Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng

PART I

SEMICONDUCTOR PHYSICS

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng

1

Physics and Properties of Semiconductors—A Review

1.1 INTRODUCTION

The physics of semiconductor devices is naturally dependent on the physics of semiconductor materials themselves. This chapter presents a summary and review of the basic physics and properties of semiconductors. It represents only a small cross section of the vast literature on semiconductors; only those subjects pertinent to device operations are included here. For detailed consideration of semiconductor physics, the reader should consult the standard textbooks or reference works by Dunlap,[1] Madelung,[2] Moll,[3] Moss,[4] Smith.[5] Böer,[6] Seeger,[7] and Wang,[8] to name a few.

To condense a large amount of information into a single chapter, four tables (some in appendixes) and over 30 illustrations drawn from experimental data are compiled and presented here. This chapter emphasizes the two most-important semiconductors: silicon (Si) and gallium arsenide (GaAs). Silicon has been studied extensively and widely used in commercial electronics products. Gallium arsenide has been intensively investigated in recent years. Particular properties studied are its direct bandgap

for photonic applications and its intervalley-carrier transport and higher mobility for generating microwaves.

1.2 CRYSTAL STRUCTURE

1.2.1 Primitive Cell and Crystal Plane

A crystal is characterized by having a well-structured periodic placement of atoms. The smallest assembly of atoms that can be repeated to form the entire crystal is called a primitive cell, with a dimension of lattice constant a. Figure 1 shows some important primitive cells.

Many important semiconductors have diamond or zincblende lattice structures which belong to the tetrahedral phases; that is, each atom is surrounded by four equidistant nearest neighbors which lie at the corners of a tetrahedron. The bond between two nearest neighbors is formed by two electrons with opposite spins. The diamond and the zincblende lattices can be considered as two interpenetrating face-centered cubic (fcc) lattices. For the diamond lattice, such as silicon (Fig. 1d), all the atoms are the same; whereas in a zincblende lattice, such as gallium arsenide (Fig. 1e), one sublattice is gallium and the other is arsenic. Gallium arsenide is a III-V compound, since it is formed from elements of groups III and V of the periodic table.

Most III-V compounds crystallize in the zincblende structure;[2,9] however, many semiconductors (including some III-V compounds) crystallize in the rock-salt or wurtzite structures. Figure 1f shows the rock-salt lattice, which again can be considered as two interpenetrating face-centered cubic lattices. In this rock-salt structure, each atom has six nearest neighbors. Figure 1g shows the wurtzite lattice, which can be considered as two interpenetrating hexagonal close-packed lattices (e.g., the sublattices of cadmium and sulfur). In this picture, for each sublattice (Cd or S), the two planes of adjacent layers are displaced horizontally such that the distance between these two planes are at a minimum (for a fixed distance between centers of two atoms), hence the name *close-packed.* The wurtzite structure has a tetrahedral arrangement of four equidistant nearest neighbors, similar to a zincblende structure.

Appendix F gives a summary of the lattice constants of important semiconductors, together with their crystal structures.[10,11] Note that some compounds, such as zinc sulfide and cadmium sulfide, can crystallize in either zincblende or wurtzite structures.

Since semiconductor devices are built on or near the semiconductor surface, the orientations and properties of the surface crystal planes are important. A convenient method of defining the various planes in a crystal is to use Miller indices. These indices are determined by first finding the intercepts of the plane with the three basis axes in terms of the lattice constants (or primitive cells), and then taking the reciprocals of these numbers and reducing them to the smallest three integers having the same ratio. The result is enclosed in parentheses (hkl) called the Miller indices for a single plane or a set of parallel planes $\{hkl\}$. Figure 2 shows the Miller indices of important planes in a cubic crystal. Some other conventions are given in Table 1. For

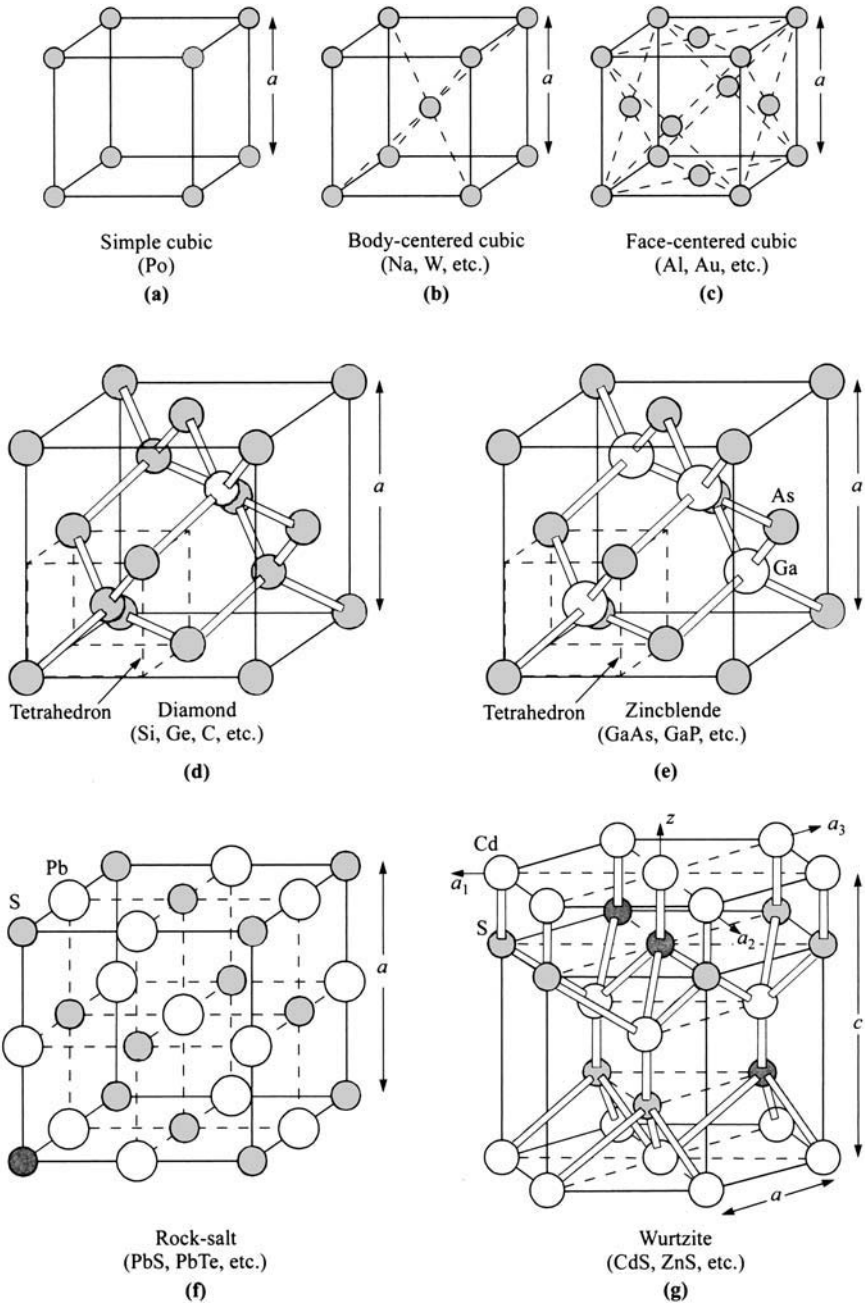


Fig. 1 Some important primitive cells (direct lattices) and their representative elements; a is the lattice constant.

