2

Review of Basic Device Physics

2.1 Introduction

Compact device models for circuit CAD (computer-aided design) requires detailed description of the transistor characteristics in the circuit environment under various biasing conditions. Transistor characteristics, however, depend on the material properties of the basic building blocks of each transistor along with its geometrical and structural information. IC (integrated circuit) transistors are fabricated on a semiconductor substrate such as silicon to achieve the desired device characteristics for the target circuit performance. These device characteristics are modulated by the transport of current-carrying fundamental constituents of matter referred to as the *electrons* and *holes*. Again, the electronic properties of semiconductors, primarily, depend on the transport of the majority carrier electrons or holes. The semiconductors with the majority carrier concentration as electrons are referred to as the *n*-type, whereas, the semiconductors with the majority carrier concentration as holes are referred to as the *p*-type. Thus, in order to understand the compact device models for circuit CAD, it is essential to understand the basic physics of the elemental *n*-type, and *p*-type semiconductors along with the transport properties of electrons and holes in building IC devices. Though a number of published titles are available on the subject, the objective of this chapter is to present a brief overview of the basic semiconductor theory along with the basics of *n*-type and *p*-type semiconductors in contact forming *pn*-junctions that are necessary to develop compact transistor models for circuit CAD. The review is brief and covers only those topics that have direct relevance to the field-effect transistor ICs. For more exclusive treatments, the readers are referred to textbooks on the subject [1–13].

2.2 Semiconductor Physics

Crystalline silicon is a widely used semiconductor-starting material in the fabrication of IC devices and chips. Thus, unless otherwise specified, in this book, the semiconductor physics is described with reference to silicon.

The silicon wafers used in the IC fabrication processes are cut parallel to either the <111> or <100> crystal planes. However, the <100> material is most commonly used due to the fact that, during IC fabrication processes, <100> wafers produce the lowest amount of charges at the silicon/silicon-dioxide (Si/SiO_2) interface and offer higher carrier mobility [14,15].

2.2.1 Energy Band Model

In a silicon crystal, each atom has four valence electrons and four nearest neighboring atoms. Each atom shares its valence electrons with its four neighbors in a paired configuration called *covalent bond*. It is predicted by quantum mechanics (QM) that the allowed energy levels of electrons in a solid is grouped into two bands, called the *valence band* (VB) and the *conduction band* (CB). These bands are separated by an energy range that the electrons in a solid cannot possess and is referred to as the *forbidden band* or *forbidden gap*. The VB is the highest energy band and its energy levels are mostly filled with electrons forming the covalent bonds. The CB is the next higher energy band with its energy levels nearly empty. The electrons that occupy the energy levels in the CB are called *free electrons* or *conduction electrons*.

Typically, the energy is a complex function of momentum in a three-dimensional space and there are many allowed energy levels for a large number of electrons in silicon, and therefore, the energy band diagram is also complex. For the simplicity of representation, only the edge levels of each of the allowed energy bands are shown in the energy band diagram in Figure 2.1. In Figure 2.1, E_c and E_v are the bottom edge of the CB and the top edge of the VB, respectively, and E_g is the bandgap energy separating E_c and E_v. And, at any ambient temperature T(K), E_g is given by

$$E_g = E_c - E_v \tag{2.1}$$

When a valence electron is given sufficient energy ($\geq E_g$), it can break out of the chemical bonding state and excite into the CB to become a free electron leaving behind a vacancy, or *hole* in the VB. A hole is associated with a positive charge since a net positive charge is associated with the atom from which the electron broke away. Note that both the electron and hole are generated simultaneously from a single event. The electrons move freely in the CB and holes move freely in the VB. In silicon, the bandgap is small (~1.12 eV); therefore, even at room temperature a small fraction of the valence electrons are excited into the CB, generating electrons and holes. This allows limited conduction to take place from the motion of the electrons in the CB and holes in the VB. As shown in Figure 2.1, when an electron in the CB gains energy, it moves up to an energy $E > E_c$, while a hole in the VB gains energy, it moves down to an energy $E < E_v$. Thus, the energy of the electrons in the CB increases upward while the energy of the holes in the VB increases downward.

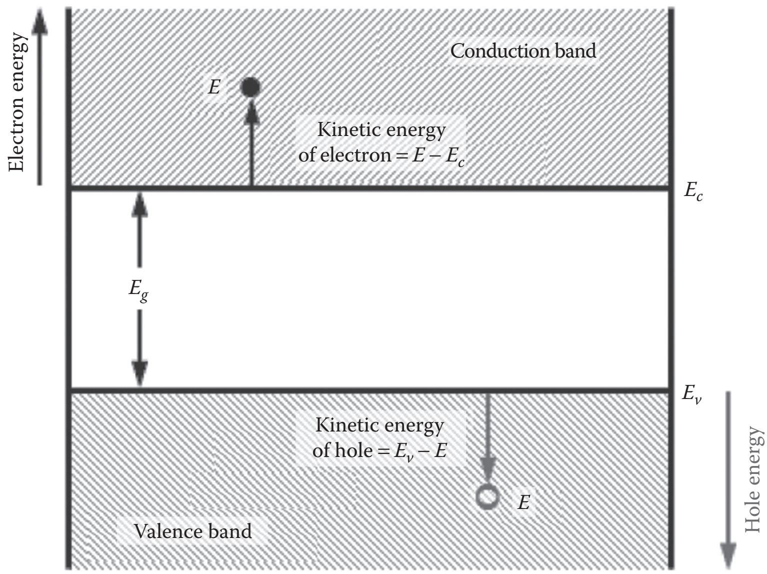

FIGURE 2.1
Energy band diagram of a semiconductor like silicon: E_c is the bottom edge of the CB and E_v is the top edge of the VB; the CB and VB are separated by an energy gap $E_g = E_c – E_v$.

The bandgap energy, E_g, for silicon at room temperature (300 K) is ~1.12 eV. As the temperature increases, the value of E_g for most semiconductors decreases due the increase in the crystal lattice spacing by thermal expansions. For silicon, the temperature coefficient of E_g at 300 K temperature is: $dE_g/dT \cong -2.73 \times 10^{-4}\,\text{eV/K}$ [16]. The temperature dependence of E_g for silicon can be modeled by using polynomial equations valid for different range of temperatures [16,17]. However, in circuit CAD tool SPICE (Simulation Program with Integrated Circuit Emphasis) [18], the temperature dependence of E_g is modeled by [19]

$$E_g(T) = 1.160 - \frac{7.02 \times 10^{-4} T^2}{1108 + \mathrm{T}} \tag{2.2}$$

where:
T is the temperature in Kelvin (K)
$E_g(T)$ is in eV

2.2.2 Carrier Statistics

The electrical properties of a semiconductor are determined by the number of carriers available for conduction. This number is determined from the density of states and the probability that these states are occupied by carriers. The probability that an available state with energy E is occupied by an electron under a thermal equilibrium condition is given by the *Fermi–Dirac* probability density function $f(E)$, also called the *Fermi function* [1–11].

$$f(E) = \frac{1}{1 + \exp\left[\left(E - E_f\right)/kT\right]} \tag{2.3}$$

where:
E_f is the *Fermi energy* or *Fermi level*
$k = 1.38 \times 10^{-23}$ J K^{-1} is the Boltzmann constant
T is the ambient temperature

The Fermi level is the energy at which the probability of finding an electron, at any $T > 0°$ K, is exactly one-half (Equation 2.3). From Equation 2.3, we find that when $E = E_f$, $f(E) = 1/2$, which means that the electron is equally likely to have an energy above E_f as below it. At absolute zero temperature ($T = 0°$ K): $f(E) = 1$ for $E < E_f$, indicating that the probability of finding an electron below E_f is unity and above E_f is zero (that is, $f(E) = 0$ for $E > E_f$). In other words, all energy levels below E_f are filled and all energy levels above E_f are empty. At finite temperatures, some states above E_f are filled and some states below E_f become empty. As T increases above absolute zero, the function $f(E)$ changes as shown in Figure 2.2. Thus, the probability that the energy levels above E_f are filled increases with temperature. It is important to note that the Fermi function or Fermi energy applies only under equilibrium conditions.

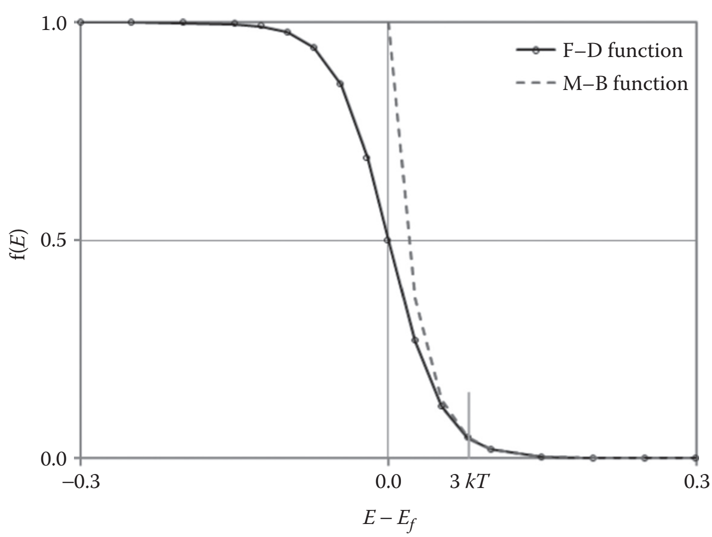

FIGURE 2.2
Fermi–Dirac (F–D) and Maxwell–Boltzmann (M–B) distribution functions in a semiconductor; the plots show that the F–D distribution can be approximated to M–B distribution at temperature, $T > 3\,kT$.

Equation 2.3 describes the probability of an allowed energy state occupied by an electron with $E > E_f$. Then the probability of a state not occupied by an electron (with $E < E_f$) is given by

$$1 - f(E) = \frac{1}{1 + \exp\left[\left(E_f - E\right)/kT\right]} \tag{2.4}$$

Equation 2.4 is the probability function describing that a hole exists.

As shown in Figure 2.2, the probability distribution $f(E)$ makes a smooth transition from unity to zero as the energy increases across the Fermi level. The width of the transition is governed by the thermal energy kT. The value of thermal energy at room temperature is about 26 mV. Thus, for all energy at least several kT (~3 kT) above E_f, the function $f(E)$ in Equations 2.3 and 2.4 can be approximated by the simple expressions

$$f(E) \cong \exp\left(-\frac{E - E_f}{kT}\right) \quad \text{for } E > E_f \tag{2.5}$$

and

$$1 - f(E) \cong \exp\left(-\frac{E_f - E}{kT}\right) \quad \text{for } E < E_f \tag{2.6}$$

Equations 2.5 and 2.6 are identical to Maxwell Boltzmann density function for classical gas particles. For most device applications at room temperature, the function $f(E)$ given by Equation 2.5 is a good approximation as shown in Figure 2.2.

Fermi level can be considered to be the chemical potential for electrons and holes. Since the condition for any system in equilibrium is that the chemical potential must be constant throughout the system, it follows that the Femi level must be constant throughout a semiconductor in equilibrium.

2.2.3 Intrinsic Semiconductors

An *intrinsic semiconductor* is a perfect single crystal semiconductor with no impurities or lattice defects. In such materials, the VB is completely filled with electrons and the CB is completely empty. Therefore, in intrinsic semiconductors, there are no charge carriers at 0° K. At higher temperatures electron–hole pairs are generated as VB electrons are thermally excited across the bandgap to the CB. In intrinsic semiconductors, all the electrons in the CB are thermally excited from the VB. In other words, at a given temperature, the number of holes in the VB equals the number of electrons in the CB of an intrinsic semiconductor. Thus, if n and p are the concentrations of free electrons and holes, respectively, then

$$n = p = n_i \tag{2.7}$$

or,

$$np = n_i^2 \tag{2.8}$$

where:

n_i is called the intrinsic carrier concentration and is the free electron (or hole) concentration in an intrinsic semiconductor

2.2.3.1 Intrinsic Carrier Concentration

From the effective densities of carriers and probability distribution function, we can derive the expression for the intrinsic carrier concentration in a semiconductor. Thus, from Equations 2.5 and 2.6, we can write the concentration of electrons in the CB as

$$n \cong N_c \exp\left(-\frac{E_c - E_f}{kT}\right) \tag{2.9}$$

and the concentration of holes in the VB as

$$p \cong N_v \exp\left(-\frac{E_f - E_v}{kT}\right) \tag{2.10}$$

where:

N_c and N_v are the effective densities of states in the CB and VB, respectively

The expressions for N_c and N_v are derived from QM considerations [5]. Both N_c and N_v are proportional to $T^{3/2}$. For an intrinsic semiconductor, $n = p = n_i$ and E_f is called the intrinsic Fermi level, or the *intrinsic energy level*, E_i. Then (using $n = p = n_i$) we can write from Equations 2.9 and 2.10,

$$N_c \exp\left(-\frac{E_c - E_f}{kT}\right) = N_v \exp\left(-\frac{E_f - E_v}{kT}\right) \tag{2.11}$$

Now, solving Equation 2.11 for $E_f = E_i$, we get the expression for the *intrinsic energy* level as

$$E_i = E_f = \frac{E_c + E_v}{2} - \frac{kT}{2}\ln\left(\frac{N_c}{N_v}\right) \tag{2.12}$$

From Equation 2.12, it can be shown that the intrinsic Fermi level, E_i, is only about 7.3 meV below the mid-gap at $T = 300°$ K. Since $kT << (E_c + E_v)$, Equation 2.12 can be simplified to

$$E_i = E_f \cong \frac{E_c + E_v}{2} \tag{2.13}$$

Thus, the intrinsic Fermi level in a semiconductor material is very close to the midpoint between the CB and the VB, and for all practical purposes, it can be assumed that E_i is in the middle of the energy gap. Thus, E_i is commonly referred to as the *mid-gap* energy level.

In order to derive an expression for the intrinsic carrier concentration as a function of T, we multiply Equations 2.9 and 2.10 to get

$$np = n_i^2(T) = N_c N_v \exp\left(-\frac{E_c - E_v}{kT}\right) = N_c N_v \exp\left[-\frac{E_g(T)}{kT}\right]$$

or (2.14)

$$n_i(T) = CT^{3/2} \exp\left(-\frac{E_g(T)}{2kT}\right)$$

where:

C is a constant

E_g is the bandgap energy defined in Equation 2.1

k is the Boltzmann constant (8.62×10^{-5} eV K^{-1})

The term kT has the dimension of energy and is called *thermal energy* and is equal to 25.86 meV at $T = 300°$ K

Substituting the values for N_c and N_v [6], we can express Equation 2.14 as

$$n_i(T) = 3.9 \times 10^{16} T^{3/2} \exp\left(-\frac{E_g(T)}{2kT}\right) \tag{2.15}$$

If $E_g(T_{NOM})$ and $n_i(T_{NOM})$ are the values of E_g and n_i at the nominal or the reference temperature T_{NOM}, respectively, then we can show

$$n_i(T) = n_i\left(T_{NOM}\right).\left(\frac{T}{T_{NOM}}\right)^{3/2} \exp\left[-\frac{E_g(T)}{2kT} + \frac{E_g\left(T_{NOM}\right)}{2kT_{NOM}}\right] \tag{2.16}$$

where $E_g(T)$ is given by Equation 2.2. The above expression is used in circuit CAD for calculation of n_i at any temperature T with $n_i = 1.45 \times 10^{10}$ cm^{-3} at $T = 300°$ K [6].

2.2.3.2 Effective Mass of Electrons and Holes

The electrons in the CB and holes in the VB move freely throughout the crystal like free particles, suffering only occasional scattering by impurities and defects present in the crystal. The free electrons experience Coulomb force

TABLE 2.1

Effective Mass Ratio for Silicon at 300 K (m_0 is the Free Electron Mass)

Carriers	Density of states effective mass (m_n^*/m_0)	Conductivity effective mass (m_n^*/m_0)
Electrons	1.08	0.26
Holes	0.81	0.386

due to the charged atomic cores of the host atoms in a regular lattice, giving rise to a periodic potential energy. The effect of the periodic potential of the crystal lattice on the motion of electrons in the CB and holes in the VB is represented by the effective masses of the electrons (m_n^*) and holes (m_p^*), respectively. In practice, there are several types of mass used for a given material and carrier type [1–11]. The effective mass required to calculate the carrier (electron and hole) concentration is called the *density of states effective mass*, whereas the mass required to calculate carrier mobility is called the *conductivity effective mass*. These effective masses depend on temperature. There is a large variation in the reported values of m_n^* and m_p^* [16]. The commonly used values for the effective mass for electrons and holes at room temperature are summarized in Table 2.1 [6].

2.2.4 Extrinsic Semiconductors

An *extrinsic semiconductor* is a semiconductor material with added elemental impurities called *dopants*. As we discussed in Section 2.2.3, the intrinsic semiconductor at room temperature has an extremely low number of free-carrier concentration, yielding very low conductivity. The added impurities introduce additional energy levels in the forbidden gap and can easily be ionized to add either electrons to the CB or holes to the VB, depending on the type of impurities and impurity levels.

