GHz, $N_d = 10^{19}\,cm^{-3}$.

12.11. (a) See Figure 12.35 for the schematic energy band diagrams for n-type QWIPs due to the B-B, B-C, B-M, and B-QB transitions.
(b) The relative magnitude of the spectral response bandwidths for different types of n-QWIPs is given as follows:

$$\Delta\lambda/\lambda(\text{B-B}) < \Delta\lambda/\lambda(\text{B-M}) < \Delta\lambda/\lambda(\text{B-QB}) < \Delta\lambda/\lambda(\text{B-C}).$$

(c) The relative magnitude of the dark currents for different types of n-QWIPs is given as follows:

$$I_d(\text{B-B}) < I_d(\text{B-M}) < I_d(\text{B-QB}) < I_d(\text{B-C}).$$

Chapter 13

13.1. The emission peak wavelength for band-to-band radiative recombination can be calculated using $\lambda_p = 1.24/E_g$
GaAs: $\lambda_p = 1.24/1.42 = 0.873\,\mu m$ (IR); GaN, $\lambda_p = 0.354\,\mu m$ (UV/blue)
$Ga_{0.3}Al_{0.7}As$: $E_g = 1.9$ eV: $\lambda_p = 0.654\,\mu m$ (red), $GaAs_{0.5}P_{0.5}$, $E_g = 2.0$ eV (red)
$In_{0.5}Ga_{0.5}As$, $E_g = 0.88$ eV, IR; $In_{0.5}Ga_{0.5}P$, $E_g = 1.76$ eV (red); $In_{0.5}Al_{0.5}P$, $E_g = 1.85$ eV (red).

13.3. $\lambda_p = 1.24/(E_d - E_a) = 1.24/(2.26 - 0.43 - 0.04) = 0.693\,\mu m$ (red).

13.5. Select your own designed parameters using the formula given in the text.

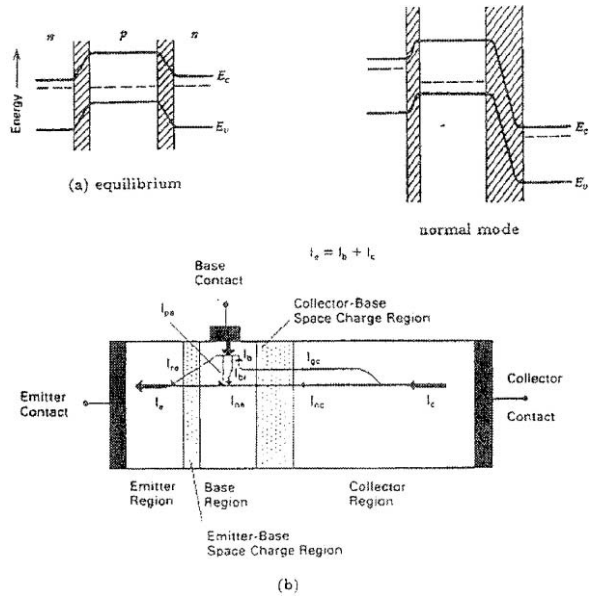
13.7. From Figure 12.28 of S. M. Sze, Physics of Semiconductor Devices, $J_{th} = 120, 400, 2 \times 10^4$ A/cm^2 and $E_g = 1.51, 1.48, 1.43$ eV, at $T = 4, 77$, and 300 K, respectively. Since $P_{th} = J_{th}V_{th}/d$, where $d = 1\,\mu m$, $V_{th} = E_g/q$, thus, $P_{th} = 1.81 \times 10^6, 5.9 \times 10^6, 2.8 \times 10^8$ J/s $\cdot$ cm^2, respectively. The rate of temperature rise $R_T = 0.01P_{th}/C_vD$.
Therefore, $R_T = 9.7 \times 10^4, 3.2 \times 10^4$, and 1.5×10^6 K/s, respectively.

13.9. Choose your own design parameters for this InP-base DH laser diode.

13.11. See Figure 13.30.

Chapter 14

14.1.



14.3. (a) 1.768 μm.
(b) 1.15×10^{18} cm^{-3}.