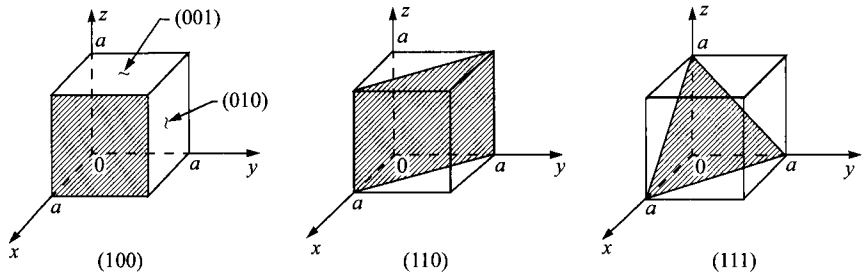


Fig. 2 Miller indices of some important planes in a cubic crystal.

silicon, a single-element semiconductor, the easiest breakage or cleavage planes are the {111} planes. In contrast, gallium arsenide, which has a similar lattice structure but also has a slight ionic component in the bonds, cleaves on {110} planes.

Three primitive basis vectors, $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ of a primitive cell, describe a crystalline solid such that the crystal structure remains invariant under translation through any vector that is the sum of integral multiples of these basis vectors. In other words, the direct lattice sites can be defined by the set[12]

$$\boldsymbol{R} = m\boldsymbol{a} + n\boldsymbol{b} + p\boldsymbol{c} \tag{1}$$

where m, n, and p are integers.

1.2.2 Reciprocal Lattice

For a given set of the direct basis vectors, a set of reciprocal lattice basis vectors $\boldsymbol{a}^*$, $\boldsymbol{b}^*$, $\boldsymbol{c}^*$ can be defined as

$$\boldsymbol{a}^* \equiv 2\pi \frac{\boldsymbol{b} \times \boldsymbol{c}}{\boldsymbol{a} \cdot \boldsymbol{b} \times \boldsymbol{c}}, \tag{2a}$$

$$\boldsymbol{b}^* \equiv 2\pi \frac{\boldsymbol{c} \times \boldsymbol{a}}{\boldsymbol{a} \cdot \boldsymbol{b} \times \boldsymbol{c}}, \tag{2b}$$

Table 1 Miller Indices and Their Represented Plane or Direction of a Crystal Surface

Miller Indices	Description of plane or direction
(hkl)	For a plane that intercepts $1/h$, $1/k$, $1/l$ on the x-, y-, and z-axis, respectively.
$(\bar{h}kl)$	For a plane that intercepts the negative x-axis.
$\{hkl\}$	For a full set of planes of equivalent symmetry, such as {100} for (100), (010), (001), $(\bar{1}00)$, $(0\bar{1}0)$, and $(00\bar{1})$ in cubic symmetry.
$[hkl]$	For a direction of a crystal such as [100] for the x-axis.
$\langle hkl \rangle$	For a full set of equivalent directions.
$[hklm]$	For a plane in a hexagonal lattice (such as wurtzite) that intercepts $1/h$, $1/k$, $1/l$, $1/m$ on the a_1-, a_2-, a_3-, and z-axis, respectively (Fig. 1g).

$$c^* \equiv 2\pi \frac{a \times b}{a \cdot b \times c} \tag{2c}$$

such that $a \cdot a^* = 2\pi$, $a \cdot b^* = 0$, and so on. The denominators are identical due to the equality that $a \cdot b \times c = b \cdot c \times a = c \cdot a \times b$ which is the volume enclosed by these vectors. The general reciprocal lattice vector is given by

$$G = ha^* + kb^* + lc^* \tag{3}$$

where h, k, and l are integers. It follows that one important relationship between the direct lattice and the reciprocal lattice is

$$G \cdot R = 2\pi \times \text{Integer}, \tag{4}$$

and therefore each vector of the reciprocal lattice is normal to a set of planes in the direct lattice. The volume V_c^* of a primitive cell of the reciprocal lattice is inversely proportional to that (V_c) of the direct lattice; that is, $V_c^* = (2\pi)^3/V_c$, where $V_c \equiv a \cdot b \times c$.

The primitive cell of a reciprocal lattice can be represented by a Wigner-Seitz cell. The Wigner-Seitz cell is constructed by drawing perpendicular bisector planes in the reciprocal lattice from the chosen center to the nearest equivalent reciprocal lattice sites. This technique can also be applied to a direct lattice. The Wigner-Seitz cell in the reciprocal lattice is called the first Brillouin zone. Figure 3a shows a typical example for a body-centered cubic (bcc) reciprocal lattice.[13] If one first draws lines from the center point (Γ) to the eight corners of the cube, then forms the bisector

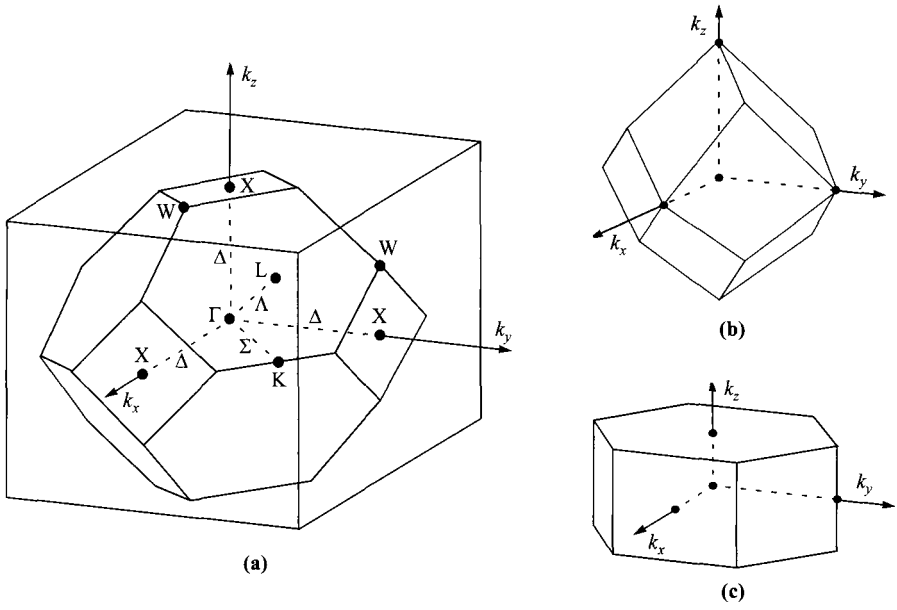


Fig. 3 Brillouin zones for (a) fcc, diamond, and zincblende lattices, (b) bcc lattice, and (c) wurtzite lattice.

planes, the result is the truncated octahedron within the cube—a Wigner-Seitz cell. It can be shown that[14] a face-centered cubic (fcc) direct lattice with lattice constant a has a bcc reciprocal lattice with spacing $4\pi/a$. Thus the Wigner-Seitz cell shown in Fig. 3a is the primitive cell of the reciprocal (bcc) lattice for an fcc direct lattice. The Wigner-Seitz cells for bcc and hexagonal direct lattices can be similarly constructed and shown in Figs. 3b and 3c.[15] It will be shown that the reciprocal lattice is useful to visualize the E-k relationship when the coordinates of the wave vectors $\boldsymbol{k}$ ($|\boldsymbol{k}| = k = 2\pi/\lambda$) are mapped into the coordinates of the reciprocal lattice. In particular, the Brillouin zone for the fcc lattice is important because it is relevant to most semiconductor materials of interest here. The symbols used in Fig. 3a will be discussed in more details.

1.3 ENERGY BANDS AND ENERGY GAP

The energy-momentum (E-k) relationship for carriers in a lattice is important, for example, in the interactions with photons and phonons where energy and momentum have to be conserved, and with each other (electrons and holes) which leads to the concept of energy gap. This relationship also characterizes the effective mass and the group velocity, as will be discussed later.