Silicon is a column-IV element with four valence electrons per atom. There are two types of impurities in silicon that are electrically active: those from column V such as arsenic (As), phosphorous (P), and antimony (Sb); and those from column III such as boron (B). A column-V atom in a silicon lattice tends to have one extra electron loosely bound after forming covalent bonds with silicon atoms as shown in Figure 2.3a. In most cases, the thermal energy at room temperature is sufficient to ionize the impurity atom and free the extra electron to the CB. Such type of impurities (P, Sb, and As) are called *donor* atoms, since they donate an electron to the crystal lattice and become positively charged. Thus, the P, Sb, and As doped silicon is called *n*-type material that contains excess electrons and its electrical conductivity is dominated by electrons in the CB. On the other hand, a column-III impurity atom in a silicon lattice tends to be deficient of one electron when forming covalent bonds with other silicon atoms as shown in Figure 2.3b. Such an impurity (B) atom can also be ionized by accepting an electron from the VB, which leaves

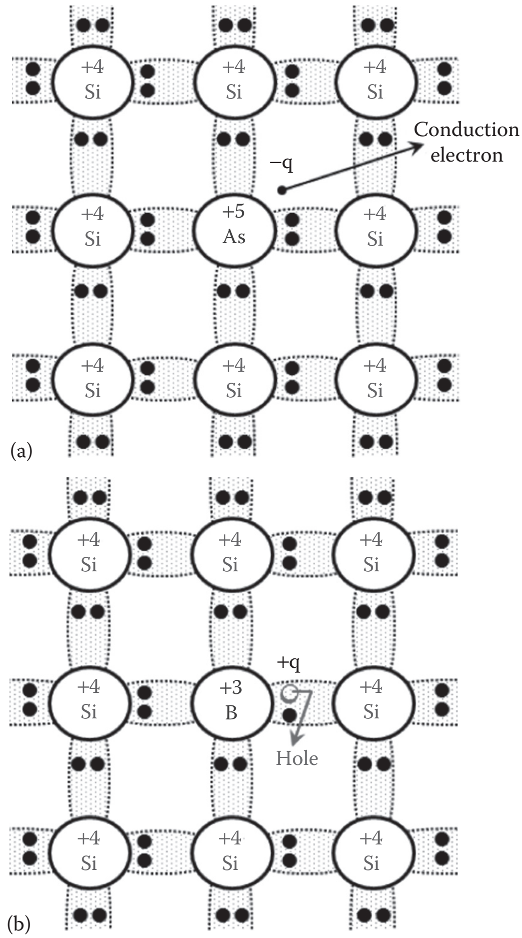

FIGURE 2.3
Extrinsic semiconductors forming covalent bonds: (a) an arsenic donor atom in silicon providing one electron for conduction in the CB and (b) a boron acceptor atom in silicon creating a hole for conduction in the VB.

a freely moving hole that contributes to electrical conduction. These impurities (e.g., B) are called acceptors, since they accept electrons from the VB, and the doped silicon is called *p*-type that contains excess holes.

Thus, we can see from Figure 2.3, the donor and acceptor atoms occupy substitutional lattice sites and the extra electrons or holes are very loosely bound, that is, can easily move to the CB or VB, respectively. In terms of energy band diagrams, donors add allowed electron states in the bandgap close to the CB edge as shown in Figure 2.4a whereas acceptors add allowed states just above the VB edge as shown in Figure 2.4b. Figure 2.4 also shows the positions of the Fermi level due to donors (Figure 2.4c) and acceptors (Figure 2.4d). Donor levels contain positive charge when ionized (emptied).

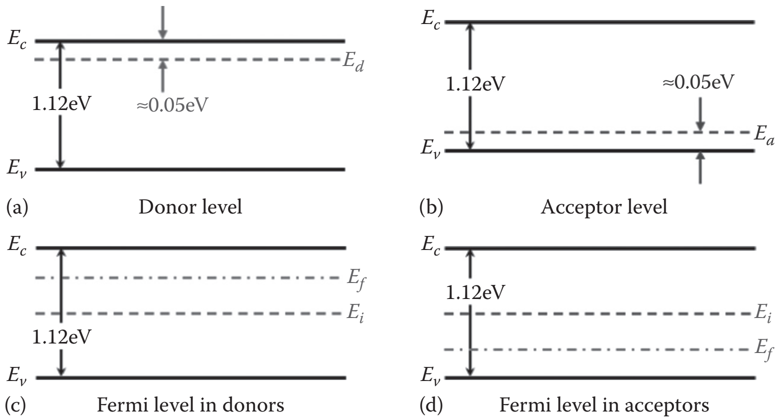

FIGURE 2.4
Energy band diagram representation in extrinsic semiconductors: (a) donor level E_d, (b) acceptor in silicon E_a, (c) intrinsic energy level and Fermi level in an *n*-type semiconductor, and (d) intrinsic energy level and Fermi level in a *p*-type semiconductor.

Acceptor levels contain negative charge when ionized (filled). A donor level E_d shown in Figure 2.4a is measured from the bottom of the CB whereas an acceptor level E_a shown in Figure 2.4b is measured from the top of the VB. The ionization energies for donors and acceptors are $(E_c–E_d)$ and $(E_a–E_v)$, respectively.

It is possible to dope silicon so that $p = n$. Material of this type is called *compensated* silicon. In practice, however, one type of impurity dominates over the other so that the semiconductor is either *n*-type or *p*-type. A semiconductor is said to be *nondegenerate* if the Fermi level lies in the bandgap more than a few kT (~3 kT) from either band edge. Conversely, if the Fermi level is within a few kT (~3 kT) of either band edge, the semiconductor is said to be *degenerate*. In the *nondegenerate* case, the carrier concentration obeys Maxwell-Boltzmann statistics given by Equations 2.5 and 2.6. However, for the degenerate case where the dopant concentration is in excess of approximately 10^{18} cm^{-3} (heavy doping), one must use Femi-Dirac distribution function given by Equations 2.3 and 2.4. Unless otherwise specified, we will assume the semiconductor to be nondegenerate.

2.2.4.1 Fermi Level in Extrinsic Semiconductor

In contrast to intrinsic semiconductor, the Fermi level in extrinsic semiconductor is not located at the mid-gap. The Fermi level in an *n*-type silicon moves up toward the CB, consistent with the increase in electron density described by Equation 2.9. On the other hand, the Fermi level in a *p*-type silicon moves toward the VB, consistent with the increase in hole density described by Equation 2.10. These cases are depicted in Figure 2.4c and d. The exact position of the Fermi level depends on both the ionization energy and concentration of dopants. For example for an n-type material with a donor impurity concentration N_d, the charge neutrality condition in silicon requires that

$$n = N_d^+ + p \tag{2.17}$$

where:
N_d^+ is the density of ionized donors

Using Equation 2.4 we can write

$$N_d^+ = N_d\left[1 - f\left(E_d\right)\right] = N_d\left\{1 - \frac{1}{1 + (1/2)\exp\left[\left(E_d - E_f\right)/kT\right]}\right\} \tag{2.18}$$

where:
$f(E_d)$ is the probability that a donor state is occupied by an electron in the normal state
E_d is the energy of the donor level

The factor 1/2 in the denominator of $f(E_d)$ arises from the spin degeneracy (up or down) of the available electronic states associated with an ionized level [20].

Substituting Equations 2.9 and 2.10 for n and p, respectively, and Equation 2.18 for N_d^+ in Equation 2.17, we get

$$N_c \exp\left(-\frac{E_c - E_f}{kT}\right) = \frac{N_d}{1 + 2\exp\left[-\left(E_d - E_f/kT\right)\right]} + N_v \exp\left(-\frac{E_f - E_v}{kT}\right) \tag{2.19}$$

Equation 2.19 can be solved for E_f. For an n-type semiconductor, $n >> p$; therefore, the second term on the right hand side of Equation 2.19 can be neglected. Now, assuming $(E_d – E_f) >> kT$, $\exp\left[-\left(E_d - E_f\right)/kT\right] << 1$. Therefore, from Equation 2.19 we get after simplification

$$E_c - E_f = kT \ln\left(\frac{N_c}{N_d}\right) \tag{2.20}$$

In this case, the Fermi level is at least a few kT below E_d and essentially all the donor levels are ionized, that is, $n = N_d^+ = N_d$ for an n-type semiconductor. Then from Equation 2.8, the hole density in an n-type semiconductor is given by

$$p = \frac{n_i^2}{N_d} \tag{2.21}$$

Similarly, for a p-type silicon with a shallow acceptor concentration N_a, the Fermi level is given by

$$E_f - E_v = kT \ln\left(\frac{N_v}{N_a}\right) \tag{2.22}$$

In this case, the hole density is $p = N_a^- = N_a$, and the electron density is

$$n = \frac{n_i^2}{N_a} \tag{2.23}$$

Instead of using Equations 2.20 and 2.22, we can express these in terms of E_f and E_i using Equations 2.9 and 2.10. From Equation 2.9, the intrinsic carrier concentration can be shown as

$$n_i \cong N_c \exp\left(-\frac{E_c - E_i}{kT}\right) \tag{2.24}$$

Or,

$$E_c = E_i + kT \ln\left(\frac{N_c}{n_i}\right) \tag{2.25}$$

Then substituting for E_c from Equation 2.25 into Equation 2.20, we get for an n-type silicon,

$$E_f - E_i = kT \ln\left(\frac{N_d}{n_i}\right) \tag{2.26}$$

Similarly, using Equation 2.10, we can express Equation 2.22 for a p-type silicon by

$$E_i - E_f = kT \ln\left(\frac{N_a}{n_i}\right) \tag{2.27}$$

Equations 2.26 and 2.27 are the measure of the Fermi level with reference to the mid-gap energy level for the n-type and p-type semiconductors, respectively.

2.2.4.2 Fermi Level in Degenerately Doped Semiconductor

For heavily doped silicon, the impurity concentration N_d or N_a can exceed the effective density of states N_c or N_v, so that $E_f \geq E_c$ and $E_f \leq E_v$ according to Equations 2.20 and 2.22. In other words, the Fermi level moves into the CB for n+ silicon, and into VB for the p+ silicon. In addition, when the impurity concentration is higher than 10^{18} cm^{-3}, the donor (or acceptor) levels broaden into bands. This results in an effective decrease in the ionization energy until finally the impurity band merges with the CB (or VB) and the ionization energy becomes zero. Under these circumstances, the silicon is said to be *degenerate*. Strictly speaking, Fermi statistics should be used for the calculation of electron concentration when $\left(E_c - E_f\right) \leq kT$ [20]. For practical purposes, it is a good approximation within a few kT to assume that the Fermi level of the degenerate $n+$ silicon is at the CB edge, and that the degenerate $p+$ silicon is at the VB edge.

2.2.5 Carrier Transport in Semiconductors

In thermal equilibrium, mobile (CB) electrons are in random thermal motion with an average velocity of thermal motion, $v_{th} \cong 1 \times 10^7$ cm sec^{-1} at 300° K. However, due to the random thermal motion of electrons, no net current flows through the material. On the other hand, in the presence of an electric field E, electrons move opposite to the direction of E. This process is called *electron drift* and causes a net current flow through the material. Also, if there is a carrier concentration gradient in the material, the carriers diffuse away from the region of higher concentration to the lower concentration, producing a net current flow in the semiconductor. Thus, the carrier transport or current flow in a semiconductor is the result of two different mechanisms: (1) the drift of carriers (electrons and holes), which is caused by the presence of an electric field and (2) the diffusion of carriers, which is caused by an electron or hole concentration gradient in the semiconductor. We will now consider factors involved in both phenomena.

2.2.5.1 Carrier Mobility and Drift Current

When an electric field is applied to a conducting medium containing free carriers, the carriers are accelerated in proportion to the force of the field. However, the accelerating carriers within a semiconductor will collide with various scattering centers including the atoms of the host lattice (lattice scattering), the impurity atoms (impurity scattering), and other carriers (carrier–carrier scattering). In the case of an electron, these different scattering mechanisms tend to redirect its momentum and in many cases tend to dissipate the energy gained from the electric field. Thus, under the influence of a uniform electric field, the process of energy gain from the field and energy loss due to the scattering balance each other and carriers attain a constant average velocity, called the *drift velocity* (v_d). At low electric fields, v_d is proportional to the electric field strength E and is given by

$$v_d = \mu E \tag{2.28}$$

where:

μ is the constant of proportionality and is called the mobility of the carriers in units of cm^2 V^{-1}sec^{-1}

The mobility is proportional to the time interval between collisions and inversely proportional to the effective mass of the carriers. The total mobility is determined by combining the mobilities for different scattering mechanisms such as mobility due to *lattice scattering* μ_L and mobility due to *ionized impurity* scattering μ_I. Assuming different scattering mechanisms are independent, we can write the expression for total mobility using Mathiessen's rule

$$\frac{1}{\mu} = \frac{1}{\mu_L} + \frac{1}{\mu_I} + \ldots \tag{2.29}$$

The measurement data show that the electron mobility (μ_n) in an n-type silicon is about three times the hole mobility (μ_p) in a p-type silicon since the effective mass of electrons in the CB is much lighter than that of holes in the VB.

Carrier mobility in bulk silicon is a function of the doping concentrations. Figure 2.5 shows plots of electron and hole mobilities in silicon as a function of doping concentration at room temperature. It is observed from the plots that at low impurity levels, the mobilities are mainly limited by carrier collisions with the silicon lattice or acoustic phonons. As the doping concentration increases beyond 1×10^{15} cm^{-3}, the mobilities decrease due to the increase in the collisions with the charged (ionized) impurity atoms through Coulomb interaction. At high temperatures, the mobility tends to be limited by lattice scattering and is proportional to $T^{-3/2}$, relatively insensitive to the doping concentration. At low temperatures, the mobility is higher; however, it strongly depends on doping concentration as it becomes more limited by

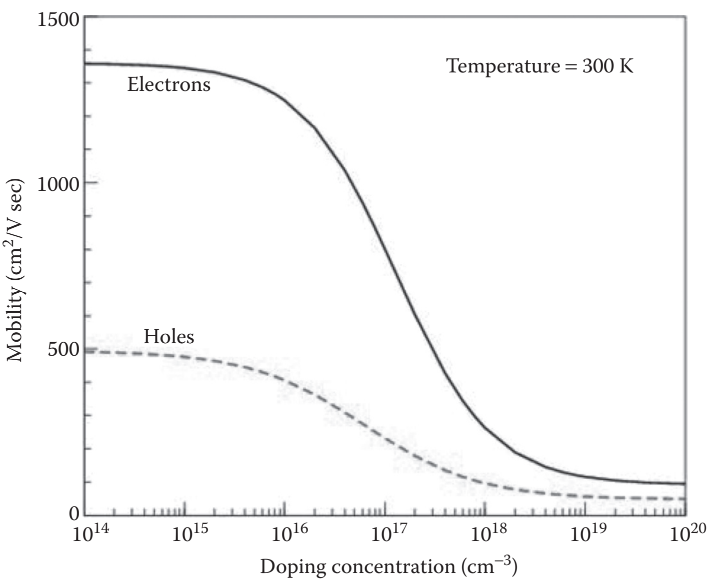

FIGURE 2.5
Electron and hole mobilities in bulk silicon at 300 K as a function of doping concentration.

impurity scattering. The detailed temperature dependence of mobility can be found in Arora and Arora et al. [17,21].

The carrier mobility discussed earlier is the *bulk mobility* applicable to conduction in the silicon substrate far away from the surface. In the channel region of MOSFET (metal-oxide-semiconductor field-effect transistor) devices, the current flow is governed by the *surface mobility*. The surface mobility is much lower than the bulk mobility due to additional scattering mechanism between the carriers and Si/SiO_2 interface in the presence of the high electric field normal to the channel as discussed in Section 5.3.1 of Chapter 5.

2.2.5.2 Electrical Resistivity

The drift of charge carriers under an applied electric field E results in a current, called the *drift current*. For a homogeneous n-type silicon, if there are n number of electrons per unit volume each carrying a charge q flow with a drift velocity v_d, then the *electron drift current* density is given by

$$J_{n,drift} = qnv_d = qn\mu_n E \tag{2.30}$$

where we have used Equation 2.28 for v_d; in Equation 2.30, $q = 1.6 \times 10^{-19}$ C is the electronic charge and μ_n is the electron mobility. From Ohm's law, the resistivity ρ of a conducting material is defined by E/J_n; therefore, from Equation 2.30, the resistivity ρ_n to electron current flow is given by

$$\rho_n = \frac{1}{qn\mu_n} \tag{2.31}$$

Similarly, for a p-type silicon, the hole drift current density, $J_{p,drift}$, and resistivity, ρ_p are given by

$$J_{p,drift} = qpv_d = qp\mu_p E \tag{2.32}$$

$$\rho_p = \frac{1}{qp\mu_p} \tag{2.33}$$

where:

μ_p is the hole mobility

If the silicon is doped with both donors and acceptors, then the total resistivity can be expressed as

$$\rho = \frac{1}{qn\mu_n + qp\mu_p} \tag{2.34}$$

Thus, the resistivity of a semiconductor depends on the electron and hole concentrations and their mobilities. Empirical resistivity versus impurity

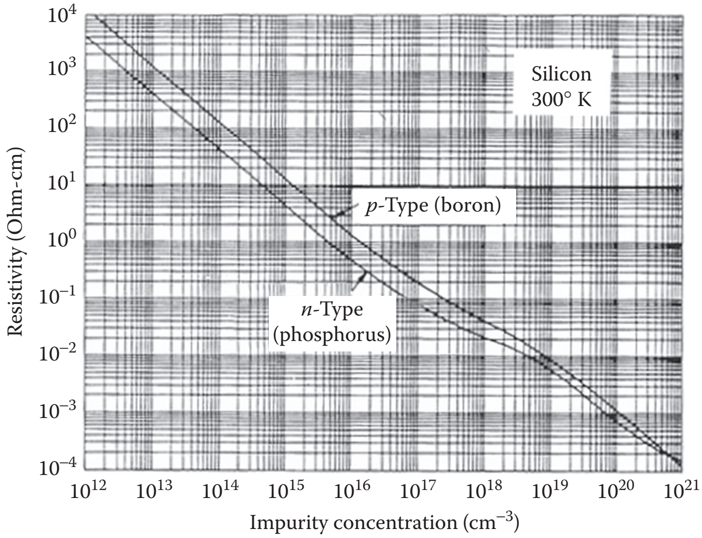

FIGURE 2.6
Impurity concentration versus resistivity for *n*-type and *p*-type silicon at 300° K. (After Sze 2007.)

concentrations plots are shown in Figure 2.6 for uniformly doped silicon at 300° K. The plot for *n*-type is lower than *p*-type doped silicon because electron mobility is higher than the hole mobility.

2.2.5.3 Sheet Resistance

The resistance of a uniform conductor of length L, width W, and thickness t is given by

$$R = \rho \frac{L}{tW} \tag{2.35}$$

where:
 ρ is the resistivity of the conductor in ohm-centimeter

Typically, in an IC technology, the thickness t of a diffusion region is uniform and normally much less than both L and W of the region. Therefore, it is useful to define a new variable ρ_{sh}, called the *sheet resistance*, which has the dimension of Ohm (Ω) and is given by

$$\rho_{sh} = \frac{\rho}{t} \tag{2.36}$$

Then Equation 2.35 becomes

$$R = \rho_{sh} \frac{L}{W} \tag{2.37}$$

From Equation 2.37, it is found that when $L = W$, the diffused layer becomes a square with $R = \rho_{sh}$. Thus, the total resistance of a diffusion line is simply ρ_{sh} times the number of squares in the path of current and is expressed in units of Ω per square (Ω/□). The process parameters that determine the sheet resistance of a layer are the resistivity and thickness t of the layer. Since the resistivity is a function of carrier concentration and mobility, both of which are functions of temperature, ρ_{sh} is temperature dependent.