14.5. (a) Since

$$\beta_T = I_F \Big/ \left[I_F + \frac{qA}{\tau_p} \int_0^W p(x)\mathrm{d}x \right] = 1 \Big/ \left[1 + \frac{qA}{\tau_p J_p} \int_0^W p(x)\mathrm{d}x \right],$$

where $I_p = qAp(x)v(x)$; $v(x)$ is the effective minority carrier velocity in the base and $\mathrm{d}x = v(x)\mathrm{d}t$. For uniformly doped base, the minority carrier lifetime in the base is given by

$$\tau_B = \int_0^W [\mathrm{d}x/v(x)] = qA/I_p \int_0^W p(x)\mathrm{d}x,$$

where $p(x) = \dfrac{I_p}{qAD_pN_d} \displaystyle\int_0^W N_d(x)\mathrm{d}x$ (see problem 14.8)

Solving above equation for τ_B one obtains

$$\tau_B = \frac{1}{D} \int_0^W (1/N_d) \int_0^W N_d(x')\mathrm{d}x' = \frac{W^2}{2D_p}$$

(b) Since a small τ_B means a shorter delay of signal or high-frequency operation. Therefore, the transistor is designed with a small base width in order to achieve a better frequency response.

14.7. (a) 5×10^{16} cm^{-3}.
(b) 6.32×10^{17} cm^{-3}.

14.9. From (13.44) and (13.45),

$$I_E = -\alpha_R I_R + I_{ES}[\exp(V_{be}/V_T) - 1],$$
$$I_C = -\alpha_F I_F + I_{CS}[\exp(V_{bc}/V_T) - 1].$$

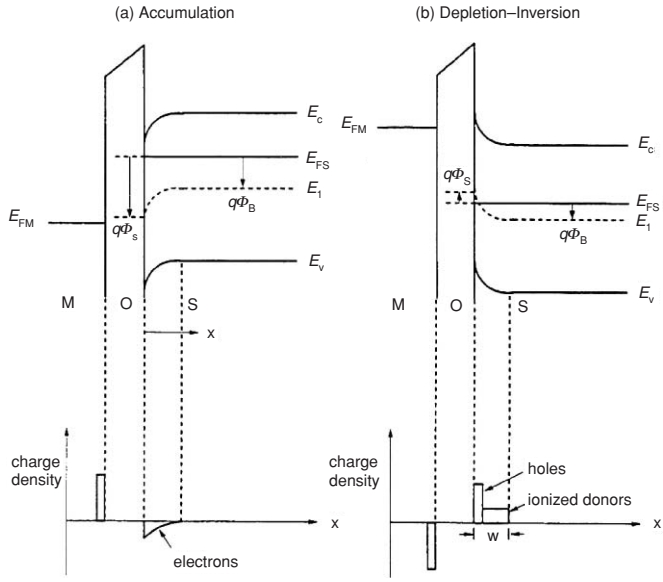
If the space-charge recombination current is negligible, then $I_R = I_C$, $I_F, \approx I_E$, and $\alpha_F I_{ES} = \alpha_R I_{CS}$.
Therefore,

$$V_{BC} = V_T\left(\frac{I_C - \alpha_F(I_B + I_C)}{I_{CS}} + 1\right),$$
$$V_{BE} = V_T\left(\frac{-(I_C + I_B) + \alpha_R I_c}{I_{ES}} + 1\right)$$
$$V_{CE} = V_{BE} - V_{BC} = V_T\left[\ln\left(\frac{I_{ES} + \alpha_R I_C - I_B - I_C}{I_{CS} - \alpha_F(I_b + I_C) + I_C}\right) + \ln(\alpha_F/\alpha_R)\right]$$

$-V_{ce}$ = given for proof.

Chapter 15

15.1.



15.3. According to (14.45), (14.46), and (14.35), we obtain (14.47):

$$I_D = C_{0x}\mu_n \frac{Z}{L}\left(V_G - V_{th} - \frac{V_D}{2}\right)V_D.$$

If considering (14.39) instead of (14.37), we include another term

$$-\frac{Q_B}{C_{0x}} = \frac{\sqrt{2q\varepsilon_0\varepsilon_s N_A(V_c + \Psi_{si})}}{C_{0x}}.$$

Integrating this additional term from $y = 0$ to $y = L$ and from $V = 0$ to $V = V_D$, we obtain

$$\frac{2\sqrt{2q\varepsilon_0\varepsilon_s N_A}}{3C_{0x}}\left[(V_D + \Psi_{si})^{3/2} - \Psi_{si}^{3/2}\right].$$

Therefore, adding this term to (14.47), we obtain (14.48).

15.5. $I_{DS} = C_{0x}\mu_n \dfrac{Z}{2L}(V_G - V_{th})^2 = 28.7 \times (3 - 0.5)^2 = 0.18\,\text{mA}$ for $V_G = 3\text{V}$ and $I_{DS} = 0.58$ mA for $V_G = 5$ V.