The band structure of a crystalline solid, that is, the energy-momentum (E-k) relationship, is usually obtained by solving the Schrödinger equation of an approximate one-electron problem. The Bloch theorem, one of the most-important theorems basic to band structure, states that if a potential energy $V(\boldsymbol{r})$ is periodic in the direct lattice space, then the solutions for the wavefunction $\psi(\boldsymbol{r},\boldsymbol{k})$ of the Schrödinger equation[14,16]

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2 + V(\boldsymbol{r})\right]\psi(\boldsymbol{r},\boldsymbol{k}) = E(\boldsymbol{k})\psi(\boldsymbol{r},\boldsymbol{k}) \tag{5}$$

are of the form of a Bloch function

$$\psi(\boldsymbol{r},\boldsymbol{k}) = \exp(j\boldsymbol{k}\cdot\boldsymbol{r})U_b(\boldsymbol{r},\boldsymbol{k}). \tag{6}$$

Here b is the band index, $\psi(\boldsymbol{r},\boldsymbol{k})$ and $U_b(\boldsymbol{r},\boldsymbol{k})$ are periodic in $\boldsymbol{R}$ of the direct lattice. Since

$$\begin{aligned}\psi(\boldsymbol{r}+\boldsymbol{R},\boldsymbol{k}) &= \exp[j\boldsymbol{k}\cdot(\boldsymbol{r}+\boldsymbol{R})]U_b(\boldsymbol{r}+\boldsymbol{R},\boldsymbol{k}) \\ &= \exp(j\boldsymbol{k}\cdot\boldsymbol{r})\exp(j\boldsymbol{k}\cdot\boldsymbol{R})U_b(\boldsymbol{r},\boldsymbol{k}),\end{aligned} \tag{7}$$

and is equal to $\psi(\boldsymbol{r},\boldsymbol{k})$, it is necessary that $\boldsymbol{k}\cdot\boldsymbol{R}$ is a multiple of 2π. It is the property of Eq. 4 that the reciprocal lattice can be used when $\boldsymbol{G}$ is replaced with $\boldsymbol{k}$ for visualizing the E-k relationship.

From the Bloch theorem one can also show that the energy $E(\boldsymbol{k})$ is periodic in the reciprocal lattice, that is, $E(\boldsymbol{k}) = E(\boldsymbol{k}+\boldsymbol{G})$, where $\boldsymbol{G}$ is given by Eq. 3. For a given band index, to label the energy uniquely, it is sufficient to use only $\boldsymbol{k}$'s in a primitive cell of the reciprocal lattice. The standard convention is to use the Wigner-Seitz cell in the reciprocal lattice (Fig. 3). This cell is the Brillouin zone or the first Brillouin zone.[13] It is thus evident that we can reduce any momentum $\boldsymbol{k}$ in the reciprocal space to a

point inside the Brillouin zone, where any energy state can be given a label in the reduced zone schemes.

The Brillouin zone for the diamond and the zincblende lattices is the same as that of the fcc and is shown in Fig. 3a. Table 2 summarizes its most-important symmetry points and symmetry lines, such as the center of the zone, the zone edges and their corresponding k axes.

The energy bands of solids have been studied theoretically using a variety of numerical methods. For semiconductors the three methods most frequently used are the orthogonalized plane-wave method,[17,18] the pseudopotential method,[19] and the $\boldsymbol{k} \cdot \boldsymbol{p}$ method.[5] Figure 4 shows results of studies of the energy-band structures of Si and GaAs. Notice that for any semiconductor there is a forbidden energy range in which allowed states cannot exist. Energy regions or energy bands are permitted above and below this energy gap. The upper bands are called the conduction bands; the lower bands, the valence bands. The separation between the energy of the lowest conduction band and that of the highest valence band is called the bandgap or energy gap E_g, which is one of the most-important parameters in semiconductor physics. In this figure the bottom of the conduction band is designated E_C, and the top of the valence band E_V. Within the bands, the electron energy is conventionally defined to be positive when measured upward from E_C, and the hole energy is positive when measured downward from E_V. The bandgaps of some important semiconductors are listed in Appendix F.

The valence band in the zincblende structure, such as that for GaAs in Fig. 4b, consists of four subbands when spin is neglected in the Schrödinger equation, and each band is doubled when spin is taken into account. Three of the four bands are degenerate at $k = 0$ (Γ point) and form the upper edge of the band, and the fourth band forms the bottom (not shown). Furthermore, the spin-orbit interaction causes a splitting of the band at $k = 0$.

Near the band edges, i.e., bottom of E_C and top of E_V, the E-k relationship can be approximated by a quadratic equation

$$E(k) = \frac{\hbar^2 k^2}{2m^*}, \tag{8}$$

where m^* is the associated effective mass. But as shown in Fig. 4, along a given direction, the two top valence bands can be approximated by two parabolic bands with different curvatures: the heavy-hole band (the wider band in k-axis with smaller $\partial^2 E/\partial k^2$)

Table 2 Brillouin Zone of fcc, Diamond, and Zincblende Lattices: Zone Edges and Their Corresponding Axes (Γ is the Center)

Point	Degeneracy	Axis
Γ, (0,0,0)	1	
X, $2\pi/a(\pm1,0,0)$, $2\pi/a(0,\pm1,0)$, $2\pi/a(0,0,\pm1)$	6	Δ, $\langle 1,0,0\rangle$
L, $2\pi/a(\pm1/2,\pm1/2,\pm1/2)$	8	Λ, $\langle 1,1,1\rangle$
K, $2\pi/a(\pm3/4,\pm3/4,0)$, $2\pi/a(0,\pm3/4,\pm3/4)$, $2\pi/a(\pm3/4,0,\pm3/4)$	12	Σ, $\langle 1,1,0\rangle$

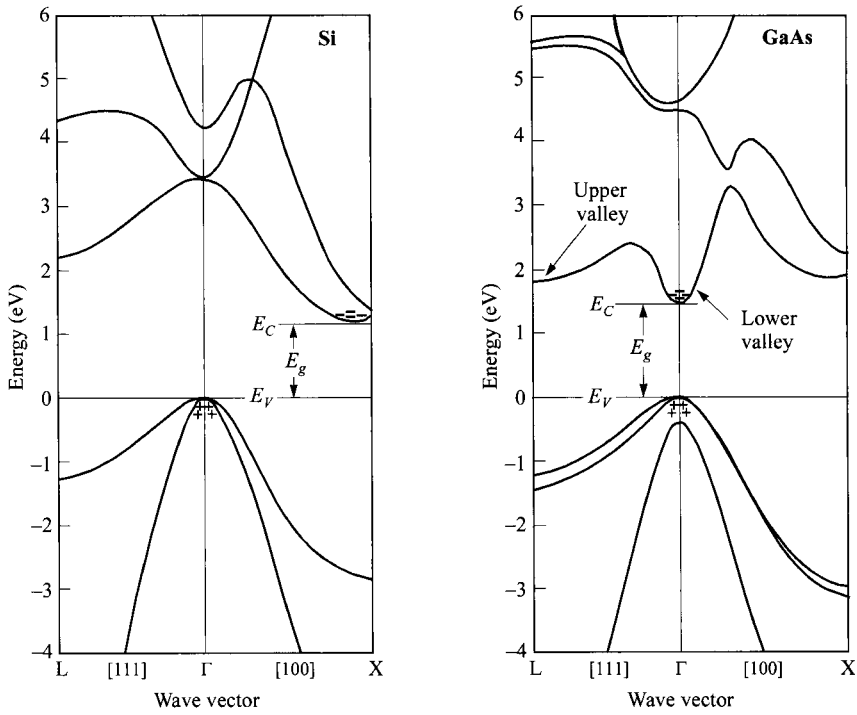


Fig. 4 Energy-band structures of (a) Si and (b) GaAs, where E_g is the energy bandgap. Plus signs (+) indicate holes in the valence bands and minus signs (–) indicate electrons in the conduction bands. (After Ref. 20.)

and the light-hole band (the narrower band with larger $\partial^2 E/\partial k^2$). The effective mass in general is tensorial with components m_{ij}^* defined as

$$\frac{1}{m_{ij}^*} \equiv \frac{1}{\hbar^2}\frac{\partial^2 E(k)}{\partial k_i \partial k_j}. \tag{9}$$

The effective masses are listed in Appendix F for important semiconductors.