2.2.5.4 Velocity Saturation

The field versus velocity linear relationship, given by Equation 2.28 in Section 2.2.5.1, is valid only for low electric field ($<1 \times 10^4$ V cm^{-1}) and carriers are in equilibrium with the lattice. At higher electric fields, the average carrier energy increases and carriers lose their energy by optical-phonon emission nearly as fast as they gain it from the field. This causes a decrease in μ from its low field value as the field increases until finally the drift velocity reaches a limiting value v_{sat}, referred to as the *saturation velocity*. This phenomenon is called the *velocity saturation*. For silicon, a typical value of $v_{sat} = 1.07 \times 10^7$ cm sec^{-1} for electrons and occurs at an electric field of about 2×10^4 V cm^{-1}. The corresponding values for holes are $v_{sat} = 8.34 \times 10^6$ cm sec^{-1} and $E \cong 5.0 \times 10^4$ V cm^{-1}.

It is found that the measured value of drift velocity for electrons and holes in silicon is a function of the applied field E and can be approximated by the following expression

$$v_d = v_{sat} \frac{E/E_c}{\left[1+\left(E/E_c\right)^{\beta}\right]^{1/\beta}} \tag{2.38}$$

where:

E_c is the critical electric field at which carrier velocity saturates

The parameters v_{sat}, E_c, and β in Equation 2.38 are given in Table 2.2.

Figure 2.7 shows the simulated value of drift velocity for electrons and holes at 300° K in silicon as a function of the applied field E obtained by Equation 2.38. It is observed from Figure 2.7 that at low fields, the carrier

TABLE 2.2

Parameters for Field Dependence of Drift Velocity for Silicon at 300 K

Parameter	v_{sat} (cm sec^{-1})	E_c (V cm^{-1})	β
Electrons	1.07×10^7	6.91×10^3	1.11
Holes	8.34×10^6	1.45×10^4	2.637

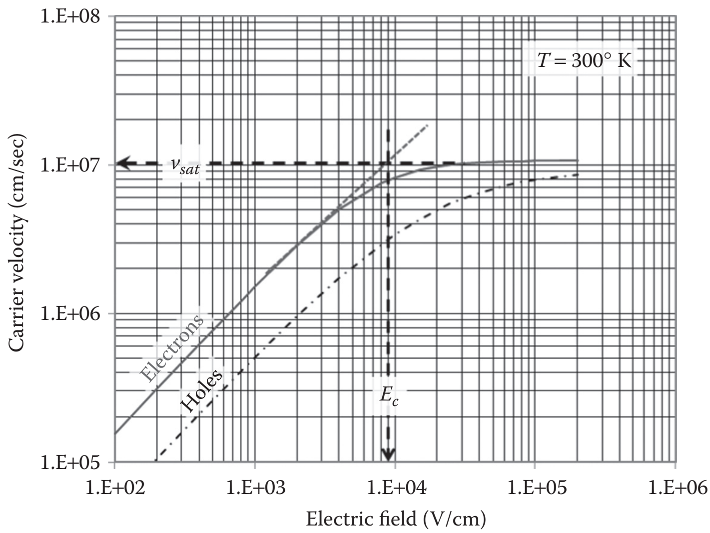

FIGURE 2.7
Drift velocities of electrons and holes in silicon at room temperature as a function of applied electric field showing velocity saturation at high electric fields.

velocity increases linearly with the electric field indicating constant mobility. When the field exceeds about 2×10^4 V cm^{-1}, carriers begin to lose energy by scattering with optical phonons and their velocity saturates. As the field exceeds 100 KV cm^{-1}, carriers gain more energy from the field than what they can lose by scattering. Consequently, their energy with respect to the bottom of the CB (for electrons) or top of the VB (for holes) begins to increase. The carriers are no longer at thermal equilibrium with the lattice. Since they acquire energy higher than the thermal energy (kT) they are called hot carriers.

It is these hot carriers that are responsible for reducing the mobility at high fields. For a more heavily doped material, the low-field mobility is lower because of the impurity scattering. However, v_{sat} remains the same, independent of impurity scattering. Also, v_{sat} is weakly dependent on temperature and decreases slightly as the temperature increases [17]. Figure 2.7 shows carrier velocity as a function of electric field. It is observed from the plots that the carrier velocity increases linearly at low electric field, then the increase in the carrier velocity slows down with the increase in electric field, and finally above a certain critical electric field the carrier velocity saturates.

2.2.5.5 Diffusion of Carriers

In addition to the drift of electrons under the influence of an electric field, the carriers also diffuse if the carrier concentration is not uniform within

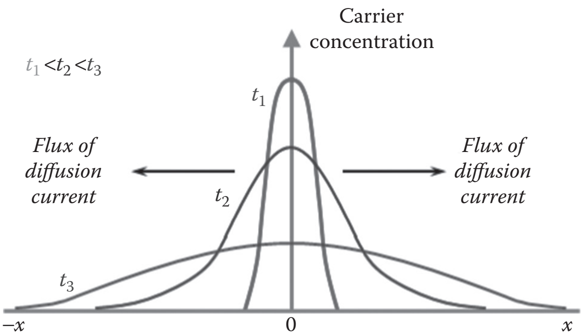

FIGURE 2.8
Diffusion of carriers from high concentration to low concentration due to concentration gradient over different time intervals $t_1 < t_2 < t_3$; t_1 is the initial time and the background concentration ≈ 0.

a semiconductor. This leads to an additional component of current in proportion to the concentration gradient and is called the *diffusion current*. Thus, the diffusion is a gradient driven motion and occurs from high-concentration regions toward low-concentration regions as shown in Figure 2.8.

The diffusion flux is given by Fix's first law,

$$F = -D\frac{dC}{dx} \tag{2.39}$$

where:

F, D, and C are the flux of carriers, diffusion constant, and carrier density, respectively

The negative sign is due to the fact that the carriers flow from the higher concentration to lower concentration; that is, dC/dx is negative. If the carrier flow in a semiconductor material is electron, then the diffusion current flow due to the electron concentration gradient dn/dx is given by

$$J_{n,diff} = qD_n\frac{dn}{dx} \tag{2.40}$$

Similarly, the hole diffusion current due to hole concentration gradient dp/dx is given by

$$J_{p,diff} = -qD_p\frac{dp}{dx} \tag{2.41}$$

where:

D_n and D_p are called the *diffusivity* or *diffusion constants* for electrons and holes in the material, respectively and are related to the respective mobility by the relationship [6]

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = \frac{kT}{q} \equiv v_{kT} \tag{2.42}$$

where:
$v_{kT} \equiv kT/q$ is called the thermal voltage

Equation 2.42 is often referred to as the Einstein's relation. For lightly doped silicon (e.g., $N_d \cong 1 \times 10^{15}$ cm^{-3}) at room temperature, $D_n = 38$ cm^2 sec^{-1} and $D_p = 13$ cm^2 sec^{-1}. The negative sign in Equation 2.41 implies that the hole current flows in a direction opposite to the hole concentration gradient.

2.2.5.6 Nonuniformly Doped Semiconductors and Built-In Electric Field

Let us consider an n-type material with nonuniformly doped N_d donor atoms as shown in Figure 2.9. Considering complete ionization of donor atoms, we have $n = N_d^+ = N_d$.

Due to the concentration gradient, electrons diffuse from the high-concentration region to the low-concentration region. Then from Equation 2.39 the diffusion flux of electrons is given by

$$F_{n,diff} = -D_n \frac{dn(x)}{dx} \tag{2.43}$$

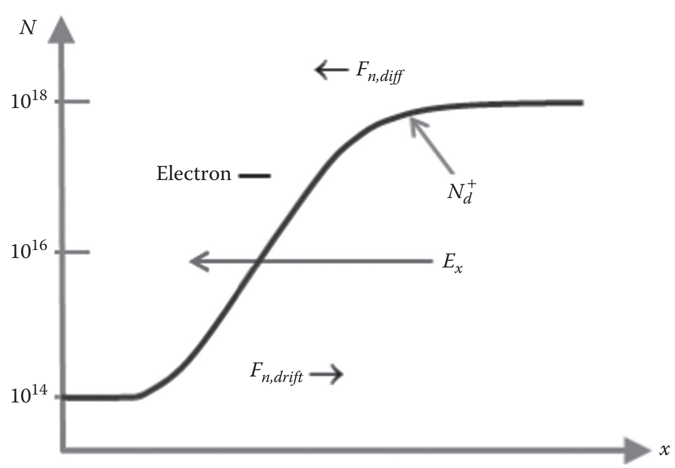

FIGURE 2.9
Drift and diffusion of carriers in a nonuniformly doped n-type semiconductor: $F_{n,diff}$ is the electron diffusion flux from the high concentration to low concentration; $F_{n,drift}$ is the drift flux of electrons due to the built-in electric field, E_x set up by the ionized donors and diffused electrons in the semiconductor.

where:
the subscript n represents the parameters for electrons

As the electrons move (diffuse) away, they leave behind positively charged donor ions (N_d^+), which try to pull electrons back causing drift flux of electrons from the low- to high-concentration region. This drift of electrons from low- to high-concentration regions sets up an electric field, E_x from the high-concentration to the low-concentration regions as shown in Figure 2.9. Then from Equation 2.30, the flux due to the drift of electrons is given by

$$F_{n,drift} = n(x)v_d = n\mu_n E_x \tag{2.44}$$

An equilibrium is established when *diffusion* = *drift*. Here $n(x)$ is the number of electrons in the diffusion flux at any point x in the distribution and $\neq N_d(x)$. Therefore, a *built-in* electric field is established that prevents diffusion of electrons. Then from Equations 2.43 and 2.44, we get the expression for the built-in electric field for electrons in an n-type nonuniformly doped substrate as

$$E_x = -\frac{D_n}{\mu_n}\frac{1}{n}\frac{dn(x)}{dx} = -v_{kT}\frac{1}{n}\frac{dn(x)}{dx} \tag{2.45}$$

Similarly, the built-in electric field for holes in a nonuniform p-type substrate is given by

$$E_x = \frac{D_p}{\mu_p}\frac{1}{p}\frac{dp(x)}{dx} = v_{kT}\frac{1}{p}\frac{dp(x)}{dx} \tag{2.46}$$

In Equations 2.45 and 2.46 we have used Einstein's relation given in Equation 2.42. This built-in electric field favors the transport of the minority carriers if created by an external source.

2.2.6 Generation–Recombination

In a semiconductor under thermal equilibrium, carriers possess an average thermal energy corresponding to the ambient temperature. This thermal energy excites some valence electrons to reach the CB. This upward transition of an electron from the VB to CB leaves behind a hole in the VB and an electron–hole pair is created. This process is called the *carrier generation* (G). On the other hand, when an electron makes a transition from the CB to the VB, an electron–hole pair is *annihilated*. This reverse process is called carrier *recombination* (R). Under thermal equilibrium, G = R so that the carrier concentration remains the same and the condition $pn = n_i^2$ is maintained. The thermal G–R process is shown in Figure 2.10.

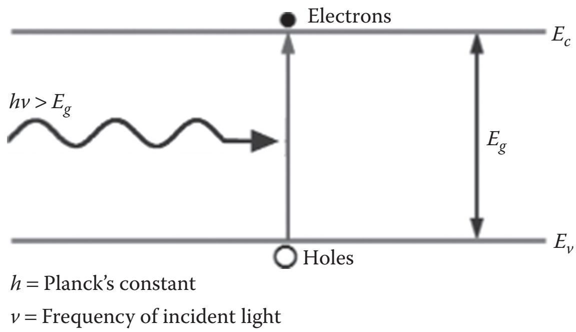

FIGURE 2.10
Band-to-band generation of electron–hole pairs under optical illumination of photon energy $h\nu$, where h and ν are the Planck's constant and the frequency of incident light, respectively.

The equilibrium condition of a semiconductor is disturbed by optically or electrically introducing free carriers exceeding their thermal equilibrium values resulting in $pn > n_i^2$ or by electrically removing carriers resulting in $pn < n_i^2$. The process of introducing carriers in access of thermal equilibrium values is called the *carrier injection* and the additional carriers are called the *excess carriers*. In order to inject excess carriers optically, we shine light with energy $E = h\nu > E_g$ on an intrinsic semiconductor so that the valence electrons can be excited into the CB by the excess energy $\Delta E = (h\nu – E_g)$, where h and ν are Planck's constant and frequency of light, respectively. In this process, we get optically generated excess electrons (n_L) and holes (p_L) in the semiconductor as shown in Figure 2.10. Therefore, the total nonequilibrium values of carrier concentration is given by

$$\left.\begin{aligned} n &= n_i + n_L \\ p &= n_i + p_L \end{aligned}\right\} \quad \text{Injection of carriers by light} \tag{2.47}$$

2.2.6.1 Injection Level

From Equation 2.47, we observe that both n and p are greater than the intrinsic carrier concentration of the semiconductor, and therefore, $pn > n_i^2$ for injection of carriers into the semiconductor. If the injected carrier density is lower than the majority carrier density at equilibrium so that the latter remains essentially unchanged while the minority carrier density is equal to the excess carrier density, then the process is called the *low-level injection*. If the injected carrier density is comparable to or exceeds the equilibrium value of the majority carrier density, then it is called the *high-level* injection.

To illustrate the injection levels, we consider an n-type *extrinsic* semiconductor with $N_d = 10^{15}$ cm^{-3}. Then from Section 2.2.4.1, the equilibrium majority carrier electron concentration is given by $n_{no} = 1 \times 10^{15}$ cm^{-3}, whereas from Equation 2.21, the minority carrier hole concentration is given by

$p_{no} = 1 \times 10^5$ cm^{-3}. Here, n_{no} and p_{no} define the equilibrium concentrations of electrons and holes, respectively, in an n-type material. Now, we shine light on the sample so that 1×10^{13} cm^{-3} electron–hole pairs are generated in the material. Then using Equation 2.47, the total number of electrons $n_n = n_{no} = 1 \times 10^{15}$ cm^{-3} and $p_n = 1 \times 10^{13}$ cm^{-3}. Thus, the majority carrier concentration n_n remains unchanged, whereas the minority carrier concentration p_n is increased significantly. This is an example of low-level injection. On the other hand, if 1×10^{17} cm^{-3} electron–hole pairs are generated by incident light, then from Equation 2.47, we get $n_n \cong 1 \times 10^{17}$ cm^{-3} and $p_n = 1 \times 10^{17}$ cm^{-3} changing both the electron and hole concentrations in the semiconductor, resulting in a high-level injection. *The mathematics for high-level injection are complex, and therefore, we will consider only low-level injection.*

2.2.6.2 Recombination Processes

The semiconductor material returns to equilibrium through recombination of injected minority carriers with the majority carriers in the case of carrier injection or through generation of electron–hole pairs in the case of extraction of carriers.

The electron–hole recombination process occurs by transition of electrons from the CB to the VB. In a direct bandgap semiconductor like GaAs where the minimum of the CB aligns with the maximum of the VB, an electron in the CB can give up its energy to move down to occupy the empty state (hole) in the VB without a change in the momentum as shown in Figure 2.11a. Since the momentum (k) must be conserved in any energy level transition, an electron in GaAs can easily make direct transition from E_c to E_v across E_g. This is called the *direct* or *band-to-band recombination*. When direct recombination happens, the energy given up by electron will be emitted as a photon, which makes it useful for light-emitting diodes.

If we generate excess carriers (Δn, Δp) at a rate G_L due to the incident light, then for low-level injection, we get $\Delta p = \Delta n = U\tau = G_L\tau$, where U is the net recombination rate and τ is the excess carrier lifetime. If p_o and n_o are the equilibrium concentrations of electrons and holes, respectively, and p and

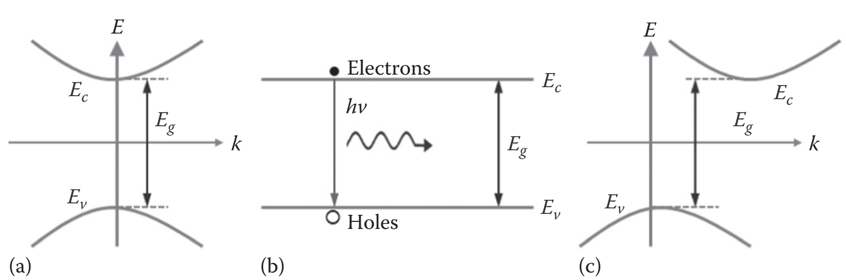

FIGURE 2.11
Bandgap in semiconductors: (a) direct bandgap, (b) band-to-band recombination in a direct bandgap semiconductor, and (c) indirect bandgap.

n are the respective total concentrations due to generation, then $\Delta p = p - p_o$ and $\Delta n = n - n_o$ and the net recombination rate due to direct recombination is given by

$$U = \frac{\Delta n}{\tau_n} = \frac{\Delta p}{\tau_p} \tag{2.48}$$

where:

τ_n and τ_p are the excess carrier electron and hole lifetime, respectively

For band-to-band recombination, the excess carrier lifetime for an electron is equal to that of a hole since the single phenomenon annihilates an electron and a hole simultaneously.

For *indirect bandgap* semiconductors such as silicon and germanium (Figure 2.11c), the probability of direct recombination is very low. Physically, this means that the minimum energy gap between E_c and E_v does not occur at the same point in the momentum space as shown in Figure 2.11c. In this case, for an electron to reach the VB, it must experience a change of momentum as well as energy to satisfy the conservation principle. This can be achieved by recombination processes through intermediate trapping levels, called the *indirect recombination* as shown in Figure 2.12.