15.7. Mobile sodium ion charges will move to the SiO_2–Si interface, and thus V_{th} will increase.

15.9. Solutions of the equations are:

$$\phi = (V_G - V_{FB}) - \varepsilon_{0x}(x + d), \quad \text{for} - d < x < 0.$$

$$\phi = \phi_{max} - \frac{qN_d}{2\varepsilon_0\varepsilon_s}(x - W_n)^2 + c_1(x - W_n), \quad \text{for } 0 < x < W_n.$$

$$\phi = \frac{qN_A}{2\varepsilon_0\varepsilon_s}(x - W_n - W_p)^2 + c_2(x - W_n - W_p), \quad \text{for } W_n < x < W_n + W_p.$$

In addition,

$$-\varepsilon_{0x} = \frac{qN_D}{\varepsilon_0\varepsilon_s}W_n,$$

$$(V_G - V_{FB}) - \varepsilon_{0x}d = \phi_{max} - \frac{qN_D}{2\varepsilon_0\varepsilon_s}W_n^2,$$

$$-\frac{qN_A}{\varepsilon_0\varepsilon_s}W_p + c_2 = c_1,$$

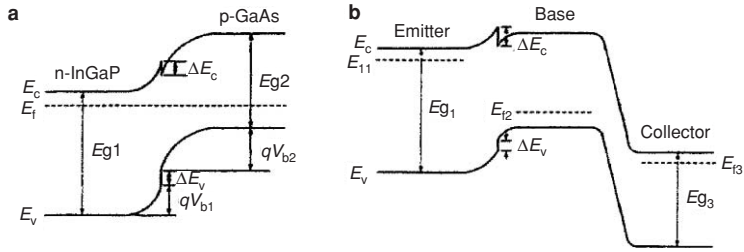
$$\phi_{max} = \frac{qN_A}{2\varepsilon_0\varepsilon_s}W_p^2 - c_2 W_p.$$

Therefore, $\phi_{max} = \dfrac{qN_A^2}{2\varepsilon_0\varepsilon_s N_D}W_p^2 + \dfrac{qN_A}{2\varepsilon_0\varepsilon_s}W_p^2 = \dfrac{qN_A}{2\varepsilon_0\varepsilon_s}W_p^2\left(\dfrac{N_A}{N_D} + 1\right)$.

Chapter 16

16.1. Since $q\Delta = k_B T \ln(N_c/N_d) = 0.022$ ev and $qV_{bi} = \frac{2}{3}E_g - q\Delta = 0.924$ eV, therefore $d = 14.44 \times 10^{-6}$ cm.

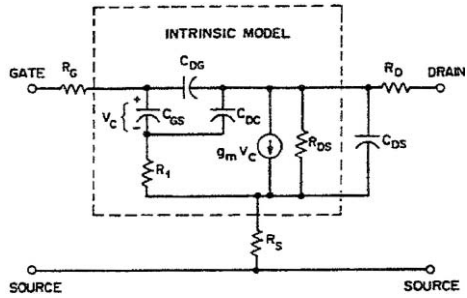
16.3.



16.5. (a) v_s equals 2.22×10^6, 1.51×10^6, and 1.02×10^6 cm/s for $L =$ 0.25, 0.50, and 1.0 μm, respectively.

(b) f_T equals 14.1, 4.8, and 1.6 GHz, respectively.

16.7. (a)



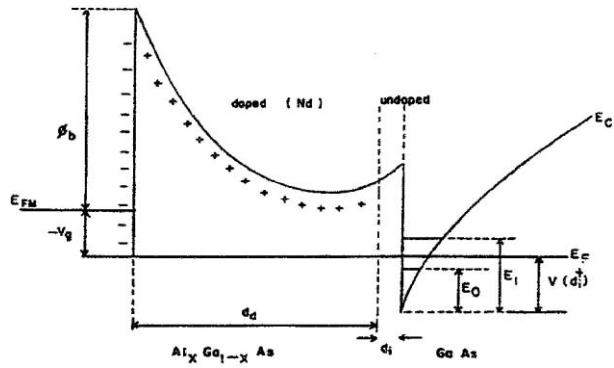
(b) Since $f_T = \dfrac{g_m}{2\pi(C_{gs} + C_{gd})}$ and $C_{gs} + C_{gd} = \dfrac{\varepsilon_0 \varepsilon_s Z L}{W_d}$, therefore,

$$f_T = \frac{\varepsilon_0 \varepsilon_s v_{sat} Z}{W_d 2\pi} \frac{W_d}{\varepsilon_0 \varepsilon_s Z L} = \frac{v_{sat}}{2\pi L}.$$

(c) f_T is 191 GHz and 19.1 GHz for $L = 0.1$ μm and 1 μm, respectively.