Carriers in motion are also characterized by a group velocity

$$v_g = \frac{1}{\hbar}\frac{dE}{dk} \tag{10}$$

and with momentum

$$p = \hbar k. \tag{11}$$

The conduction band consists of a number of subbands (Fig. 4). The bottom of the conduction band can appear at the center $k = 0$ (Γ) or off center along different k axes. Symmetry considerations alone do not determine the location of the bottom of the conduction band. Experimental results show, however, that in Si it is off center and

along the [100] axis (Δ), and in GaAs the bottom is at $k = 0$ (Γ). Considering that the valence-band maximum (E_V) occurs at Γ, the conduction-band minimum can be aligned or misaligned in k-space in determining the bandgap. This results in direct bandgap for GaAs and indirect bandgap for Si. This bears significant consequences when carriers transfer between this minimum gap in that momentum (or k) is conserved for direct bandgap but changed for indirect bandgap.

Figure 5 shows the shapes of the constant-energy surfaces. For Si there are six ellipsoids along the ⟨100⟩-axes, with the centers of the ellipsoids located at about three-fourths of the distance from the Brillouin zone center. For GaAs the constant energy surface is a sphere at the zone center. By fitting experimental results to parabolic bands, we obtain the electron effective masses; one for GaAs and two for Si, m_l^* along the symmetry axes and m_t^* transverse to the symmetry axes. Appendix G also includes these values.

At room temperature and under normal atmospheric pressure, the values of the bandgap are 1.12 eV for Si and 1.42 eV for GaAs. These values are for high-purity materials. For highly doped materials the bandgaps become smaller. Experimental results show that the bandgaps of most semiconductors decrease with increasing temperature. Figure 6 shows variations of bandgaps as a function of temperature for Si and GaAs. The bandgap approaches 1.17 and 1.52 eV respectively for these two semiconductors at 0 K. The variation of bandgaps with temperature can be expressed approximately by a universal function

$$E_g(T) \approx E_g(0) - \frac{\alpha T^2}{T + \beta} \tag{12}$$

where $E_g(0)$, α, and β are given in the inset of Fig. 6. The temperature coefficient dE_g/dT is negative for both semiconductors. Some semiconductors have positive dE_g/dT; for example, the bandgap of PbS (Appendix F) increases from 0.286 eV at 0 K to 0.41 eV at 300 K. Near room temperature, the bandgap of GaAs increases with pressure P,[24] and dE_g/dP is about 12.6×10^{-6} eV-cm^2/N, while the Si bandgap decreases with pressure, with $dE_g/dP = -2.4\times10^{-6}$ eV-cm^2/N.

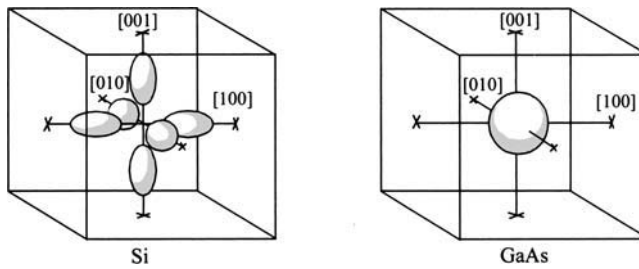
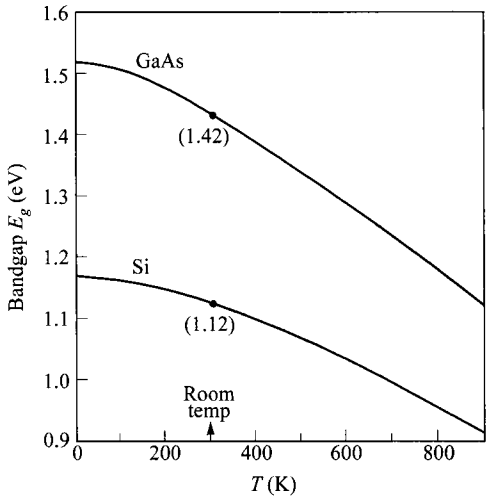


Fig. 5 Shapes of constant-energy surfaces for electrons in Si and GaAs. For Si there are six ellipsoids along the ⟨100⟩-axes with the centers of the ellipsoids located at about three-fourths of the distance from the Brillouin zone center. For GaAs the constant-energy surface is a sphere at zone center. (After Ref. 21.)



	$E_g(0)$ (eV)	α (eV/K)	β (K)
GaAs	1.519	5.4×10^{-4}	204
Si	1.169	4.9×10^{-4}	655

$$E_g(T) = E_g(0) - \frac{\alpha T^2}{T+\beta}$$

Fig. 6 Energy bandgaps of Si and GaAs as a function of temperature. (After Refs. 22–23.)

1.4 CARRIER CONCENTRATION AT THERMAL EQUILIBRIUM

One of the most-important properties of a semiconductor is that it can be doped with different types and concentrations of impurities to vary its resistivity. Also, when these impurities are ionized and the carriers are depleted, they leave behind a charge density that results in an electric field and sometimes a potential barrier inside the semiconductor. Such properties are absent in a metal or an insulator.

Figure 7 shows three basic bond representations of a semiconductor. Figure 7a shows intrinsic silicon, which is very pure and contains a negligibly small amount of impurities. Each silicon atom shares its four valence electrons with the four neigh-

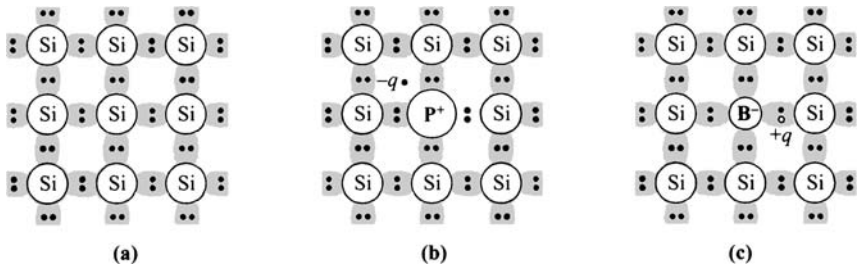


Fig. 7 Three basic bond pictures of a semiconductor. (a) Intrinsic Si with no impurity. (b) n-type Si with donor (phosphorus). (c) p-type Si with acceptor (boron).

boring atoms, forming four covalent bonds (also see Fig. 1). Figure 7b shows an *n*-type silicon, where a substitutional phosphorous atom with five valence electrons has replaced a silicon atom, and a *negative*-charged electron is *donated* to the lattice in the conduction band. The phosphorous atom is called a *donor*. Figure 7c similarly shows that when a boron atom with three valence electrons substitutes for a silicon atom, a *positive*-charged *hole* is created in the valence band, and an additional electron will be *accepted* to form four covalent bonds around the boron. This is *p*-type, and the boron is an *acceptor*.

These names of *n*- and *p*-type had been coined when it was observed that if a metal whisker was pressed against a *p*-type material, forming a Schottky barrier diode (see Chapter 3), a *positive* bias was required on the semiconductor to produce a noticeable current.[25,26] Also, when exposed to light, a *positive* potential was generated with respect to the metal whisker. Conversely, a *negative* bias was required on an *n*-type material to produce a large current.

1.4.1 Carrier Concentration and Fermi Level

We first consider the intrinsic case without impurities added to the semiconductor. The number of electrons (occupied conduction-band levels) is given by the total number of states $N(E)$ multiplied by the occupancy $F(E)$, integrated over the conduction band,

$$n = \int_{E_C}^{\infty} N(E)F(E)dE. \tag{13}$$

The density of states $N(E)$ can be approximated by the density near the bottom of the conduction band for low-enough carrier densities and temperatures:[5]

$$N(E) = M_C \frac{\sqrt{2}m_{de}^{3/2}(E-E_C)^{1/2}}{\pi^2 \hbar^3}. \tag{14}$$

M_C is the number of equivalent minima in the conduction band and m_{de} is the density-of-state effective mass for electrons:[5]

$$m_{de} = (m_1^* m_2^* m_3^*)^{1/3} \tag{15}$$

where m_1^*, m_2^*, m_3^* are the effective masses along the principal axes of the ellipsoidal energy surface. For example, in silicon $m_{de} = (m_l^* m_t^{*2})^{1/3}$. The occupancy is a strong function of temperature and energy, and is represented by the Fermi-Dirac distribution function

$$F(E) = \frac{1}{1 + \exp[(E-E_F)/kT]} \tag{16}$$

where E_F is the Fermi energy level which can be determined from the charge neutrality condition (see Section 1.4.3).