Impurities that form electronic states deep in the energy gap assist the recombination of electrons and holes in the indirect bandgap semiconductors. Here the word *deep* indicates that the states are far away from the band edges and near the center of the energy gap. These deep states are commonly referred to as *recombination centers* or *traps*. Such recombination centers are usually unintentional impurities, which are not necessarily ionized at room temperature. These deep level impurities have concentrations far below the concentration of donor or acceptor impurities, which have shallow energy levels. Gold (Au) is a deep level impurity intentionally used in silicon to increase the recombination rate. This recombination via deep level impurities or traps is often referred to as the *indirect recombination*. The process shown in Figure 2.12 consists of (1) an electron capture by an empty center, (2) electron emission from an occupied center, (3) hole capture by an occupied center, and (4) hole emission by an empty center.

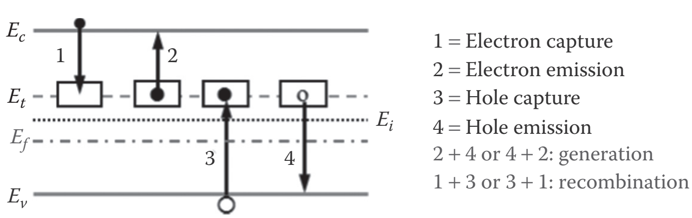

FIGURE 2.12
Generation and recombination in an indirect bandgap semiconductor; E_t is the trap level deep into the bandgap; 1, 2, 3, and 4 represent the generation and recombination processes.

Let us consider the following example where an impurity like Au is introduced that provides a *trapping level* or a set of allowed states at energy E_t. The trap level E_t is assumed to act like an acceptor (it can be neutral or negatively charged). Recombination is accomplished by trapping an electron and a hole. (The analysis can be easily extended to the case where the trap acts like a donor, that is, positively charged or neutral charge states.) The indirect recombination process was originally proposed by Shockley and Read [22] and independently suggested by Hall [23] and, therefore, is often referred to as the *Shockley–Read–Hall* (SRH) recombination. By considering the transition processes shown in Figure 2.12, Shockley, Read, and Hall showed that for low-level injection, the net recombination rate is given by

$$U = \frac{v_{th}\sigma N_t\left(pn - n_i^2\right)}{n + p + 2n_i \cosh\left[\left(E_t - E_i\right)/kT\right]} \tag{2.49}$$

where:

v_{th} is the carrier thermal velocity ($\approx 1 \times 10^7$ cm sec^{-1})
σ is the carrier capture cross section ($\approx 10^{-15}$ cm^2)
N_t is the density of trap centers
$v_{th}\sigma N_t$ is the *capture probability* or capture cross section

From Equation 2.49 we observe the following:

1. The "driving force" or the rate of recombination is proportional to $\left(pn - n_i^2\right)$, that is, the deviation from the equilibrium condition
2. $U = 0$ when $\left(np = n_i^2\right)$, that is, equilibrium condition
3. U is maximum when $E_t = E_i$, that is, trap levels near the mid-band are the most efficient recombination centers

Thus, for the simplicity of understanding, let us consider the case when $E_t = E_i$. Then from Equation 2.49, the net recombination rate is given by

$$U = \frac{v_{th}\sigma N_t\left(pn - n_i^2\right)}{n + p + 2n_i} \tag{2.50}$$

For an *n-type* semiconductor with low-level injection, $n >> p + 2n_i$; denoting $p = p_n$ as the total excess minority carrier concentration and $\left(p_{no} = n_i^2/n\right)$ as the equilibrium minority carrier concentration, we get after simplification of Equation 2.50

$$U = v_{th}\sigma N_t\left(p_n - p_{no}\right) = \frac{\Delta p}{\tau_p} \tag{2.51}$$

where the minority carrier hole lifetime in an n-type semiconductor is given by

$$\tau_p = \frac{1}{v_{th}\sigma_p N_t} \tag{2.52}$$

In an n-type material, lots of electrons are available for capture. Therefore, Equation 2.51 shows that the minority carrier hole lifetime τ_p is the limiting factor in recombination process in an n-type material.

Similarly, for a *p-type* semiconductor, we can show from Equation 2.50 that the net recombination rate for electrons is given by

$$U = \frac{\Delta n}{\tau_n} \tag{2.53}$$

where

$$\tau_n = \frac{1}{v_{th}\sigma_n N_t} \tag{2.54}$$

is the minority carrier electron lifetime. Thus, for a p-type semiconductor the minority carrier electron lifetime is the limiting factor in the recombination process.

The other recombination process in silicon that does not depend on deep level impurities and that sets an upper limit on lifetime is *Auger recombination*. In this process, the electrons and holes recombine without trap levels and the released energy (of the order of energy gap) is transferred to another majority carrier (a hole in a p-type and electron in an n-type silicon). Usually, Auger recombination is important when the carrier concentration is very high ($>5 \times 10^{18}$ cm^{-3}) as a result of high doping or high-level injection.

2.2.7 Basic Semiconductor Equations

2.2.7.1 Poisson's Equation

Poisson's equation is a very general differential equation governing the operation of IC devices and is based on Maxwell's field equation that relates the charge density to the electric field potential. Conventionally, the electrostatic potential, ϕ in a semiconductor is defined in terms of the intrinsic Fermi level (E_i) such that

$$\phi = -\frac{E_i}{q} \tag{2.55}$$

The negative sign in Equation 2.55 is due to the fact that E_i is defined as the electron energy whereas ϕ is defined for a positive charge. The electric field E, which is defined as the electrostatic force per unit charge, is equal to the negative gradient of ϕ, such that

$$E = -\frac{d\phi}{dx} \tag{2.56}$$

Mathematically, Poisson's equation (for silicon) is stated as

$$\frac{dE}{dx} = \frac{\rho(x)}{K_{si}\varepsilon_0} \tag{2.57}$$

or, using Equation 2.56,

$$\frac{d^2\phi}{dx^2} = -\frac{\rho(x)}{K_{si}\varepsilon_0} \tag{2.58}$$

where

$\rho(x)$ is the net charge density at any point x

ε_0 (=8.854 × 10^{-14} F cm^{-1}) is the permittivity of free space

K_{si} (=11.8) is the relative permittivity of silicon

If n and p are the free electron and hole concentrations, respectively, corresponding to N_d^+ and N_a^- ionized acceptor and donor concentrations, respectively, in silicon, we can express Equation 2.58 as

$$\frac{d^2\phi}{dx^2} = -\frac{dE}{dx} = -\frac{q}{K_{si}\varepsilon_0}\left\{\left[p(x)-n(x)\right]+\left[N_d^+(x)-N_a^-(x)\right]\right\} \tag{2.59}$$

Assuming complete ionization of dopants, $N_d^+ = N_d$ and $N_a^- = N_a$, we can write Poisson's equation as

$$\frac{d^2\phi}{dx^2} = -\frac{q}{K_{si}\varepsilon_0}\left\{\left[p(x)-n(x)\right]+\left[N_d(x)-N_a(x)\right]\right\} \tag{2.60}$$

Equation 2.60 is a one-dimensional (1D) equation and can easily be extended to three-dimensional (3D) space. 1D-Poisson equation is adequate for describing most of the basic device operations. However, for small geometry advanced devices 2D (two-dimensional) or 3D Poisson's equation must be used.

Another form of Poisson's equation is Gauss's law, which is obtained by integrating Equation 2.57:

$$E = \frac{1}{K_{si}\varepsilon_0}\int \rho(x)dx = \frac{Q_s}{K_{si}\varepsilon_0} \tag{2.61}$$

It is to be noted that the semiconductor as a whole is charged neutral, that is, ρ must be zero. However, when the space charge neutrality does not apply, Poisson's equation must be used.

2.2.7.2 Carrier Concentration in Terms of Electrostatic Potential

In an n-type nondegenerate semiconductor the Fermi level E_f (or Fermi potential $\phi_f = -E_f/q$) lies above the intrinsic level E_i (or intrinsic potential $\phi_i = -E_i/q$) as shown in Figure 2.4c. Then from Equation 2.26 we can write

$$N_d = n_i \exp\left(\frac{E_f - E_i}{kT}\right) = n_i \exp\left[\frac{q}{kT}\left(\phi_i - \phi_f\right)\right] \tag{2.62}$$

while in a p-type semiconductor the Fermi level E_f (or Fermi potential ϕ_f) lies below the intrinsic level E_i (or intrinsic potential ϕ_i) as shown in Figure 2.4d, and from Equation 2.27 we can show

$$N_a = n_i \exp\left(\frac{E_i - E_f}{kT}\right) = n_i \exp\left[\frac{q}{kT}\left(\phi_f - \phi_i\right)\right] \tag{2.63}$$

At room temperature, the available thermal energy is sufficient to ionize nearly all acceptor and donor atoms due to their low ionization energies. Hence it is safe approximation to say that in a nondegenerate silicon at room temperature:

$$n \approx N_d\left(n - \text{type}\right) \tag{2.64}$$

$$p \approx N_a\left(p - \text{type}\right) \tag{2.65}$$

where:

N_d is the concentration of donor atoms
N_a is the concentration of acceptor atoms

In an n-type material, where $N_d >> n_i$, electrons are majority carriers whose concentration is given by Equation 2.64, while the hole concentration p_n (representing concentration of p in an n-type material) from Equation 2.64 is given by

$$p_n \cong \frac{n_i^2}{N_d} \tag{2.66}$$

The hole concentration p_n is much smaller than n_n in an n-type semiconductor. Thus, holes are minority carriers in an n-type semiconductor. Similarly, in a p-type semiconductor where $N_a >> n_i$, holes are the majority carriers given by Equation 2.65, while the minority carrier electron concentration is given by

$$n_p \cong \frac{n_i^2}{N_a} \tag{2.67}$$

Since $n_p << p$, electrons are minority carriers in a p-type semiconductor. Consequently, we often use the terminology of majority and minority carriers.

From Equation 2.62, we can write for an n-type semiconductor

$$\phi_i - \phi_f = \frac{kT}{q}\ln\left(\frac{N_d}{n_i}\right) = v_{kT}\ln\left(\frac{N_d}{n_i}\right) \equiv -\phi_B \tag{2.68}$$

where:

$\phi_B \equiv (\phi_f - \phi_i)$ is called the *bulk potential* and is negative for n-type semiconductors

Similarly, from Equation 2.63, for p-type semiconductor, we can show

$$\phi_f - \phi_i = v_{kT}\ln\left(\frac{N_a}{n_i}\right) \equiv \phi_B \tag{2.69}$$

Thus, we can write a generalized expression for bulk potential in semiconductors as

$$\phi_B = \left(\phi_i - \phi_f\right) = \pm v_{kT}\ln\left(\frac{N_b}{n_i}\right) \tag{2.70}$$

where:

the "+" sign is for p-type semiconductors with $N_b = N_a$
the "–" sign is for n-type semiconductors with $N_b = N_d$

Note that the Fermi potential, ϕ_f, is not only a function of carrier concentration but also dependent on temperature through n_i. From Equation 2.70, we observe that since n_i increases with temperature according to Equation 2.15, the magnitude of ϕ_B decreases and as n_i approaches to N_b, ϕ_f approaches to ϕ_i. Thus, with an increase of temperature, the Fermi level approaches the mid-gap position, that is, the intrinsic Fermi level, showing thereby that the semiconductor becomes intrinsic at high temperature. Thus, the doped or extrinsic silicon will become intrinsic if the temperature is high enough. The temperature at which this happens depends upon the dopant concentration. When the material becomes intrinsic, the device can no longer function, and therefore, the intrinsic region is avoided in device operation.

The temperature coefficient of ϕ_f can be obtained by differentiating Equation 2.70 giving

$$\frac{d\phi_f}{dT} = \frac{1}{T}\left[\phi_f - \left(\frac{E_g}{2} + \frac{3}{2}v_{kT}\right)\right] \tag{2.71}$$

Equation 2.71 gives $d\phi_f/dT \sim 1$ mV K^{-1}. If we use Equation 2.15 for n_i, then ϕ_f with reference to $\phi_i = 0$ at any temperature T can be written in terms of T_{NOM} as

$$\phi_f(T) = \phi_f(T_{NOM}).\left(\frac{T}{T_{NOM}}\right) - v_{kT}\left\{\frac{3}{2}\ln\left(\frac{T}{T_{NOM}}\right) + \left[-\frac{E_g(T)}{2kT} + \frac{E_g(T_{NOM})}{2kT_{NOM}}\right]\right\} \quad (2.72)$$

Equation 2.72 is used in circuit CAD tools for modeling the temperature dependence of ϕ_f.

2.2.7.3 Quasi-Fermi Level

Under thermal equilibrium conditions, the electron and hole concentrations are given by Equations 2.62 and 2.63 (using $n = N_d$ and $p = N_a$), respectively, maintaining the condition $pn = n_i^2$. However, when carriers are injected into the semiconductor or extracted out from the semiconductor, the equilibrium condition is disturbed. In nonequilibrium conditions: (1) injection, $np > n_i^2$ or (2) extraction, $np < n_i^2$, we cannot use Equations 2.62 and 2.63. And, the carrier densities can no longer be described by a constant Fermi level through the system. Here, we define *quasi-Fermi* levels such that Equations 2.62 and 2.63 hold as given by

$$n = n_i \exp\left(\frac{E_{fn} - E_i}{kT}\right) = n_i \exp\left[\frac{q}{kT}(\phi_i - \phi_{fn})\right] \quad (2.73)$$

$$p = n_i \exp\left(\frac{E_i - E_{fp}}{kT}\right) = n_i \exp\left[\frac{q}{kT}(\phi_{fp} - \phi_i)\right] \quad (2.74)$$

where:
E_{fn} and E_{fp} are the electron and hole quasi-Fermi levels, respectively

It is to be noted that $\mathrm{E_{fn}}$ *and* $\mathrm{E_{fp}}$ *are the mathematical tools; their values are chosen so that the accurate carrier concentrations are given in the nonequilibrium situations.* In general, $E_{fn} \neq E_{fp}$.

From Equations 2.73 and 2.74, we can show

$$pn = n_i^2 \exp\left(\frac{E_{fn} - E_{fp}}{kT}\right) \quad (2.75)$$

In equilibrium condition, $E_{fn} = E_{fp} = E_f$ and $\phi_{fn} = \phi_{fp}$ so that Equations 2.73 and 2.74 become same as Equations 2.62 and 2.63 for $n = N_d$ and $p = N_a$, respectively. And, Equation 2.75 becomes $pn = n_i^2$.

2.2.7.4 Transport Equations

In Section 2.2.5.5, we have shown that the electron diffusion current density $J_{n,diff}$ due to concentration gradient in a semiconductor is given by Equation 2.40. On the other hand, the electron current density due to drift of electrons by an applied electric field described in Section 2.2.5.2 is given by Equation 2.30. Thus, when an electric field is present in addition to a concentration gradient, both the drift and diffusion current will flow through the semiconductor. The total electron current density J_n at any point x is then simply the sum of the diffusion and drift currents, that is, J_n $(=J_{n,drift} + J_{n,diff})$. Therefore, the total electron current in a semiconductor is given by

$$J_n = qn\mu_n E + qD_n \frac{dn}{dx} \tag{2.76}$$

Similarly, the total hole current density J_p $(=J_{p,drift} + J_{p,diff})$ is given by

$$J_p = qp\mu_p E - qD_p \frac{dp}{dx} \tag{2.77}$$

so that the total current density $J = J_n + J_p$. The current Equations 2.76 and 2.77 are often referred to as the transport equations.

Under thermal equilibrium no current flows inside the semiconductor and therefore, $J_n = J_p = 0$. However, under nonequilibrium conditions J_n and J_p can be written in terms of quasi-Fermi potentials ϕ_n and ϕ_p for electric field, E, in Equations 2.76 and 2.77, respectively, to get

$$\begin{aligned} J_n &= -qn\mu_n \frac{d\phi_n}{dx} \\ J_p &= -qp\mu_p \frac{d\phi_p}{dx} \end{aligned} \tag{2.78}$$

2.2.7.5 Continuity Equations

When carriers diffuse through a certain volume of semiconductor, the current density leaving the volume may be smaller or larger depending upon the recombination or generation taking place inside the volume. Let us consider a small length Δx of a semiconductor as shown in Figure 2.13 with cross-sectional area A in the yz plane.

From Figure 2.13, the hole current density entering the volume $A.\Delta x$ is $J_p(x)$ whereas the density leaving is $J_p(x + \Delta x)$. From the conservation of charge, the rate change of hole concentration in the volume is the sum of (1) net holes flowing out of the volume and (2) net recombination rate. That is,

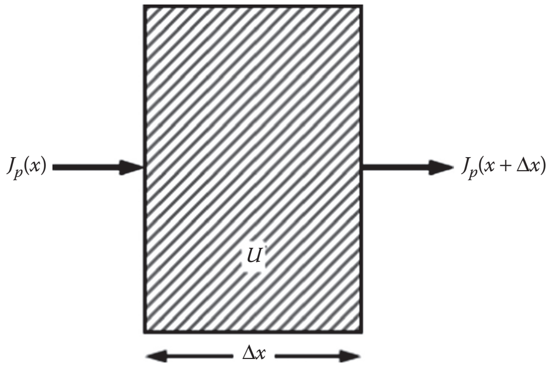

FIGURE 2.13
Current continuity in a semiconductor: J_p(x) is the hole currents flowing into an elemental length Δx of the semiconductor and $J_p(x + \Delta x)$ is the net current flowing out after carrier generation–recombination processes inside the element; U is the net recombination rate.

$$-\frac{\partial p}{\partial t}\Delta x = \left[\frac{1}{q}J_p(x+\Delta x) - \frac{1}{q}J_p(x)\right] + \left(G_p - R_p\right)\Delta x \tag{2.79}$$

The negative sign is due to the decrease of holes due to recombination; and G_p and R_p are the generation and recombination rate of holes in the volume, respectively. Then from Equation 2.79, we can show

$$-\frac{\partial p}{\partial t} = \frac{1}{q}\frac{\partial J_p}{\partial x} + \left(G_p - R_p\right) \tag{2.80}$$

Similarly, for electrons we can show

$$-\frac{\partial n}{\partial t} = -\frac{1}{q}\frac{\partial J_n}{\partial x} + \left(G_n - R_n\right) \tag{2.81}$$

where:
R_n and G_n are the recombination and generation rate of electrons, respectively

Equations 2.80 and 2.81 are called the *continuity equations* for holes and electrons, respectively, and describe the time-dependent relationship between current density, recombination and generation rates, and space. They are used for solving transient phenomena and diffusion with recombination–generation of carriers.