(d) $g_{me} = \dfrac{g_{mi}}{1 + g_{mi} R_s}$ and $g_{dse} = \dfrac{g_{dxi}}{1 + g_{dsi} R_s}$.

16.9. (a)

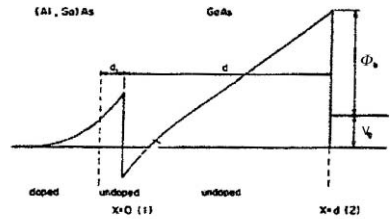


Due to modulation doping, a 2-DEG charge sheet can be formed in the triangle well of the undoped GaAs buffer layer.

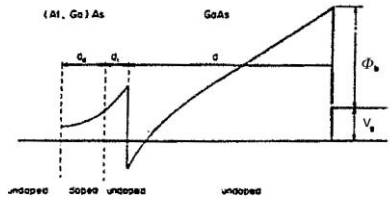
(b) Since no ionized impurity scatterings are expected in the 2-DEG well.

(c) Due to high electron mobility.

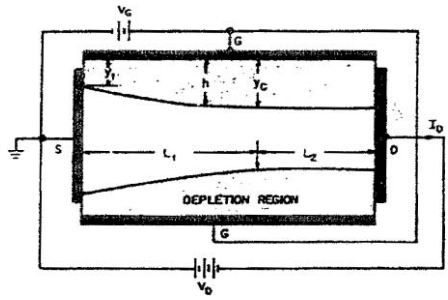
(d)



MODFET I

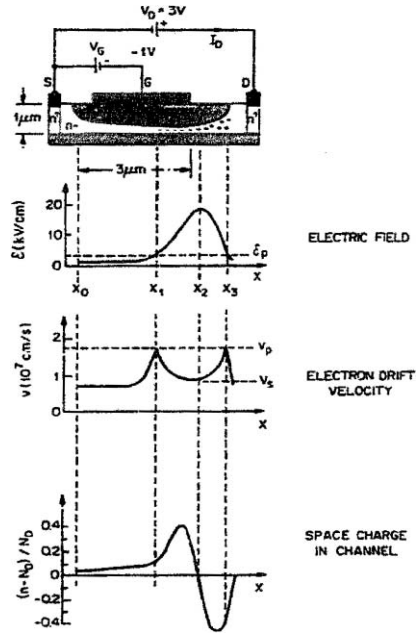


16.11. (a)



Refer to Figure 16.1.

(b)



Refer to Figure 16.4.

Appendix

TABLE A.1. Physical constants.

Avogadro's number	$N_A = 6.02214 \times 10^{23}$ atoms/g mol
Bohr radius	$a_B = 0.52917$ Å
Boltzmann constant	$k_B = 1.38066 \times 10^{-23}$ J/K
Electronic charge	$e = 1.60218 \times 10^{-19}$ C
Free electron rest mass	$m_0 = 9.11 \times 10^{-31}$ kg
Gas constant	$R = 1.98719$ cal/mol K
Permeability in vacuum	$\mu_0 = 1.25664 \times 10^{-8}$ H/cm $(=4\pi \times 10^{-9})$
Permittivity in vacuum	$\varepsilon_0 = 8.85418 \times 10^{-14}$ F/cm
Planck constant	$h = 6.62607 \times 10^{-34}$ J s
Reduced Planck constant	$\hbar(h/2\pi) = 1.05457 \times 10^{-34}$ J s
Proton rest mass	$M_p = 1.67262 \times 10^{-27}$ kg
Speed of light in vacuum	$c = 2.99792 \times 10^8$ m/s
Thermal voltage	$k_B T = 0.025852$ eV

TABLE A.2. International system of units (SI units).

Quantity	Unit	Symbol	Dimension
Current	ampere	A	
Length	meter	m	
Mass	kilogram	kg	
Time	second	s	
Temperature	kelvin	K	
Luminous intensity	candela	Cd	
Luminous flux	lumen	lm	
Frequency	hertz	Hz	1/s
Force	newton	N	kg m/s^2
Pressure	pascal	Pa	N/m^2
Energy	joule	J	N m
Power	watt	W	J/s
Electric charge	coulomb	C	A s
Potential	volt	V	J/C
Conductance	siemens	S	A/V
Resistance	ohm	Ω	V/A
Capacitance	farad	F	C/V
Inductance	henry	H	Wb/A
Magnetic flux	weber	Wb	V s
Magnetic flux density	tesla	T	Wb/m^2

Index