The integral of Eq. 13 can be evaluated to be

$$n = N_C \frac{2}{\sqrt{\pi}} F_{1/2}\left(\frac{E_F - E_C}{kT}\right) \tag{17}$$

where N_C is the effective density of states in the conduction band and is given by

$$N_C \equiv 2\left(\frac{2\pi m_{de} kT}{h^2}\right)^{3/2} M_C. \tag{18}$$

The Fermi-Dirac integral, changing variables with $\eta \equiv (E - E_C)/kT$ and $\eta_F \equiv (E_F - E_C)/kT$, is given by

$$F_{1/2}\left(\frac{E_F - E_C}{kT}\right) \equiv F_{1/2}(\eta_F) = \int_{E_C}^{\infty} \frac{[(E - E_C)/kT]^{1/2}}{1 + \exp[(E - E_F)/kT]} \frac{dE}{kT}$$

$$= \int_0^{\infty} \frac{\eta^{1/2}}{1 + \exp(\eta - \eta_F)} d\eta \tag{19}$$

whose values are plotted in Fig. 8. Note that for $\eta_F < -1$, the integral can be approximated by an exponential function. At $\eta_F = 0$ when the Fermi level coincides with the band edge, the integral has a value of ≈ 0.6 such that $n \approx 0.7 N_C$.

Nondegenerate Semiconductors. By definition, in nondegenerate semiconductors, the doping concentrations are smaller than N_C and the Fermi levels are more than several kT below E_C (negative η_F), the Fermi-Dirac integral approaches

$$F_{1/2}\left(\frac{E_F - E_C}{kT}\right) = \frac{\sqrt{\pi}}{2} \exp\left(-\frac{E_C - E_F}{kT}\right) \tag{20}$$

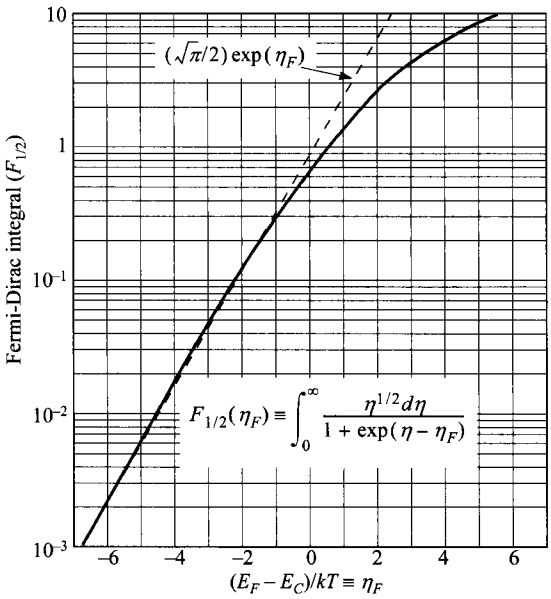


Fig. 8 Fermi-Dirac integral $F_{1/2}$ as a function of Fermi energy. (After Ref. 27.) Dashed line is approximation of Boltzmann statistics.

and Boltzmann statistics apply. Equation 17 becomes

$$n = N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \quad \text{or} \quad E_C - E_F = kT \ln\left(\frac{N_C}{n}\right). \tag{21}$$

Similarly, for p-type semiconductors we can obtain the hole density and its Fermi level near the top of the valence band;

$$p = N_V \frac{2}{\sqrt{\pi}} F_{1/2}\left(\frac{E_V - E_F}{kT}\right) \tag{22}$$

which can be simplified to

$$p = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) \quad \text{or} \quad E_F - E_V = kT \ln\left(\frac{N_V}{p}\right), \tag{23}$$

where N_V is the effective density of states in the valence band and is given by

$$N_V \equiv 2\left(\frac{2\pi m_{dh} kT}{h^2}\right)^{3/2}. \tag{24}$$

Here m_{dh} is the density-of-state effective mass of the valence band:[5]

$$m_{dh} = (m_{lh}^{*\,3/2} + m_{hh}^{*\,3/2})^{2/3} \tag{25}$$

where the subscripts refer to *light* and *heavy* hole masses previously referenced in Eq. 9.

Degenerate Semiconductors. As shown in Fig. 8, for degenerate levels where n- or p-concentrations are near or beyond the effective density of states (N_C or N_V), the value of Fermi-Dirac integral has to be used instead of the simplified Boltzmann statistics. For $\eta_F > -1$, the integral has weaker dependence on the carrier concentration. Note that also the Fermi levels are outside the energy gap. A useful estimate of the Fermi level as a function of carrier concentration is given by, for n-type semiconductor[28]

$$E_F - E_C \approx kT\left[\ln\left(\frac{n}{N_C}\right) + 2^{-3/2}\left(\frac{n}{N_C}\right)\right], \tag{26a}$$

and for p-type

$$E_V - E_F \approx kT\left[\ln\left(\frac{p}{N_V}\right) + 2^{-3/2}\left(\frac{p}{N_V}\right)\right]. \tag{26b}$$

Intrinsic Concentration. For intrinsic semiconductors at finite temperatures, thermal agitation occurs which results in continuous excitation of electrons from the valence band to the conduction band, and leaving an equal number of holes in the valence band. This process is balanced by recombination of the electrons in the conduction band with holes in the valence band. At steady state, the net result is $n = p = n_i$, where n_i is the intrinsic carrier density.

The Fermi level for an intrinsic semiconductor (which by definition is nondegenerate) is obtained by equating Eqs. 21 and 23:

$$E_F = E_i = \frac{E_C + E_V}{2} + \frac{kT}{2}\ln\left(\frac{N_V}{N_C}\right)$$
$$= \frac{E_C + E_V}{2} + \frac{3kT}{4}\ln\left(\frac{m_{dh}}{m_{de}M_C^{2/3}}\right). \tag{27}$$

Hence the Fermi level E_i of an intrinsic semiconductor generally lies very close to, but not exactly at, the middle of the bandgap. The intrinsic carrier density n_i can be obtained from Eq. 21 or 23:

$$n_i = N_C\exp\left(-\frac{E_C - E_i}{kT}\right) = N_V\exp\left(-\frac{E_i - E_V}{kT}\right) = \sqrt{N_C N_V}\exp\left(-\frac{E_g}{2kT}\right)$$
$$= 4.9\times10^{15}\left(\frac{m_{de}m_{dh}}{m_0^2}\right)^{3/4} M_C^{1/2}T^{3/2}\exp\left(-\frac{E_g}{2kT}\right). \tag{28}$$

Figure 9 shows the temperature dependence of n_i for Si and GaAs. As expected, the larger the bandgap is, the smaller the intrinsic carrier density will be.[30]

It also follows that for nondegenerate semiconductors, the product of the majority and minority carrier concentrations is fixed to be

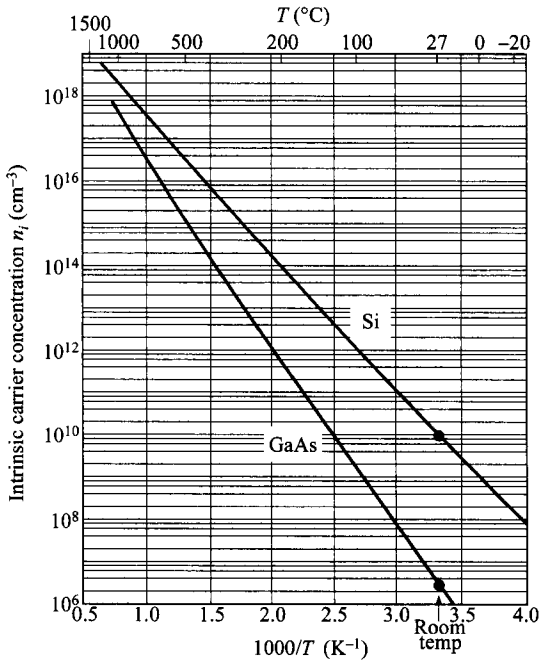


Fig. 9 Intrinsic carrier concentrations of Si and GaAs as a function of reciprocal temperature. (After Refs. 22 and 29.)