Equations 2.60, 2.78, 2.80, and 2.81 constitute a complete set of 1D equations to describe carrier, current, and field distributions in a semiconductor; however, they can easily be extended to 3D space. Given appropriate boundary conditions, we can solve them for any arbitrary device structure. Generally, we will be able to simplify them based on physical approximations.

2.3 Theory of *n*-Type and *p*-Type Semiconductors in Contact

We have discussed the basic theory of intrinsic, *n*-type, and *p*-type semiconductors in Section 2.2. In this section, we will discuss the underlying physics of a semiconductor substrate when one region is *n*-type and the immediate adjacent region is *p*-type, forming a junction called the *pn-junction* or *pn-junction diode* or simply *diode*. In reality, a silicon *pn*-junction is formed by counter doping a local region of a larger region of doped silicon as shown in Figure 2.14. The *pn* junctions form the basis for all advanced semiconductor devices. Therefore, understanding their operation is basic to the understanding of most advanced IC devices.

2.3.1 Basic Features of *pn*-Junctions

A silicon *pn*-junction structure is an alternating type of *p*-type and *n*-type doped silicon layers. The *pn*-junctions can be fabricated in a variety of techniques on a silicon substrate using photo mask → Implant → Drive-in. A typical final impurity profile along the active region can be simplified as an *erfc* or *Gaussian* as shown in Figure 2.14b and c.

As shown in Figure 2.14a, the basic structure includes an *n*-region doped on a *p*-type substrate. The vertical cross section of the intrinsic or active *pn*-junction is shown in Figure 2.14a by a vertical cutline *A*. The 1D-doping

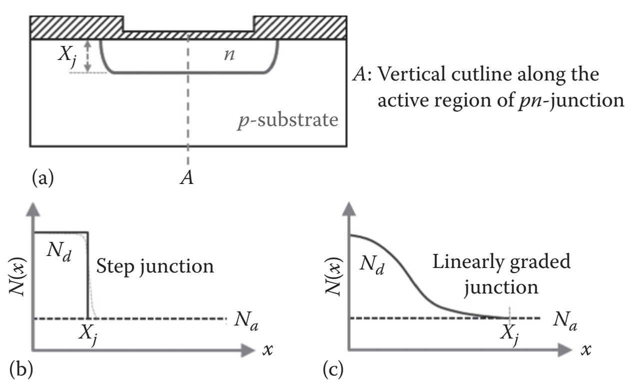

FIGURE 2.14
A typical *pn*-junction: (a) 2D cross section showing the cutline along the depth of the structure to obtain 1D doping profiles, (b) 1D-doping profile of an abrupt junction, and (c) 1D-doping profile of a graded junction.

profile along the cutline of the active device is shown in Figure 2.14b and c. The metallurgical junction depth X_j is indicated as the point where the net impurity concentrations of donors and acceptors are equal. For compact modeling, the actual impurity profile is approximated by a *step* or *abrupt* (high–low) shallow junctions, Figure 2.14b or a *linearly graded* (deep) junctions, Figure 2.14c, so that a tractable circuit model can be developed. A step doping profile is characterized by constant p-type dopant concentration N_a that changes with position in a stepwise fashion to a constant n-type dopant concentration N_d.

From the 1D impurity profiles in Figure 2.14b and c, we find that there is a large carrier concentration gradient at the junction resulting in carrier diffusion. Holes from the p-side diffuse into the n-side, leaving behind negatively charged acceptor ions $\left(N_a^-\right)$ and electrons from the n-side diffuse into the p-side leaving behind positively charged donor ions $\left(N_d^+\right)$. Consequently, a space charge region is formed (negative charge on the p-side and positive charge on the n-side), creating thereby an electric field E, and, hence, a potential difference as shown in Figure 2.15. The direction of the field (n-region to p-region) is such that it opposes further diffusion of carriers so that, in thermal equilibrium, the net flow of carriers is zero; that is, an electric field is set up, which tends to pull *electrons* and *holes* back to the original positions. The internal potential difference between the two sides of the junction is called the *built-in potential* or *barrier height*, ϕ_{bi}. The space charge region on two sides of the metallurgical junction is often called the *depletion region*, because the region is depleted of the free carriers.

Figure 2.16a shows the energy-band diagram of a p-type silicon and n-type silicon physically separated from each other. As discussed in Section 2.2.4, the Fermi level for an n-type silicon lies close to its CB, and for a p-type silicon lies close to its VB. Also, as we will show later, the Fermi level of a semiconductor is flat, that is, spatially constant, when there is no current flow in it. Therefore, as the p-type region and the n-type region are brought

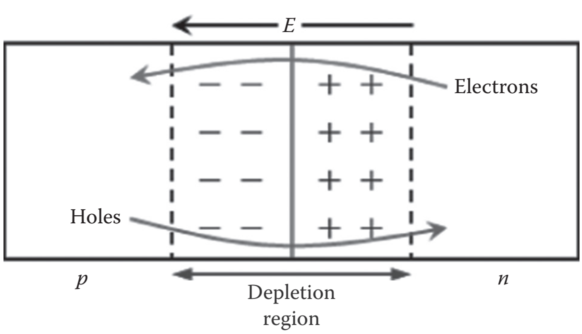

FIGURE 2.15
Formation of built-in electric field due to the space charges left behind by mobile carriers after diffusion from the high- to the low-concentration region on either side of the junction.

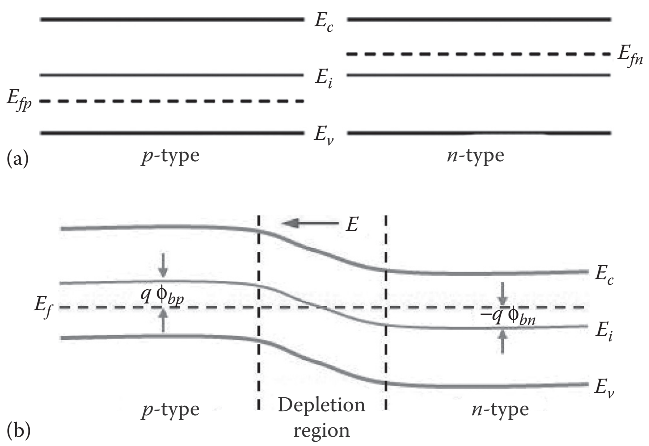

FIGURE 2.16
Energy band diagram of a *pn*-junction at equilibrium: (a) isolated *n*- and *p*-regions and (b) *p*-*n* regions are in contact to form a *pn*-junction.

together to form a *pn*-junction, the Fermi level must remain flat across the entire structure if there is no current flow in and across the junction. This causes the energy band bending, as shown in Figure 2.16b. The potential difference between the corresponding energy bands on the *p*- and *n*-sides is called the *built-in potential*, ϕ_{bi}, of the *pn*-junction as shown in Figure 2.16b.

2.3.2 Built-In Potential

In *pn*-junctions at equilibrium, the diffusion of carriers is balanced by the drift of carriers by the *built-in* electric field. To facilitate the description of both the *n*-side and the *p*-side of a *pn*-junction simultaneously, when necessary for clarity, we will distinguish the parameters on the *n*-side from the corresponding ones on the *p*-side by adding a subscript *n* to the symbols associated with the parameters on the *n*-side, and subscript *p* to the symbols associated with the parameters on the *p*-side. For example, E_{fp} and E_{fn} denote the Fermi level, respectively, on the *p*-side and *n*-side. Similarly, n_n and p_n denote the electron concentration and hole concentration, respectively, on the *n*-side, and n_p and p_p denote the electron concentration and hole concentration, respectively, on the *p*-side. Thus, n_n and p_p specify the majority carrier concentrations, while n_p and p_n specify the minority carrier concentrations.

Consider the *n*-side of a *pn*-junction at thermal equilibrium. If the *n*-side is nondegenerately doped to a concentration of N_d, then the separation between its Fermi level, which is flat across the junction, and its intrinsic Fermi level is given by Equations 2.62 and 2.63:

$$E_{fn} - E_{in} = kT \ln\left(\frac{N_d}{n_i}\right) = kT \ln\left(\frac{n_{no}}{n_i}\right) \equiv -q\phi_{bn}$$
$$E_{ip} - E_{fp} = kT \ln\left(\frac{N_a}{n_i}\right) = kT \ln\left(\frac{p_{po}}{n_i}\right) \equiv q\phi_{bp} \tag{2.82}$$

where:

n_{no} and p_{po} represent the equilibrium concentrations in the n-type and p-type semiconductors, respectively

Since at equilibrium, E_f is a constant across the pn-junction, that is, $E_{fp} = E_{fn}$, therefore, the built-in potential across the pn-junction is given by

$$q\phi_{bi} = E_{ip} - E_{in} = kT \ln\left(\frac{n_{no}p_{po}}{n_i^2}\right) \tag{2.83}$$

From pn-product equation, $n_{no}p_{no} = n_i^2 = n_{po}p_{po}$, therefore, Equation 2.83 can also be written as

$$\phi_{bi} = \phi_{bp} - \phi_{bn} = v_{kT} \ln\left(\frac{N_a N_d}{n_i^2}\right) = v_{kT} \ln\left(\frac{n_{no}p_{po}}{n_i^2}\right) \tag{2.84}$$

$$\phi_{bi} = \phi_{bp} - \phi_{bn} = v_{kT} \ln\left(\frac{n_{no}}{n_{po}}\right) = v_{kT} \ln\left(\frac{p_{po}}{p_{no}}\right) \tag{2.85}$$

Thus, ϕ_{bi} given by Equation 2.84 or 2.85 exists across a pn-junction without an applied bias at thermal equilibrium to counteract diffusion. The typical value of ϕ_{bi} is in between 0.5 and 0.9 V for silicon junctions and is strongly dependent on temperature due to dependence on n_i. And, ϕ_{bi} across a pn-junction increases as N_d or N_a increases.

2.3.3 Step Junctions

The analysis of pn-junction is much simpler if the junction is assumed to be abrupt, that is, the doping impurities are assumed to change abruptly from p-type on one side to n-type on the other side of the junction. The abrupt junction approximation is reasonable for modern VLSI (very-large-scale-integrated) devices, where the use of ion implantation for doping the junctions, followed by low thermal cycle diffusion and/or annealing, resulting in junctions that are fairly abrupt. Besides, the abrupt-junction approximation often leads to closed-form solutions for easier understanding of device physics.

2.3.3.1 Junction Potential and Electric Field

The analysis of an abrupt junction becomes even simpler in the depletion approximation in which the *pn*-junction is approximated by three regions as illustrated in Figure 2.17. Both the bulk *p*-region, that is, the region with $x < -x_p$, and the bulk *n*-region, that is, the region with $x > x_n$, are assumed to be charge neutral, while the transition region, that is, the region with $-x_p < x < x_n$, is assumed to be depleted of mobile electrons and holes. The width W_d of the depletion region can be obtained by solving Poisson's equation 2.60 as repeated below:

$$\frac{d^2\phi}{dx^2} = -\frac{q}{K_{si}\varepsilon_0}\left\{\left[p(x)-n(x)\right]+\left[N_d(x)-N_a(x)\right]\right\} \tag{2.86}$$

Let us assume that the free carrier concentrations n and p are negligibly small compared to the fixed ionized impurities $N_a^- \cong N_a$ and $N_d^+ \cong N_d$ over the entire region defined by the depletion width bounded by $-x_p$ and x_n, that is, $N_d >> n_n$ or p_n and $N_a >> p_p$ or n_p as shown in Figure 2.17. This assumption is often referred to as the *depletion approximation*. It is often used during the development of analytical device models.

For the simplicity of modeling, we will assume that all the donors and acceptors within the depletion region are ionized, and that the junction is abrupt and not compensated; that is, there are no donor impurities on the *p*-side and no acceptor impurities on the *n*-side. With these assumptions, Equation 2.86 becomes

$$\frac{d^2\phi}{dx^2} = \frac{qN_a(x)}{K_{si}\varepsilon_0} \quad \text{for } -x_p < x < 0 \tag{2.87}$$

and,

$$\frac{d^2\phi}{dx^2} = -\frac{qN_d(x)}{K_{si}\varepsilon_0} \quad \text{for } 0 < x < x_n \tag{2.88}$$

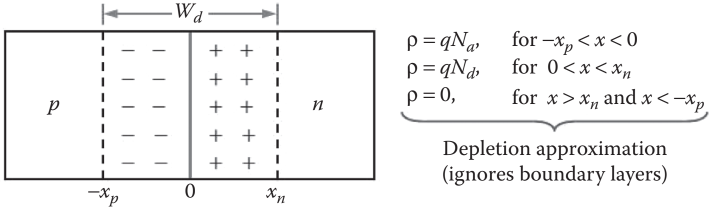

FIGURE 2.17
The *pn*-junction charge condition under *depletion approximation* in three different regions: the equilibrium depletion region is bounded by $-x_p$ and x_n on the *p*-region and *n*-regions, respectively; the depletion region is assumed to be free of mobile carriers with $\rho = 0$.

Integrating Equation 2.87 from $x = -x_p$ to at any point $x < 0$ and Equation 2.88 from $x > 0$ to $x = x_n$ using the boundary condition $d\phi/dx = 0$ at $x = -x_p$ and $x = x_n$, we get the electric field distribution in the depletion region. Thus, assuming a step *pn*-junction so that N_a and N_d are uniform in *p*- and *n*-regions, respectively, and depletion approximation the electric field, $E(x)$ distribution within the depletion region can be shown as

$$E(x) = -\frac{qN_a}{K_{si}\varepsilon_0}\left(x_p - x\right) \quad \text{for } -x_p < x < 0 \tag{2.89}$$

$$E(x) = -\frac{qN_d}{K_{si}\varepsilon_0}\left(x_n - x\right) \quad \text{for } 0 < x < x_n \tag{2.90}$$

Since the electric field must be continuous at $x = 0$, we get from Equations 2.89 and 2.90 the maximum electric field E_{max} as

$$E_{max} = -\frac{qN_a}{K_{si}\varepsilon_0}x_p = -\frac{qN_d}{K_{si}\varepsilon_0}x_n \tag{2.91}$$

or

$$qN_a x_p = qN_d x_n \tag{2.92}$$

which gives the distribution of charge on either side of the junction and shows that the negative charge on the *p*-side exactly equals the positive charge on the *n*-side. Equation 2.92 also shows that the width of the depletion region on each side of the junction varies inversely with the dopant concentration; the higher the doping concentration, the narrower the depletion region. Equations 2.89 and 2.90 also show that E varies linearly between 0 and E_{max} as shown in Figure 2.17.

Let ϕ_m is the total potential drop across the *pn*-junction; that is, $\phi_m = \left[\phi\left(x_n\right) - \phi\left(x_p\right)\right]$. Then the total potential drop can be obtained by integrating Equations 2.89 and 2.90 from $x = -x_p$ to $x = x_n$. Now, we can get:

$$\begin{aligned}\phi_m &= \int_{-x_p}^{x_n} d\phi(x) = -\int_{-x_p}^{x_n} E(x)dx \\ &= \frac{E_{max}\left(x_n + x_p\right)}{2} = \frac{E_{max}}{2}W_d\end{aligned} \tag{2.93}$$

where:

$W_d = (x_n + x_p)$ is the total width of the depletion layer

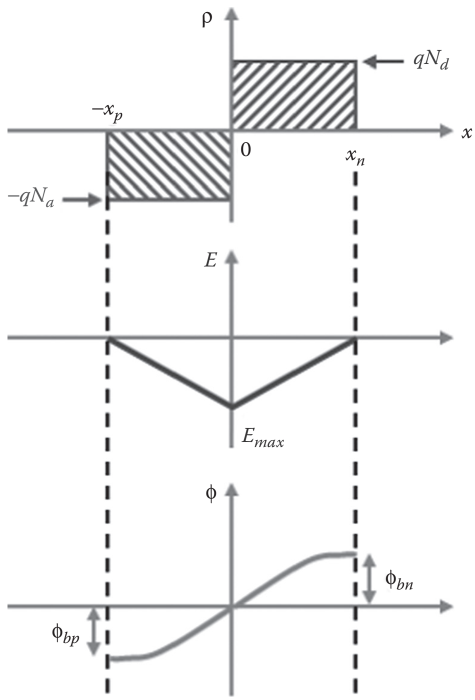

FIGURE 2.18
Depletion approximation of a *pn*-junction: the equilibrium distribution of charge, ρ; electric field, *E*; and electrostatic potential, ϕ within the depletion region.

It can be seen from Equation 2.93 that ϕ_m is equal to the area under the $E(x)$ versus x plot, that is, Figure 2.18. Eliminating E_{max} from Equations 2.91 and 2.93, we can show that

$$W_d = \sqrt{\frac{2\varepsilon_0 K_{si}(N_a + N_d)}{qN_aN_d}\phi_m} \tag{2.94}$$

In order to derive expressions for x_p and x_n, we integrate Equations 2.89 and 2.90 once again. Remembering that $E = -d\phi/dx$, and the potential difference between the p and n sides is ϕ_{bi}, it can be shown that

$$x_p = \sqrt{\frac{2K_{si}\varepsilon_0}{q}\frac{N_d}{N_a(N_a + N_d)}\phi_{bi}} \tag{2.95}$$

And,

$$x_n = \sqrt{\frac{2K_{si}\varepsilon_0}{q}\frac{N_a}{N_d(N_a + N_d)}\phi_{bi}} \tag{2.96}$$

So that the total depletion width W_d ($=x_p + x_n$) becomes

$$W_d = \sqrt{\frac{2K_{si}\varepsilon_0}{q}\left(\frac{1}{N_a}+\frac{1}{N_d}\right)\phi_{bi}} \tag{2.97}$$

Note that Equation 2.97 shows that W_d strongly depends on the doping on the lightly doped side and particularly W_d is inversely proportional to the square root of the doping concentration on the lightly doped side. The value of W_d given above is at thermal equilibrium without any external voltage applied to the *pn*-junction.

From Equations 2.91 and 2.92, the charge per unit area on either side of the depletion region is

$$Q_d = qN_d x_p = qN_d x_n = E_{max}K_{si}\varepsilon_0 \tag{2.98}$$

We can show that, the depletion layer capacitance per unit area is given by

$$C_d = \frac{d|Q_d|}{d\phi_m} = \frac{K_{si}\varepsilon_0}{W_d} \tag{2.99}$$

Equation 2.99 shows that the depletion capacitance of a *pn*-junction is equivalent to a parallel-plate capacitor of separation W_d and dielectric constant K_{si}. Physically, this is due to the fact that only the mobile charge at the edges of the depletion layer, but not the space charge within the depletion region, responds to changes of the applied voltage.

2.3.4 *pn*-Junctions under External Bias

An externally applied voltage, V_d across a *pn*-junction has the effect of shifting the Fermi level of the bulk neutral *n*-region relative to that of the bulk neutral *p*-region. That is, the total potential drop is the sum of the built-in potential and the externally applied potential:

$$\phi_m = \phi_{bi} \pm V_d \tag{2.100}$$

where:

"+" sign is for the case where the junction is reverse biased and $\phi_m > \phi_{bi}$
the "–" sign is for the case where the junction is forward biased and $\phi_m < \phi_{bi}$

Thus, when the *pn*-junction is in a nonequilibrium condition, with voltage V_d applied to it, then, as stated earlier, the potential barrier height becomes ($\phi_{bi} - V_d$), so that the depletion width as a function of voltage becomes

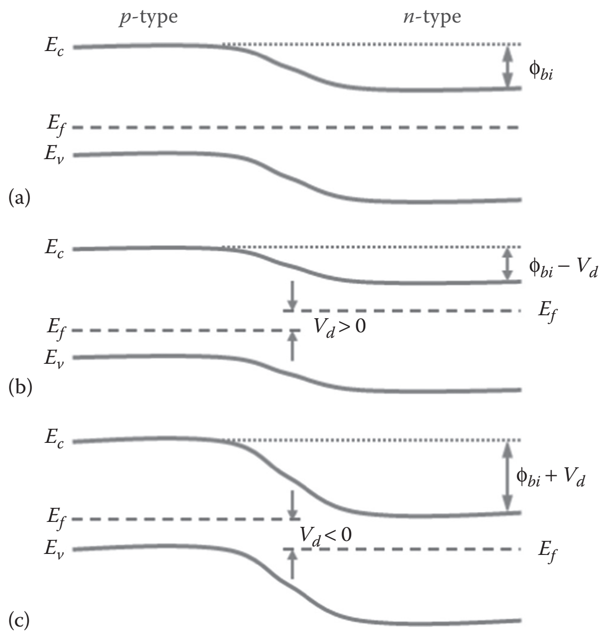

FIGURE 2.19
The *pn*-junction in equilibrium and under external bias: (a) equilibrium, (b) forward bias, and (c) reverse bias.

$$W_d = \sqrt{\frac{2K_{si}\varepsilon_0}{q}\left(\frac{1}{N_a}+\frac{1}{N_d}\right).\left(\phi_{bi}-V_d\right)} \tag{2.101}$$

This shows that a forward bias V_d ($\equiv V_f$) will result in a decrease in the depletion width due to the decrease in the barrier height, while a reverse bias $-V_d$ ($\equiv V_r$) will result in an increase in the depletion width due to a higher barrier height as shown in Figure 2.19.

Using Equation 2.95 for x_p or 2.96 for x_n in Equation 2.91, the maximum electric field E_{max} in the depletion region becomes

$$E_{max} = \sqrt{\frac{2q}{K_{si}\varepsilon_0}\frac{N_aN_d}{\left(N_a+N_d\right)}\left(\phi_{bi}-V_d\right)} \tag{2.102}$$

Equation 2.102 shows that the higher the reverse voltage (e.g., $-V_d$), the higher is the electric field across the *pn*-junction.

2.3.4.1 One-Sided Step Junctions

If the impurity concentration on one side of a *pn*-junction is much higher than the other side, the junction is called a *one-sided step junction*. In this case, the depletion region extends almost totally into the lighter doped side. For example, in the case of an $n+ p$ junction ($N_d >> N_a$ and $x_n << x_p$), the depletion width W_d is almost entirely in the p-side. Thus, from Equation 2.101, we can show that the general expression for W_d for a one-sided step junction is

$$W_d = \sqrt{\frac{2K_{si}\varepsilon_0}{qN_b}.(\phi_{bi} \pm V_d)} \tag{2.103}$$

where:
$N_b = N_a$ for $n+ p$ junction
$N_b = N_d$ for $p+ n$ junction

A more accurate result for the depletion width can be obtained by considering the majority carrier distribution tails or spillover (electrons in the n-side and holes in the p-side by Debye length, L_d) as shown by dashed lines in Figure 2.20. Each contributes a correction factor v_{kT} to ϕ_{bi}. Thus, the depletion width is still given by Equation 2.103 except that ϕ_{bi} is replaced by ($\phi_{bi} - 2v_{kT}$) so that, using this more accurate expression, W_d for a one-sided step junction becomes

$$W_d = \sqrt{\frac{2K_{si}\varepsilon_0}{qN_b}.(\phi_{bi} - 2v_{kT} \pm V_d)} \tag{2.104}$$

However, Equation 2.103 is accurate to within about 3% for the biases normally encountered in the VLSI circuits.

2.3.5 *pn*-Junction Equations

In considering I–V characteristics of a pn-junction, it is much more convenient to work with the quasi-Fermi potentials, instead of the intrinsic potential.

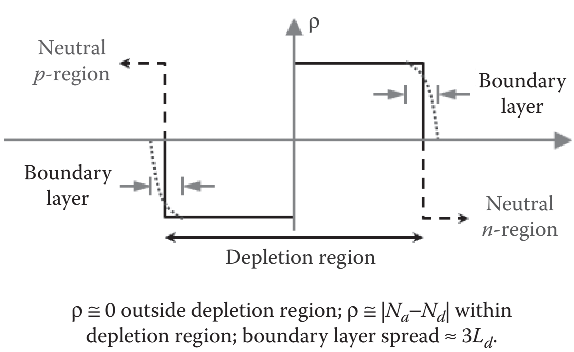

FIGURE 2.20
Majority carrier spillover (broken lines) outside the depletion region forming a boundary layer of about $3L_d$ at the boundary of the neutral bulk region; L_d is the Debye length defining the abruptness of the junction.

The quasi-Fermi potentials and the current densities for doped semiconductors given by Equations 2.73, 2.74, and 2.78 can be expressed as

$$\begin{aligned} J_n &= -qn\mu_n \frac{d\phi_n}{dx} \\ J_p &= -qp\mu_p \frac{d\phi_p}{dx} \end{aligned} \tag{2.105}$$

where:

$$\begin{aligned} \phi_n &= \phi_i - v_{kT} \ln\left(\frac{n}{n_i}\right) \\ \phi_p &= \phi_i + v_{kT} \ln\left(\frac{p}{n_i}\right) \end{aligned} \tag{2.106}$$

where:
ϕ_n and ϕ_p are the quasi-Fermi potentials for electrons and holes, respectively

2.3.5.1 Relationship between Minority Carrier Density and Junction Voltage

Under forward bias V_d, the barrier to majority carrier flow is reduced. And, electrons are injected from n-region to p-region and holes are injected from p-region to n-region. The electrons going from n-region to p-region become minority carriers in the p-region. Similarly, *holes* going from p-region to n-region become minority carriers in the n-region. Therefore, the minority carrier behavior is of fundamental importance to understand the behavior of a pn-junction. The minority carriers injected across the barrier will tend to recombine if given sufficient time. They will also tend to diffuse away from the region of the junction.

In order to calculate diode current in thermal equilibrium, let us consider n_{no} and p_{po} are the equilibrium majority carrier concentrations in the neutral n- and p-regions, respectively; and n_{po} and p_{no} are the equilibrium minority carrier electron and hole concentrations in the neutral p- and n-regions, respectively, as shown in Figure 2.21. Then from carrier statistics discussed in Section 2.2.7.2, we have in the *neutral* n-region

$$n_{no} \cong N_d; \;\; p_{no} \cong \frac{n_i^2}{N_d} \tag{2.107}$$

and, in the *neutral* p-region

$$p_{po} \cong N_a; \; n_{po} \cong \frac{n_i^2}{N_a} \tag{2.108}$$

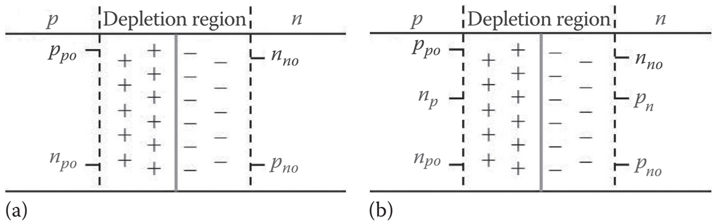

FIGURE 2.21
Carrier concentrations at the edge of depletion region: (a) *pn*-junction at equilibrium where p_{po} and n_{no} are the equilibrium majority carrier hole and electron concentrations in the *p*-type and *n*-type regions, respectively, whereas n_{po} and p_{no} are the equilibrium minority carrier electron and hole concentrations in the *p*-type and *n*-type regions, respectively and (b) *pn*-junction after minority carrier n_p and p_n injection in the bulk *p*-region and *n*-region, respectively.

From Equation 2.85, the equilibrium carrier concentrations in a *pn*-junction are given by the expressions

$$\phi_{bi} = \begin{cases} v_{kT} \ln\left(\dfrac{n_{no}}{n_{po}}\right) \\ v_{kT} \ln\left(\dfrac{p_{po}}{p_{no}}\right) \end{cases} \tag{2.109}$$

Therefore, from Equation 2.109, we can write for a *pn*-junction at equilibrium

$$\begin{aligned} n_{no} &= n_{po} \exp\left(\frac{\phi_{bi}}{v_{kT}}\right) \\ p_{po} &= p_{no} \exp\left(\frac{\phi_{bi}}{v_{kT}}\right) \end{aligned} \tag{2.110}$$

Now, under the applied bias V_d, we replace ϕ_{bi} by $(\phi_{bi} \pm V_d)$; therefore, from Equation of 2.110, the nonequilibrium carrier concentrations are given by

$$\begin{aligned} n_n &= n_p \exp\left(\frac{\phi_{bi} - V_d}{v_{kT}}\right) \\ p_p &= p_n \exp\left(\frac{\phi_{bi} - V_d}{v_{kT}}\right) \end{aligned} \tag{2.111}$$

where:
- n_p is the nonequilibrium minority electron concentration at the edge of the depletion region in the neutral *p*-region
- p_n is the nonequilibrium hole concentration at the edge of the depletion region in the neutral *n*-region as shown in Figure 2.21b

Let us further assume *low-level injection*, that is, the injected carrier densities are lower than the background concentrations, so that $n_n = n_{n0}$ and $p_p = p_{p0}$. Then from Equations 2.110 and 2.111, we get

$$\begin{aligned} n_p &= n_{po}\exp\left(\frac{V_d}{v_{kT}}\right) \\ p_n &= p_{no}\exp\left(\frac{V_d}{v_{kT}}\right) \end{aligned} \tag{2.112}$$

In Equation 2.112 n_p and p_n are the injected minority carrier concentrations at the edge of the depletion region in the *p*- and *n*-regions, respectively. The expressions in Equation 2.112 define the minority carrier densities at the *edge* of the space charge region under an applied bias and are the most important boundary conditions governing a *pn*-junction. They relate the minority carrier concentrations at the boundaries of the depletion layer to their thermal equilibrium values and to the applied voltage across the junction. They apply to both a forward-biased ($V_d > 0$) junction resulting in $n_p >> n_{po}$ at $x = -x_p$ and $p_n >> p_{no}$ at $x = x_n$, and to a reverse-biased ($V_d < 0$) junction resulting in $n_p << n_{po}$ at $x = -x_p$ and $p_n << p_{no}$ at $x = x_n$. Expressions in Equations 2.112 can be expressed as

$$\begin{aligned} n_p &= \frac{n_i^2}{p_{po}}\exp\left(\frac{V_d}{v_{kT}}\right) \\ p_n &= \frac{n_i^2}{n_{no}}\exp\left(\frac{V_d}{v_{kT}}\right) \end{aligned} \tag{2.113}$$

Again, for low-level injection in the *p*-region, $p_{po} = p$ and $n_p = n$; similarly, in the *n*-region, $n_{no} = n$ and $p_n = p$; therefore, we get from Equation 2.112 or Equation 2.113

$$pn = n_i^2\exp\left(\frac{V_d}{v_{kT}}\right) \tag{2.114}$$

Equation 2.114 defines the *pn*-product of carriers at the depletion edge under the applied voltage V_d as shown in Figure 2.21. Thus, the applied bias in a *pn*-junction sets up the following processes as shown in Figure 2.22:

- The injected carriers in the *n*- and *p*-regions momentarily set up an electric field (from *n* to *p*)
- This field draws in majority carriers in each region

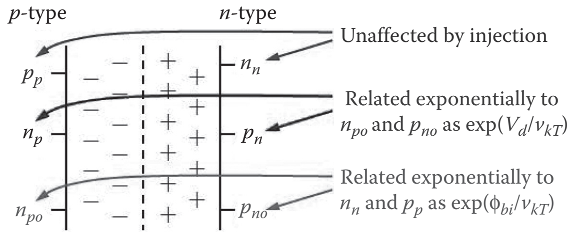

FIGURE 2.22
Carriers in a *pn*-junction under applied bias showing the corresponding dependence on built-in potential and applied bias.

- These majority carriers neutralize the injected carriers and reestablish the charge neutrality
- While this process is going on, the injected minority carriers diffuse into the *n*- and *p*-regions; that is, recombination process takes place over some distance

The distribution of carriers in the *n*-region of the *pn*-junction is shown in Figure 2.23. The majority carrier concentration shown by broken line remains unchanged whereas the minority carrier concentration decays exponentially and approaches to the equilibrium concentration in each side of the junction.

The injected excess carriers set up a momentary electric field, E, in the regions of excess carrier concentration. Then the current due to this drift electric field in the *n*-region is $I_{drift} = q\mu_n nE$ for majority carrier *electrons* and $I_{drift} = q\mu_p pE$ for minority carrier *holes.* Since $n >> p$, the hole drift current is negligible in the *n*-region. Similarly, electron drift current is negligible in the neutral *p*-region. The minority carriers move primarily by diffusion while the majority carriers are pulled to the junction by drift. Since the injected

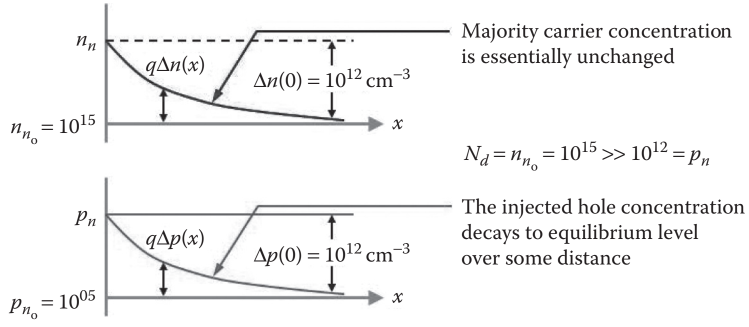

FIGURE 2.23
The carrier profile in the *n*-region of a *pn*-junction with applied bias; the majority carrier electron concentration, n_{no} is 1×10^{15} cm^{-3} and injected carrier concentration is 1×10^{12} cm^{-3} describing low-level injection.

minority carriers control the current flow in a *pn*-junction, the current flow in *pn*-junctions can be considered as the diffusion current only. *Thus, we see that the minority carriers really control the behavior of pn-junctions.*

2.3.6 *pn*-Junctions *I–V* Characteristics

We discussed in Section 2.3.2 that the drift component of the current caused by the electric field in the depletion region is exactly balanced out by the diffusion component of the current caused by the electron and hole concentration gradient across the junction, resulting in zero current flow in the *pn*-junction device. When an external voltage is applied, this current component balance is upset, and current will flow in the diode. If carriers are generated by light or some other external means, thermal equilibrium is disturbed, and current can also flow in a *pn*-junction. Here, the current flow in a *pn*-junction as a result of an external applied voltage is described.

Let us consider a forward-biased *pn*-junction. Electrons are injected from the *n*-side into the *p*-side, and holes are injected from the *p*-side to *n*-side. If the generation and recombination in the depletion region are negligible, then the hole current leaving *p*-side is the same as the hole current entering the *n*-side. Similarly, the electron current leaving the *n*-side is equal to the electron current entering the *p*-side. To determine the total current flowing in the *pn*-junction, we need to determine either hole current entering the *p*-side or electron current entering to *n*-side of the *pn*-junction.

The starting point for describing *I–V* characteristics of a *pn*-junction is the continuity equations. From Equation 2.81, the electron continuity equation is given by

$$-\frac{\partial n}{\partial t} = -\frac{1}{q}\frac{\partial J_n}{\partial x} + \left(G_n - R_n\right) \tag{2.115}$$

where:

R_n and G_n are the electron recombination and generation rates, respectively

Equation 2.115 can be rewritten as

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{n - n_o}{\tau_n} \tag{2.116}$$

where τ_n is the electron lifetime defined in terms of the excess electron concentration n over the thermal equilibrium value n_o in Equations 2.48 and 2.53 and is given by

$$\tau_n \equiv \frac{n - n_o}{R_n - G_n} \tag{2.117}$$

Now, substituting Equation 2.76 for J_n in Equation 2.116, we get

$$\frac{\partial n}{\partial t} = n\mu_n \frac{\partial E}{\partial x} + \mu_n E \frac{\partial n}{\partial x} + D_n \frac{\partial^2 n}{\partial x^2} - \frac{n - n_0}{\tau_n} \tag{2.118}$$

Equation 2.118 is the general equation that is solved under appropriate boundary conditions to derive an expression for electron current flow across a *pn*-junction under an applied bias.