$$pn = N_C N_V \exp\left(-\frac{E_g}{kT}\right) = n_i^2 \quad , \tag{29}$$

which is known as the mass-action law. But for degenerate semiconductors, $pn < n_i^2$. Also using Eq. 28 and E_i as the reference energy, we have the alternate equations for n-type materials;

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right) \quad \text{or} \quad E_F - E_i = kT \ln\left(\frac{n}{n_i}\right) , \tag{30a}$$

and for p-type materials;

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right) \quad \text{or} \quad E_i - E_F = kT \ln\left(\frac{p}{n_i}\right) . \tag{30b}$$

1.4.2 Donors and Acceptors

When a semiconductor is doped with donor or acceptor impurities, impurity energy levels are introduced that usually lie within the energy gap. A donor impurity has a donor level which is defined as being neutral if filled by an electron, and positive if empty. Conversely, an acceptor level is neutral if empty and negative if filled by an electron. These energy levels are important in calculating the fraction of dopants being ionized, or electrically active, as discussed in Section 1.4.3.

To get a feeling of the magnitude of the impurity ionization energy, we use the simplest calculation based on the hydrogen-atom model. The ionization energy for the hydrogen atom in vacuum is

$$E_H = \frac{m_0 q^4}{32\pi^2 \varepsilon_0^2 \hbar^2} = 13.6 \text{ eV} . \tag{31}$$

The ionization energy for a donor $(E_C - E_D)$ in a lattice can be obtained by replacing m_0 by the conductivity effective mass of electrons[5]

$$m_{ce} = 3\left(\frac{1}{m_1^*} + \frac{1}{m_2^*} + \frac{1}{m_3^*}\right)^{-1} \tag{32}$$

and by replacing ε_0 by the permittivity of the semiconductor ε_s in Eq. 31:

$$E_C - E_D = \left(\frac{\varepsilon_0}{\varepsilon_s}\right)^2 \left(\frac{m_{ce}}{m_0}\right) E_H . \tag{33}$$

The ionization energy for donors as calculated from Eq. 33 is 0.025 eV for Si and 0.007 eV for GaAs. The hydrogen-atom calculation for the ionization level for the acceptors is similar to that for the donors. The calculated acceptor ionization energy (measured from the valence-band edge, $E_a \equiv (E_A - E_V)$ is 0.05 eV for Si and GaAs.

Although this simple hydrogen-atom model given above certainly cannot account for the details of ionization energy, particularly the deep levels in semiconductors,[31–33] the calculated values do predict the correct order of magnitude of the true ionization energies for shallow impurities. These calculated values are shown to be

much smaller than the energy gap, and often are referred to as shallow impurities if they are close to the band edges. Also, since these small ionization energies are comparable to the thermal energy kT, ionization is usually complete at room temperature. Figure 10 shows the measured ionization energies for various impurities in Si and GaAs. Note that it is possible for a single atom to have many levels; for example, gold in silicon has both an acceptor level and a donor level in the forbidden energy gap.

1.4.3 Calculation of Fermi Level

The Fermi level for the intrinsic semiconductor (Eq. 27) lies very close to the middle of the bandgap. Figure 11a depicts this situation, showing schematically from left to right the simplified band diagram, the density of states $N(E)$, the Fermi-Dirac distribution function $F(E)$, and the carrier concentrations. The shaded areas in the conduction band and the valence band represent electrons and holes, and their numbers are the same; i.e., $n = p = n_i$ for the intrinsic case.

When impurities are introduced to the semiconductor crystals, depending on the impurity energy level and the lattice temperature, not all dopants are necessarily ionized. The ionized concentration for donors is given by[36]

$$N_D^+ = \frac{N_D}{1 + g_D \exp[(E_F - E_D)/kT]} \tag{34}$$

where g_D is the ground-state degeneracy of the donor impurity level and equal to 2 because a donor level can accept one electron with either spin (or can have no electron). When acceptor impurities of concentration N_A are added to a semiconductor crystal, a similar expression can be written for the ionized acceptors

$$N_A^- = \frac{N_A}{1 + g_A \exp[(E_A - E_F)/kT]} \tag{35}$$

where the ground-state degeneracy factor g_A is 4 for acceptor levels. The value is 4 because in most semiconductors each acceptor impurity level can accept one hole of either spin and the impurity level is doubly degenerate as a result of the two degenerate valence bands at $\boldsymbol{k} = 0$.

When impurity atoms are introduced, the total negative charges (electrons and ionized acceptors) must equal the total positive charges (holes and ionized donors), represented by the charge neutrality

$$n + N_A^- = p + N_D^+ . \tag{36}$$

With impurities added, the mass-action law ($pn = n_i^2$) in Eq. 29 still applies (until degeneracy), and the pn product is always independent of the added impurities.

Consider the case shown in Fig. 11b, where donor impurities with the concentration N_D (cm^{-3}) are added to the crystal. The charge neutrality condition becomes

$$\begin{aligned} n &= N_D^+ + p \\ &\approx N_D^+ \end{aligned} \quad . \tag{37}$$