In order to calculate the diode current, we assume that the injected minority carriers move away from the depletion region by diffusion only—*diffusion approximation*. We calculate the diode current under the following assumptions:

1. The step junction profile is applicable
2. The depletion approximation is valid
3. Low-level injection is maintained in the bulk
4. No generation–recombination takes place in the depletion region
5. There is no voltage drop in the bulk region so that V_d is sustained entirely across the depletion region
6. The width of the bulk *p*- and *n*-regions outside the depletion region is much longer than the minority carrier diffusion length for holes and electrons L_p and L_n, respectively (long-base diode)

With the above simplifying assumptions, the current through a *pn*-junction can be shown to be

$$I_d = I_s\left[\exp\left(\frac{V_d}{v_{kT}}\right) - 1\right] \tag{2.119}$$

where I_s is called the reverse saturation current and is given by

$$I_s = \begin{cases} qA_d n_i^2\left[\dfrac{D_p}{N_d L_p} + \dfrac{D_n}{N_a L_n}\right]; & W_n > L_p \quad \text{and} \quad W_p > L_n \\[2ex] qA_d n_i^2\left[\dfrac{D_p}{N_d W_n} + \dfrac{D_n}{N_a W_p}\right]; & W_n < L_p \quad \text{and} \quad W_p < L_n \end{cases} \tag{2.120}$$

where:

A_d is the active area of the *pn*-junction
W_n and W_p are the width of the neutral *n*- and *p*-regions, respectively
D_n and D_p are the minority carrier electron and hole diffusion constants, respectively
L_n and L_p are the minority carrier electron and hole diffusion lengths, respectively

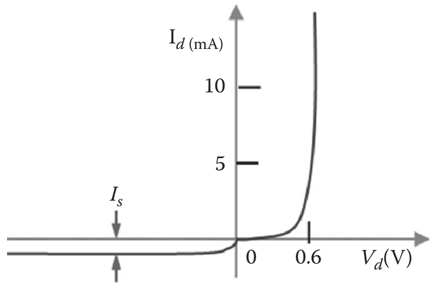

FIGURE 2.24
Current voltage characteristics of a typical *pn*-junction; I_s is the reverse saturation current; an applied voltage of about 0.6 V is required to overcome the built-in voltage and device conduction.

Actual diodes may represent intermediate cases, that is, $W_n > L_p$ and $W_p < L_n$ and vice versa. In either case, the lightly doped side of the junction largely determines the diode current I_d in Equation 2.119. Figure 2.24 shows a typical *I–V* characteristics of a *pn*-junction.

2.3.6.1 Temperature Dependence of pn-*Junction Leakage Current*

From Equation 2.120 we see that the temperature dependence of the electron and hole diffusion currents is dominated by the temperature dependence of the parameter n_i^2, which is proportional to exp($-E_g/kT$) as shown in Equation 2.14, where E_g is the bandgap energy. Then substituting for $n_i(T)$ from Equation 2.14 in Equation 2.120, we can show the temperature dependence of I_s with reference to T_{NOM} as

$$\begin{aligned} I_s(T) &= I_s(T_{NOM})\left(\frac{T}{T_{NOM}}\right)^3 \exp\left[\frac{E_g(T_{NOM})}{kT_{NOM}} - \frac{E_g(T)}{kT}\right] \\ &= I_s(T_{NOM})\left(\frac{T}{T_{NOM}}\right)^{XTI} \exp\left[\frac{E_g(T_{NOM})}{kT_{NOM}} - \frac{E_g(T)}{kT}\right] \end{aligned} \tag{2.121}$$

where exponent 3 is replaced by the parameter XTI. In advance *pn*-junction model for circuit CAD, two parameters XTI and NJ, called *temperature exponent coefficient fitting* parameters, are used to express Equation 2.121 as

$$I_s(T) = I_s(T_{NOM})\cdot \exp\left\{\frac{\left[E_g(T_{NOM})/kT_{NOM}\right] - \left[E_g(T)/kT\right] + XTI\ln\left(T/T_{NOM}\right)}{NJ}\right\} \tag{2.122}$$

2.3.6.2 Limitations of pn-*Junction Current Equation*

The ideal *pn*-junction current Equation 2.119 accurately describes the device characteristics of *pn*-junctions over a certain range of applied voltage. However, Equation 2.119 becomes inaccurate over a significant range of device operations both in the forward- and reverse-biased modes.

The current voltage characteristics of a *forward-biased* silicon *pn*-junction diode are shown in Figure 2.25 where the ideal diode current is shown by the broken line. Two different regions of nonideal behavior are shown in this plot. At a very low value of the forward bias ($V_d < 0.3$ V), the injected carrier densities are relatively small. When these carriers move through the depletion region, some of them may be lost by recombination in this region, thereby forming a recombination current I_{rec}, which is added to the ideal diode diffusion current. The result is a larger total current than that predicted by the ideal diode Equation 2.119, particularly in the low current level, and violates assumption 4. Thus, I_{rec} dominates in the silicon diode at very low current levels and negligibly small at higher current levels.

In deriving Equation 2.119, we have assumed that all the minority carriers cross the depletion region. In practice, some recombine through trapping centers. Then, using the SRH theory of generation and recombination, it can be shown that the space-charge recombination current I_{rec} is

$$I_{rec} = \frac{qA_d n_i W}{2\tau_{rec}} \exp\left(\frac{V_d}{2v_{kT}}\right) \tag{2.123}$$

In Equation 2.123, τ_{rec} is the lifetime associated with the recombination of excess carriers in the depletion region. τ_{rec} is analogous to, but usually greater than, τ_n and τ_p for the neutral regions and is generally approximately equal

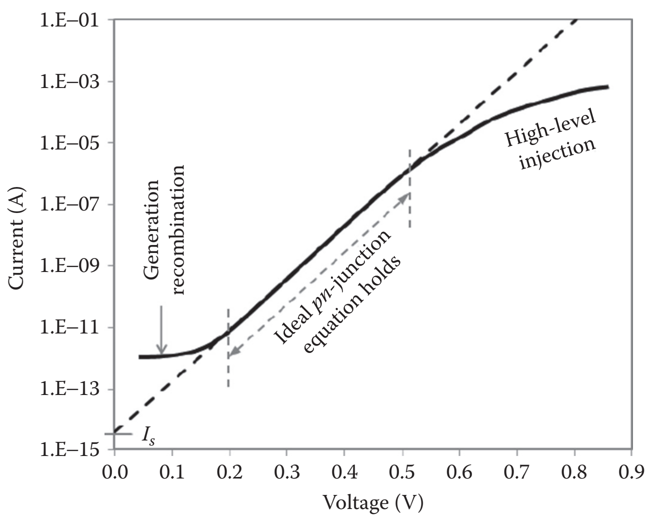

FIGURE 2.25
Forward characteristics of a real *pn*-junction: plot shows the deviation of ideal current equation at the low- and high-current levels due to generation–recombination and high-level injections, respectively.

to $2\sqrt{\tau_p \tau_n}$. Thus, the total diode saturation current, I_s, is the sum of Equations 2.120 and 2.123. In general, until V_d reaches a value of about 0.4 V, the neutral region diffusion current will be less than I_{rec}.

At high current levels, the injected minority carrier density is comparable to the majority carrier concentration (high-level injection), and therefore, assumption 3 is invalid. For high-level injection, majority carrier concentration increases significantly above its equilibrium value, giving rise to an electric field. Thus, in such cases both drift and diffusion components must be considered. The presence of the electric field results in a voltage drop across this region and thus reduces the applied voltage across the junction, resulting in a lower current than expected. It can be shown that under high-level injection the diode current I_d is

$$I_d = \frac{qA_d n_i D_p}{W} \exp\left(\frac{V_d}{2v_{kT}}\right) \quad \left(\text{high-level injection}\right) \tag{2.124}$$

which indicates that high-level current depends on $1/2v_{kT}$ rather than on $1/v_{kT}$ as shown in Figure 2.25. Thus, depending on the magnitude of the applied forward voltage, the current through a *pn*-junction can be represented by an empirical expression

$$I_d = I_s\left[\exp\left(\frac{V_d}{n_E v_{kT}}\right) - 1\right] \tag{2.125}$$

where n_E is called the *ideality factor* and is a measure of the deviation of the real and the ideal *I*–*V* plots. When recombination current dominates or when there is high-level injection $n_E = 2$ and when diffusion current dominates $n_E = 1$.

In the case of a *reverse-biased pn*-junction, Figure 2.26 shows the current through the *pn*-junction where I_s is the current due to an ideal *pn*-junction (Equation 2.119). Clearly, the current in a real *pn*-junction does not saturate at $-I_s$ as predicted by Equation 2.119. This is because when the *pn*-junction is reverse biased, generation of electron–hole pairs in the depletion region takes place, which was neglected in the ideal *pn*-junction equation. In fact, the generation current dominates because carrier concentrations are smaller than their thermal equilibrium values. Again, using SRH theory, it can be shown that the generation current I_{gen} is

$$I_{gen} = \frac{qA_d n_i W_d}{2\tau_{gen}} \tag{2.126}$$

where:

τ_{gen} is the generation lifetime of the carriers in the depletion region and is approximately equal to $2\tau_p$ if we assume $\tau_p = \tau_n$

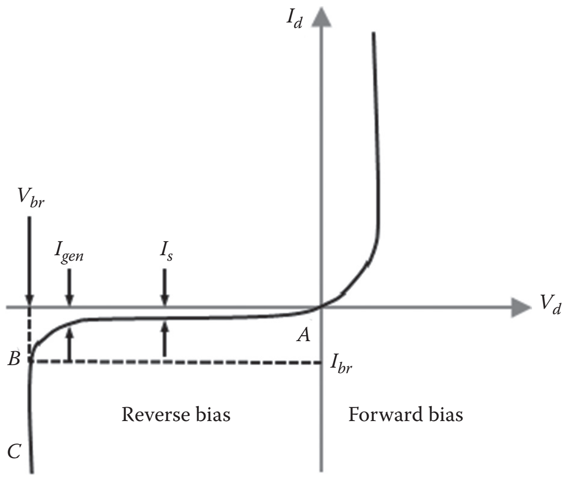

FIGURE 2.26
Reverse characteristics of a real *pn*-junction; V_{br} and I_{br} are the breakdown voltage and current, respectively; I_s is the ideal reverse saturation current; and I_{gen} is the generation current in the depletion region.

Note that while I_s is proportional to n_i^2, I_{gen} is proportional to n_i only. Thus, I_{gen} will dominate when n_i is small as is the case at room and low temperatures. Further, since the space charge width W_d increases as the square root of the reverse bias (Equation 2.103), the generation current increases with reverse bias voltage as shown in Figure 2.26. Thus, taking into account I_{gen}, the total reverse current I_r becomes $I_r \equiv -I_d = -(I_s + I_{gen})$. This value of I_r agrees well with the measured value of reverse current and also it provides proper voltage dependence of the reverse current in properly constructed silicon planar *pn*-junctions.

In real *pn*-junctions there is a third component of leakage current, called the surface leakage current I_{sl}. This current can be treated as a special case of I_{gen} modeled at the surface where a high concentration of dislocations at the oxide-silicon interface, often referred to as fast surface states, provides additional generation centers over those present in the bulk. It is very much process dependent and is responsible for large variation in the leakage current. Both process-induced and electrically induced defects at the surface generally increase the generation rate by an order of magnitude compared with the bulk recombination–generation rate. In that case I_{sl} dominates over the other components of I_r and is thus responsible for higher leakage current for a *pn*-junction compared to that predicted by the sum of I_{gen} and I_s. Leakage current is highly temperature dependent due to the presence of n_i term. Also, note that the generation limited leakage current is proportional to n_i while diffusion limited leakage current is proportional to n_i^2.

2.3.6.3 Bulk Resistance

At high current levels, bulk resistance and the metal–silicon contact resistance can produce a significant voltage drop (assumption 5), resulting in a smaller voltage across the junction and thus a lower current. Usually, the bulk resistance and contact resistance are combined into one resistor called series resistance r_s (Figure 2.27). Thus, if V_d is the applied voltage to the diode terminals and V_d' is the voltage across the diode junction, resulting in the current I_d as shown in Figure 2.24, we have

$$V_d = V_d' + r_s I_d \tag{2.127}$$

Under the ideal conditions when $r_s = 0$, $V_d = V_d'$, that is related to I_d by Equations 2.119 or 2.125. Thus, in the presence of the series resistance, *I–V* expression of a *pn*-junction becomes

$$I_d = I_s\left[\exp\left(\frac{V_d - I_d r_s}{n_E v_{kT}}\right) - 1\right] \tag{2.128}$$

Rearranging this equation yields

$$V_d = n_E v_{kT} \ln\left(1 + \frac{I_d}{I_s}\right) + I_d r_s \tag{2.129}$$

Clearly, when I_d is large, the terminal voltage V_d will increase linearly with I_d because $I_d r_s$ increases faster than the logarithmic term.

2.3.6.4 Junction Breakdown Voltage

From Equation 2.126, we observe that the *reverse* (or leakage) current of a *pn*-junction depends on W_d, and from Equation 2.101 we observe that W_d depends on the reverse bias $V_d = V_r$. Also, we notice from Equation 2.102 that the electric field in the depletion region increases with the increase of V_r. When the field reaches a certain critical field E_c corresponding to the reverse voltage $V_r = V_{br}$, called the *breakdown voltage*, a slight increase of reverse

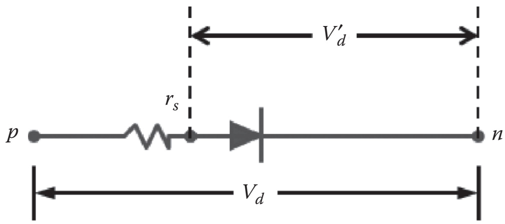

FIGURE 2.27
Diode model at high-level current: r_s is the diode resistance due to contact and the neutral bulk regions.

voltage causes a very large increase of current as shown in Figure 2.26 (region BC). This condition is often called the *breakdown condition* and is a most important consideration in device design. The breakdown occurs because carriers, while moving through the depletion region, acquire sufficient energy to create new electron–hole pairs through impact ionization [24,25]. The newly generated electron–hole pairs can also acquire sufficient energy from the field to create additional electron–hole pairs. Since the electrons and holes travel in opposite directions, the carriers can multiply a few times in the depletion region before they reach the electrodes. This multiplicative process results in an avalanche effect. The resulting breakdown voltage, V_{br}, is called the avalanche breakdown voltage and can be obtained using Equation 2.102.

$$E_{max} = \sqrt{\frac{2q}{K_{si}\varepsilon_0}\frac{N_a N_d}{(N_a + N_d)}(\phi_{bi} + V_r)} \tag{2.130}$$

At the breakdown condition, $E_{max} = E_c$ and $V_r = V_{br}$; since $V_{br} >> \phi_{bi}$, we can safely neglect ϕ_{bi} in Equation 2.130 to obtain the expression for breakdown voltage for a *pn*-junction

$$V_{br} = \frac{K_{si}\varepsilon_0 E_c^2}{2q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right) \tag{2.131}$$

Equation 2.131 shows that any increase in the doping, either of *n*- or *p*-region, results in a decrease in the breakdown voltage V_{br}. Further, it shows that V_{br} is controlled by the concentration N_b of the lightly doped region and is proportional to $1/N_b$. In a *pn*-junction, V_{br} generally varies as $N^{-2/3}$ [13]. For moderately doped silicon (1×10^{14} to 1×10^{16} cm^{-3}), the value of the critical field is $E_c \sim 4 \times 10^5$ V cm^{-1} and for a first approximation V_{br} is independent of doping [26].

If the *pn*-junction is heavily doped (concentration $>1 \times 10^{18}$ cm^{-3}) on both sides, the depletion layer is very narrow. Carriers cannot gain enough energy within the depletion region so that avalanche breakdown is not possible. However, in the depletion region, the electric field is high; E_{max} can be close to 1×10^6 V cm^{-1}. In such a heavily doped $p+$ $n+$ junction under reverse bias, electrons at the VB of the $p+$ side *tunnel* through the forbidden gap into the CB of the $n+$ side. This tunneling process can be approximated by a particle penetrating a triangular potential barrier, with a height higher than its energy by the semiconductor bandgap E_g. This *tunneling* process contributes to the current resulting in breakdown of the junction. This mechanism of breakdown is called the *Zener breakdown*. In the source-drain *pn*-junction of a MOSFET, the avalanche breakdown dominates [27,28].

2.3.7 *pn*-Junction Dynamic Behavior

Besides electrostatic behavior, *pn*-junctions are often subject to varying voltages. In such dynamic operations, charges in the *pn*-junction vary, resulting in an additional current not predicted by the DC current (Equation 2.119). There are two types of stored charge in a *pn*-junction: (1) the charge Q_{dep} due to the depletion or space-charge region on each side of the junction and (2) the charge Q_{dif} due to minority carrier injection. Remember that it is these injected (excess) mobile carriers that generate current I_d and also represent a stored charge Q_{dif} in a *pn*-junction. The latter is given by the area between the curve representing p_n (or n_p) and the steady state level p_{no} (or n_{po}) as shown in Figure 2.23. These two types of stored charges result in two types of capacitances: the junction capacitance C_j due to Q_{dep} and the diffusion capacitance due to Q_{dif}, as discussed in Sections 2.3.7.1 and 2.3.7.2, respectively.

2.3.7.1 *Junction Capacitance*

In a *pn*-junction, a small change in the applied voltage causes an incremental change in the depletion region charge Q_{dep} due to the corresponding change in the depletion width. If the applied voltage is returned to its original value, carriers flow in such a direction that the previous increment of charge is neutralized. The response of the *pn*-junction to the incremental voltage thus results in a generation of an effective capacitance C_j referred to as the *transition capacitance*, *junction capacitance*, or *depletion layer capacitance*. Recalling the definition of capacitance per unit area in terms of an incremental charge dQ_{dep} per unit area induced by an applied voltage dV_d, we have

$$C_j = \frac{dQ_{dep}}{dV_d} \tag{2.132}$$

Considering, $Q_{dep} = qN_a x_p = qN_d x_n$ from Equation 2.92, we can show

$$C_j = qN_a \frac{dx_p}{dV_d} = qN_d \frac{dx_n}{dV_d} \tag{2.133}$$

Then using Equation 2.95 or 2.96, the *pn*-junction capacitance per unit area can be shown as

$$C_j = \sqrt{\frac{qK_{si}\varepsilon_0}{2(\phi_{bi} - V_d)}\left(\frac{N_a N_d}{N_a + N_d}\right)} \tag{2.134}$$

Equation 2.134 is the expression for the diode capacitance for a step profile in terms of the physical parameters of the device. Remember that Equation 2.134 is valid for $V_d < \phi_{bi}$, that is, for reverse bias only. Comparing Equations 2.134 and 2.101, it is easy to see that

$$C_j = \frac{K_{si}\varepsilon_0}{W_d} \tag{2.135}$$

Equation 2.135 states that the junction capacitance is equivalent to that of a parallel plate capacitor with silicon as the dielectric and separated by a distance W_d, the depletion width. Though the derivation of Equation 2.134 is based on a step profile, it can be shown that *the relationship is valid for any arbitrary doping profile*.

It should be pointed out that although the *pn*-junction capacitance can be calculated using the parallel plate capacitor formula, there are differences between the two types of capacitors. While true parallel plate capacitance is independent of applied voltage, *pn*-junction capacitance given by Equation 2.134 becomes voltage dependent through W_d. Therefore, the total charge in a *pn*-junction cannot be obtained by simply multiplying the capacitance by the applied voltage, although a small variation in the charge can still be obtained by multiplying a small variation in the voltage by the instantaneous capacitance value. Another difference is that, in a *pn*-junction, the dipoles in the transition region have their positive charge in the *n*-side depletion region and negative charge in the *p*-side depletion region, while in a parallel plate capacitor the separation between the charges in the dipoles is much less and the dipoles are distributed homogenously throughout the dielectric.

For a one-sided step junction, for example, $n+p$ diode with $N_d >> N_a$, Equation 2.134 becomes

$$C_j = \sqrt{\frac{qK_{si}\varepsilon_0 N_a}{2\left(\phi_{bi} - V_d\right)}} \tag{2.136}$$

For the circuit CAD, it is more convenient to express capacitance in terms of model parameters. If C_{j0} is the junction capacitance at equilibrium, that is, at $V_d = 0$, then from Equation 2.134 we get

$$C_{j0} = \sqrt{\frac{qK_{si}\varepsilon_0}{2\phi_{bi}}\left(\frac{N_a N_d}{N_a + N_d}\right)} \tag{2.137}$$

Then using Equation 2.137 in Equation 2.134, the junction capacitance for a *pn*-junction is given by

$$C_j = \frac{C_{j0}}{\sqrt{1 - \left(V_d/\phi_{bi}\right)}} \tag{2.138}$$

In IC *pn*-junctions, the doping profile is neither abrupt nor linearly graded as assumed in the derivation for C_j, and therefore, to calculate the capacitance for real devices, we replace the one-half power in Equation 2.138 by m_j, called the junction *grading coefficient*, resulting in the following generalized equation for C_j as

$$C_j = \frac{C_{j0}}{\left[1-\left(V_d/\phi_{bi}\right)\right]^{m_j}} \tag{2.139}$$

For IC *pn*-junctions, m_j ranges between 0.2 and 0.6. Figure 2.28 shows a plot of the junction capacitance C_j as a function of junction voltage V_d. Note that the capacitance C_j decreases as the reverse-biased $|V_d|$ increases (V_d is negative). When the diode is forward biased (V_d is positive), the capacitance C_j increases and becomes infinite at $V_d = \phi_{bi}$ as shown in Figure 2.28 (Curve 1). This is because Equation 2.139 no longer applies due to the depletion approximation becoming invalid. A more exact analysis of the C_j as a function of the behavior of the forward bias V_d is shown by Curve 2. However, in SPICE a straight line is used instead of Curve 2 in Figure 2.28. In this case, we define a parameter F_c, $0 < F_c < 1$, such that when the *pn*-junction is forward biased and $V_d \geq F_c\phi_{bi}$, the following equation for C_j is used. By Taylor series expansion of $\left[1-\left(V_d/\phi_{bi}\right)\right]^{-m_j}$ at $V_d = F_c.\phi_{bi}$, we can show

$$\left(1-\frac{V_d}{\phi_{bi}}\right)^{-m_j} = \left(1-FC\right)^{-\left(m_j+1\right)}\left[m_j\left(\frac{V_d}{\phi_{bi}}\right)+1-FC\left(m_j+1\right)\right] \tag{2.140}$$

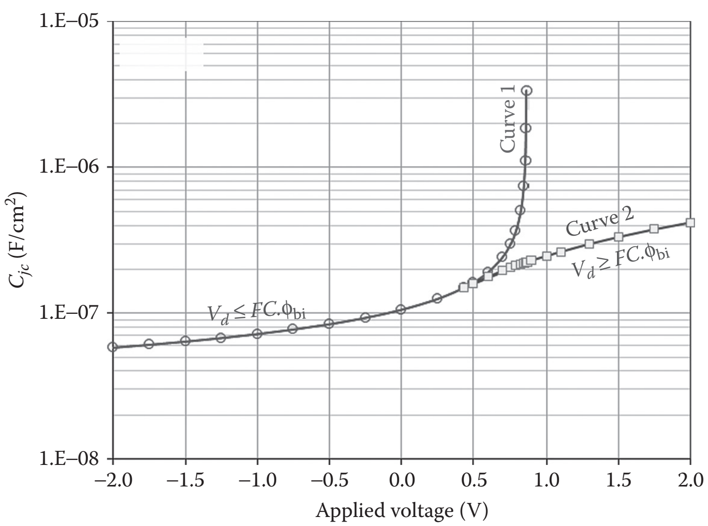

FIGURE 2.28
Junction capacitance of a typical *pn*-junction obtained by using the expressions in Equation 2.141; curve 1 represents Equation 2.138 for $V_d < \phi_{bi}$ and curve 2 is obtained by analytical expression to ensure convergence in circuit simulation during forward biasing a *pn*-junction.

Then we can show

$$C_j = \begin{cases} \dfrac{C_{j0}}{\left(1-\dfrac{V_d}{\phi_{bi}}\right)^{m_j}}; & V_d \le FC\phi_{bi} \\[2ex] \dfrac{C_{j0}}{(1-FC)^{m_j+1}}\left[\dfrac{m_j}{\phi_{bi}}V_d + 1 - FC\left(m_j+1\right)\right]; & V_d \ge FC\phi_{bi} \end{cases} \tag{2.141}$$

2.3.7.2 Diffusion Capacitance

The diffusion capacitance C_{dif} is associated with the rearrangement of the excess minority carriers in response to an incremental change in the applied forward voltage. The variation in the stored charge Q_{dif}, associated with the excess minority carrier injection in the bulk region under forward bias, is modeled by the capacitance C_{dif}. The capacitance C_{dif} is called the diffusion capacitance, because the minority carriers move across the bulk region by diffusion; since Q_{dif} is proportional to the current I_d, for an $n+\ p$ junction we can write

$$Q_{dif} = \frac{1}{A_d}\tau_p I_d \tag{2.142}$$

For a short base diode, τ_p is replaced by τ_t, the transit time of the pn-junction. For the case of a long base diode the transit time is the excess minority carrier lifetime. Differentiating Equation 2.142 gives

$$C_{dif} = \frac{dQ_{dif}}{dV_d} = \frac{\tau_{pI_s}}{A_d v_{kT}}\exp\left(\frac{V_d}{v_{kT}}\right) \tag{2.143}$$

where we have used Equation 2.119 for I_d. A more accurate derivation shows that the value of C_{dif} is half of the value in Equation 2.143.

EXAMPLE:
Let us compare the magnitude of the two capacitances for a forward bias of 0.3 V; assume we have an $n+\ p$ diode with $N_a = 1 \times 10^{15}$ cm^{-3} and $N_d = 1 \times 10^{19}$ cm^{-3}; then Equation 2.84 gives $\phi_{bi} = 0.814$ V. For a forward bias of 0.3 V, Equation 2.101 gives $W_d = 8.15 \times 10^{-5}$ cm and Equation 2.134 gives $C_j = 1.27 \times 10^{-8}$ F cm^{-2}.

Again, assuming $\tau_t = 1 \times 10^{-7}$ sec, and $I_s = 4 \times 10^{-12}$ A for a junction area of 20×20 μm^2 gives $C_{dif} = 4 \times 10^{-7}$ F cm^{-2}, which is much larger than C_j.

It should be noted that under forward bias, C_{dif} increases much faster with increasing V_d (=V_f), due to the exponential dependence on V_d, as compared to C_j. However, under reverse bias, C_j decreases much more slowly with increasing $V_d\,(=-V_r)$, as compared to C_{dif}. Therefore, C_j is the dominant capacitance for reverse bias and small for forward bias ($V_d < \phi_{bi}/2$), while diffusion capacitance C_{dif} is dominant for forward bias ($V_d > \phi_{bi}/2$).

2.3.7.3 Small Signal Conductance

In the model discussed in Section 2.3.7.2, referred to as the large-signal model, we did not place any restriction on the allowed voltage variation. However, in some circuit situations, voltage variations are sufficiently small so that the resulting small current variations can be expressed using linear relationships. This is the so called small signal behavior of a *pn*-junction. An example of linear relations are the capacitances C_j and C_{dif} in Equations 2.141 and 2.143, respectively, as they represent an overall nonlinear charge storage effect in terms of linear circuit elements (capacitors), although we did not label them as such.

For small variations about the operating point, which is set by the DC condition, the nonlinear junction current can be linearized so that the incremental diode current is proportional to the incremental applied bias. This linear relationship is used to calculate the small signal conductance g_d

$$g_d = \frac{dI_d}{dV_d} \tag{2.144}$$

Using (2.119) for I_d, we have

$$g_d = \frac{I_s}{v_{kT}} \exp\left(\frac{V_d}{v_{kT}}\right) = \frac{1}{v_{kT}}\left(I_d + I_s\right) \tag{2.145}$$

Thus, Equation 2.145 clearly shows that g_d is proportional to the slope of the DC characteristics at the operating point. When the diode is forward biased, I_d is much larger than I_s and therefore, g_d is proportional to I_d. However, when the diode is reverse biased, $I_d = -I_s$ and therefore, from Equation 2.145, g_d becomes zero. But in real diodes, $g_d \neq 0$ in the reverse bias condition due to the fact that the generation current I_{gen} (Equation 2.126) is dominant conduction mechanism.

2.3.8 Diode Equivalent Circuit for Circuit CAD

The small signal equivalent circuit of a *pn*-junction is shown in Figure 2.29. In Figure 2.29, r_s represents the series resistance due to ohmic drop across the neutral *n*- and *p*-regions; C_j is junction capacitance; C_d is the diffusion capacitance due to the minority carrier diffusion through the neutral regions; and g_d is the small signal conductance of the *pn*-junctions.

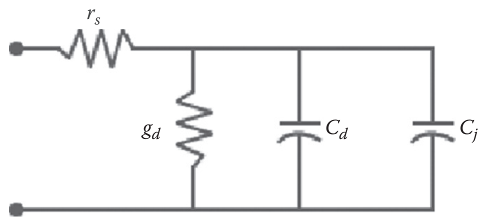

FIGURE 2.29
An equivalent circuit for a *pn*-junction showing the relevant circuit elements: r_s is the series resistance of the neutral *n*- and *p*-regions; C_j is junction capacitance; C_d is the diffusion capacitance; and g_d is the small-signal conductance.

2.4 Summary

This chapter presented a brief overview of the basic semiconductor physics and basic theory of extrinsic semiconductors forming *pn*-junctions. First of all, the basic properties of intrinsic semiconductor materials including bond and band structures, intrinsic carrier concentration, and energy levels are discussed. Then the behavior of extrinsic semiconductors, carrier statistics of electrons and holes, carrier transport, and transport equations are discussed. After the discussion of *p*-type and *n*-type semiconductors, the basic properties of *n*- and *p*-type semiconductors forming *pn*-junctions are described. Then the basic theory of *pn*-junctions, current transport, and dynamic characteristics are discussed. Finally, a basic equivalent circuit model of *pn*-junction for circuit CAD is presented.

Physical Constants

Constants	Symbol	Magnitude	Units
Electronic charge	q	1.602×10^{-19}	C
Free-electron mass	m	9.11×10^{-28}	g
Boltzmann's constant	k	1.38×10^{-23}	J K^{-1}
		8.62×10^{-5}	eV K^{-1}
Planck's constant	h	6.25×10^{-34}	J s
Permittivity of free space	ε_0	8.854×10^{-14}	F cm^{-1}
Thermal voltage at 300 K	$kT/q = v_{kT}$	0.02586	V
Thermal energy at 300 K	kT	0.02586	eV

Exercises

2.1 Experimental results show that the bandgap energy (E_g) in silicon decreases with temperature (T). The E_g versus T behavior is modeled by an empirical relation in circuit CAD given by

$$E_g(T) = 1.160 - \frac{7.02\text{x}10^{-4}T^2}{1108 + \text{T}}(\text{eV}) \tag{E2.1}$$

Here, T is in Kelvin.

a. Compute and plot E_g for $0 \le T \le 600$ K.

b. From the plot extract $E_g(T = 300\text{ K})$.

c. E_g versus T is also modeled by polynomial equations given below:

$$E_g(T) = 1.206 - 2.73 \times 10^{-4}T \quad (\text{eV}) \tag{E2.2}$$

and

$$E_g(T) = 1.16 - 3 \times 10^{-4}T \quad (\text{eV}) \tag{E2.3}$$

Calculate $E_g(T)$ using the polynomial equations and plot E_g versus T characteristics on the same graph in part (a) (superimpose). From the plots, show the range of temperature at which the polynomial equations are valid. Extract the values of $E_g(T = 300°\text{ K})$ from the polynomial equations and compare with that in part (a).

2.2 Use Equation E2.1 to compute and plot n_i versus T for $0 \le T \le 600$ K from the following equation:

$$n_i(T) = 3.9 \times 10^{16} T^{3/2} \exp\left[-\frac{E_g(T)}{2kT}\right] \tag{E2.4}$$

From the plot, extract n_i at $T = 300°$ K and compare your results with that obtained for silicon.

2.3 A p-type semiconductor is doped with $N_a = 1 \times 10^{16}$ cm^{-3} and has the minority carrier lifetime = 10 μsec.

a. Calculate the steady state electron and hole concentrations under light that creates 10^{18} cm^{-3} sec^{-1} electron–hole pairs.

b. Calculate and sketch the position of equilibrium Fermi level E_f relative to E_i.

c. Calculate and sketch the position of quasi-Fermi levels E_{fn} and E_{fp} relative to E_i.

d. Compare the position of equilibrium Fermi level in part (b) with that of the steady state quasi-Fermi levels under the light in part (c). What are the similarities and differences? Explain.

e. Calculate and compare the *pn*-products under the equilibrium and nonequilibrium conditions at room temperature.

2.4 Consider an abrupt $n+p$-junction with $N_d = 10^{20}$ cm^{-3}, $N_a = 1 \times 10^{16}$ cm^{-3}, and area = 20×20 μm^2:

a. Calculate the built-in potential (ϕ_{bi}) and zero-bias capacitance (C_{j0}).

b. Calculate the junction capacitance for an applied bias $V = -5$ V.

2.5 An IC resistor is shown in Figure E2.1. The doping concentrations for the *n*- and *p*-type regions are $N_d = 2.5 \times 10^{16}$ cm^{-3} and $N_a = 2.5 \times 10^{15}$ cm^{-3}, respectively. The junction depth $X_j = 0.4$ μm, the width of the *n*-type region $W = 2.5$ μm, and its length is $L = 20$ μm. The contact regions are each $3W \times 3W$ in area as shown in Figure E2.1.

a. Calculate the depletion width into the *n*- and *p*-sides of the *pn*-junction at $V_d = 0$.

b. Calculate the sheet resistance of the *n*-type region. Assume that the depletion region does not contribute to resistivity.

c. Calculate and sketch the position of quasi-Fermi levels E_{fn} and E_{fp} relative to E_i.

d. Calculate the maximum electric field at the *pn*-junction.

e. Assuming the DC voltage $V_d = 0$, calculate the depletion capacitance C_d in fF between the *n*-region and the *p*-type substrate.

f. Compute and plot C_j–V characteristics for applied bias range –2.0 V to ϕ_{bi} of the *pn*-junction for the doping gradient factor $m = 0.3$, 0.4, and 0.5. Explain your results.

g. Use series expansion to show that the expression in Equation 2.141 is valid for $V_d \geq FC.\phi_{bi}$.

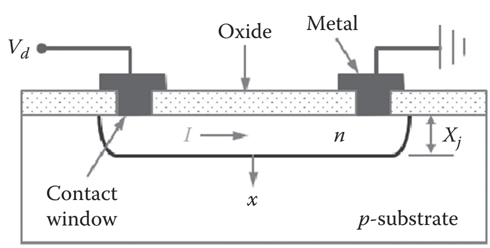

FIGURE E2.1
pn-junction capacitance modeling.

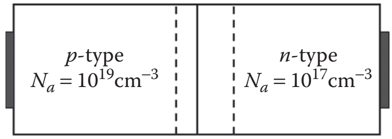

FIGURE E2.2
pn-junction *I–V* characteristics.

h. Compute and plot C_j–V characteristics for $-1.2\ V \leq V_d \leq 1.2\ V$ using Equation 2.141 for $V_d \leq FC.\phi_{bi}$ and $V_d \geq FC.\phi_{bi}$. Consider $m = 0.36$. State any assumptions you make including the fitting parameter FC. Explain your results.

2.6 In the derivation of the forward *I–V* characteristic of a *pn*-junction, we assumed *quasi-equilibrium*; that is, we assumed that we could simply subtract V_d as a small perturbation on the equilibrium situation. We will examine the validity of this assumption in this problem. Consider the diode shown in Figure E2.2 (the contacts are remote).

a. Assuming $D_n = 25\ \text{cm}^2\ \text{sec}^{-1}$, $D_p = 10\ \text{cm}^2\ \text{sec}^{-1}$, and $L_n = L_p = 10\ \mu\text{m}$, calculate the current that flows across the junction at an applied forward bias of 0.4 V.

b. With $V_d = 0$, electrons and holes will flow across the junction due to drift and diffusion, such that the currents due to drift and diffusion exactly cancel each other ($I = 0$). Estimate the hole diffusion current that would flow if there were no electric field to stop it.

c. What do your answer in part (a) and (b) tell you about the validity of our quasi-equilibrium assumption.