With substitution, we obtain

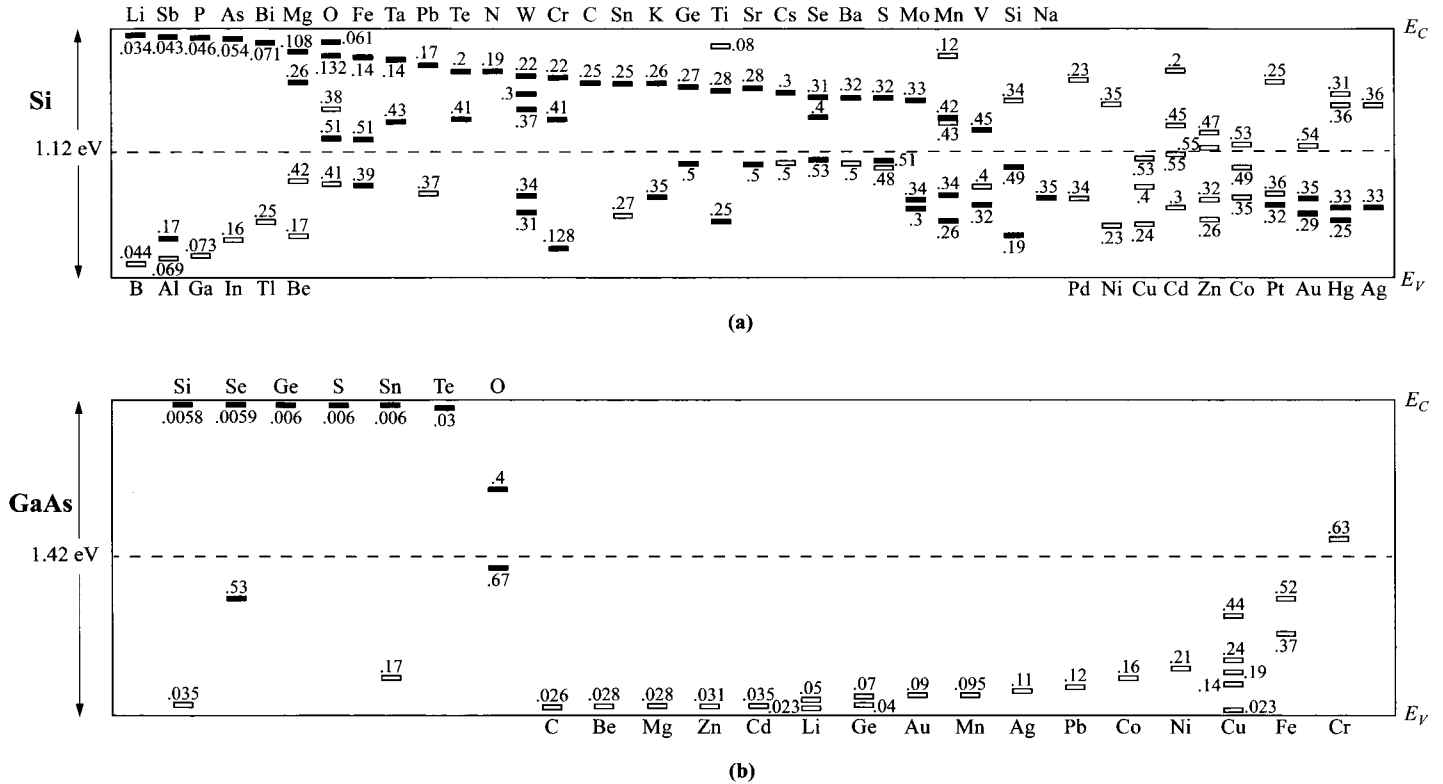


Fig. 10 Measured ionization energies for varies impurities in (a) Si and (b) GaAs. Levels below the gap center are measured from E_V. Levels above the gap center are measured from E_C. Solid bars represent donor levels and hollow boxes represent acceptor levels. (After Refs. 29, 31, 34, and 35.)

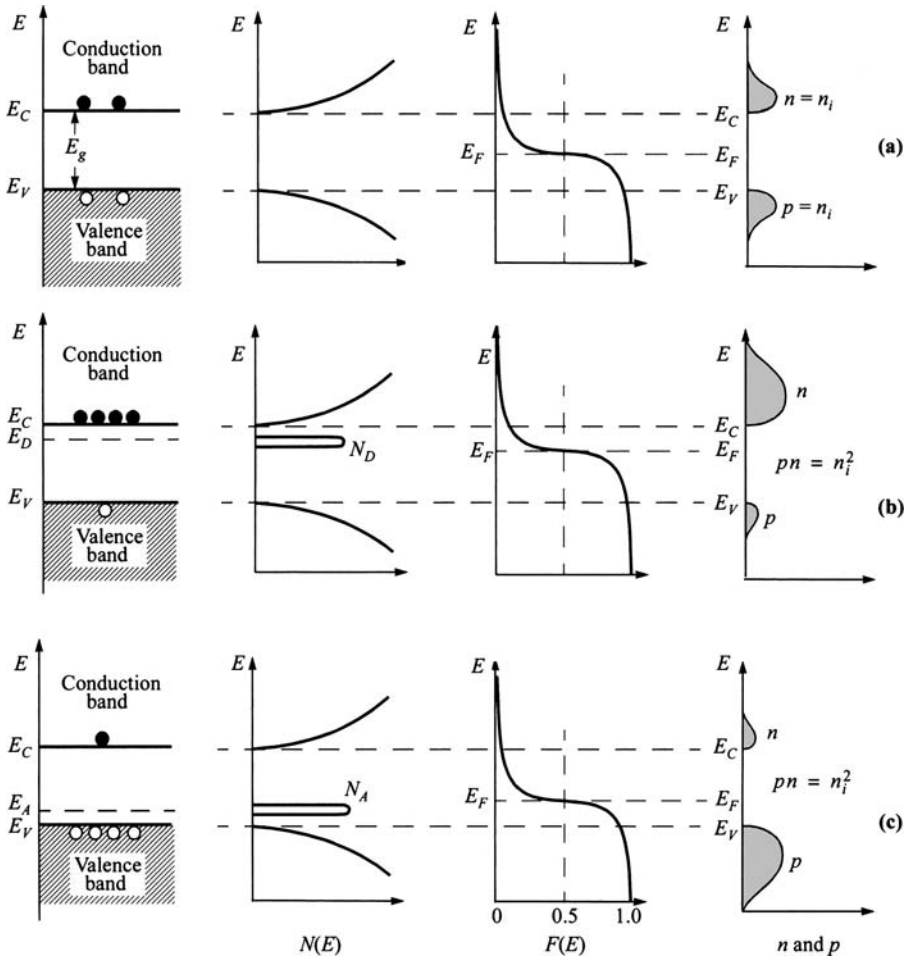


Fig. 11 Schematic band diagram, density of states, Fermi-Dirac distribution, and carrier concentrations for (a) intrinsic, (b) n-type, and (c) p-type semiconductors at thermal equilibrium. Note that $pn = n_i^2$ for all three cases.

$$N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \approx \frac{N_D}{1 + 2\exp[(E_F - E_D)/kT]}. \tag{38}$$

Thus for a set of given N_D, E_D, N_C, and T, the Fermi level E_F can be uniquely determined implicitly. Knowing E_F, the carrier concentrations n can be calculated. Equation 38 can also be solved graphically. In Fig. 12, the values of n and N_D^+ are plotted as a function of E_F. Where the two curves meet determines the position of E_F.

Without solving for Eq. 38, it can be shown that for $N_D \gg \frac{1}{2}N_C\exp[-(E_C-E_D)/kT] \gg N_A$, the electron concentration can be approximated by[5]

$$n \approx \sqrt{\frac{N_D N_C}{2}}\exp\left[-\frac{(E_C-E_D)}{2kT}\right]. \tag{39}$$

For compensated n-type material ($N_D > N_A$) with nonnegligible acceptor concentration, when $N_A \gg \frac{1}{2}N_C\exp[-(E_C-E_D)/kT]$, the approximate expression for the electron density is then

$$n \approx \left(\frac{N_D-N_A}{2N_A}\right)N_C\exp\left[-\frac{(E_C-E_D)}{kT}\right]. \tag{40}$$

Figure 13 shows a typical example, where n is plotted as a function of the reciprocal temperature. At high temperatures we have the intrinsic range since $n \approx p \approx n_i \gg N_D$. At medium temperatures, $n \approx N_D$. At very low temperatures most impurities are frozen out and the slope is given by either Eq. 39 or Eq. 40, depending on the compensation conditions. The electron density, however, remains essentially constant over a wide range of temperatures (≈100 to 500 K).

Figure 14 shows the Fermi level for Si and GaAs as a function of temperature and impurity concentration, as well as the dependence of the bandgap on temperature (see Fig. 6).

At relatively high temperatures, most donors and acceptors are ionized, so the neutrality condition can be approximated by

$$n + N_A = p + N_D. \tag{41}$$

Equations 29 and 41 can be combined to give the concentrations of electrons and holes. In an n-type semiconductor where $N_D > N_A$:

$$\begin{aligned} n_{no} &= \frac{1}{2}[(N_D-N_A) + \sqrt{(N_D-N_A)^2+4n_i^2}] \\ &\approx N_D \qquad \text{if } |N_D-N_A| \gg n_i \text{ or } N_D \gg N_A, \end{aligned} \tag{42}$$

$$p_{no} = \frac{n_i^2}{n_{no}} \approx \frac{n_i^2}{N_D}. \tag{43}$$

The Fermi level can be obtained from

$$n_{no} = N_D = N_C\exp\left(-\frac{E_C-E_F}{kT}\right) = n_i\exp\left(\frac{E_F-E_i}{kT}\right). \tag{44}$$

Similarly, the carrier concentrations in a p-type semiconductor ($N_A > N_D$) are given by

$$\begin{aligned} p_{po} &= \frac{1}{2}[(N_A-N_D) + \sqrt{(N_A-N_D)^2+4n_i^2}] \\ &\approx N_A \qquad \text{if } |N_A-N_D| \gg n_i \text{ or } N_A \gg N_D, \end{aligned} \tag{45}$$

$$n_{po} = \frac{n_i^2}{p_{po}} \approx \frac{n_i^2}{N_A}, \tag{46}$$

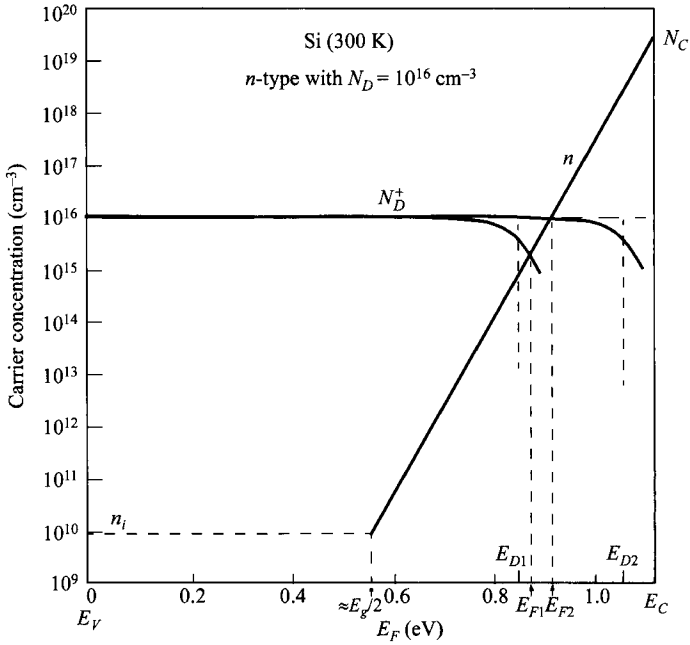


Fig. 12 Graphical method to determine the Fermi energy level E_F and electron concentration n, when ionization is not complete. Examples with two different values of impurity levels E_D are shown.

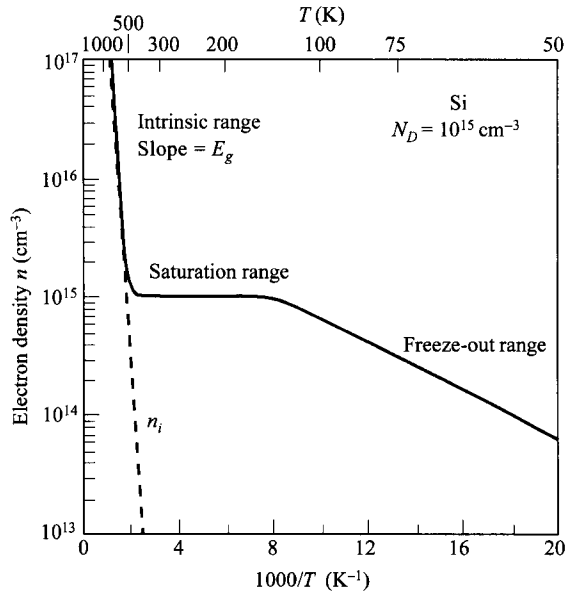


Fig. 13 Electron density as a function of temperature for a Si sample with donor impurity concentration of 10^{15} cm^{-3}. (After Ref. 5.)

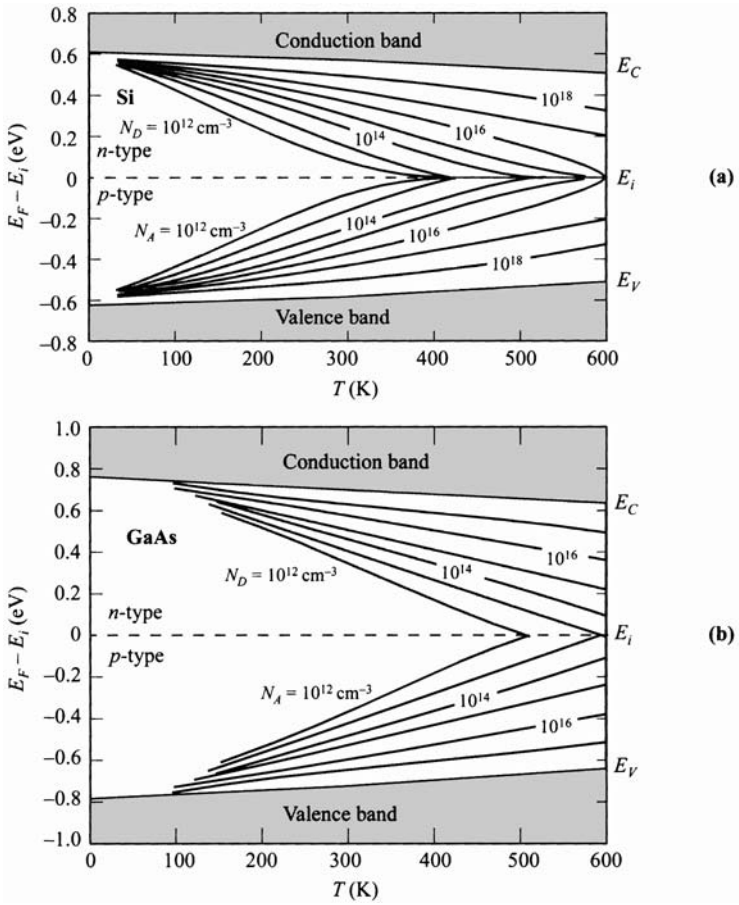


Fig. 14 Fermi level for (a) Si and (b) GaAs as a function of temperature and impurity concentration. The dependence of the bandgap on temperature is also shown. (After Ref. 37.)

and

$$p_{po} = N_A = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) = n_i \exp\left(\frac{E_i - E_F}{kT}\right). \tag{47}$$

In the formulas above, the subscripts n and p refer to the type of semiconductors, and the subscript "o" refers to the thermal equilibrium condition. For n-type semiconductors the electron is referred to as the majority carrier and the hole as the minority carrier, since the electron concentration is the larger of the two. The roles are reversed for p-type semiconductors.

1.5 CARRIER-TRANSPORT PHENOMENA

1.5.1 Drift and Mobility

At low electric fields, the drift velocity v_d is proportional to the electric field strength $\mathscr{E}$ and the proportionality constant is defined as the mobility μ in cm^2/V-s, or

$$v_d = \mu \mathscr{E}. \tag{48}$$

For nonpolar semiconductors, such as Ge and Si, the presence of acoustic phonons (see Section 1.6.1) and ionized impurities results in carrier scattering that significantly affects the mobility. The mobility from interaction with acoustic phonon of the lattice, μ_l, is given by[38]

$$\mu_l = \frac{\sqrt{8\pi}\, q\hbar^4 C_l}{3E_{ds}^2 m_c^{*5/2}(kT)^{3/2}} \propto \frac{1}{m_c^{*5/2}T^{3/2}} \tag{49}$$

where C_l is the average longitudinal elastic constant of the semiconductor, E_{ds} the displacement of the band edge per unit dilation of the lattice, and m_c^* the conductivity effective mass. From Eq. 49 mobility decreases with the temperature and with the effective mass.

The mobility from ionized impurities μ_i can be described by[39]

$$\mu_i = \frac{64\sqrt{\pi}\,\varepsilon_s^2(2kT)^{3/2}}{N_I q^3 m^{*1/2}}\left\{\ln\left[1+\left(\frac{12\pi\varepsilon_s kT}{q^2 N_I^{1/3}}\right)^2\right]\right\}^{-1} \propto \frac{T^{3/2}}{N_I m^{*1/2}} \tag{50}$$

where N_I is the ionized impurity density. The mobility is expected to decrease with the effective mass but to increase with the temperature because carriers with higher thermal velocity are less deflected by Coulomb scattering. Note the common dependence of the two scattering events on the effective mass but opposite dependence on temperature. The combined mobility, which includes the two mechanisms above, is given by the Matthiessen rule

$$\mu = \left(\frac{1}{\mu_l} + \frac{1}{\mu_i}\right)^{-1}. \tag{51}$$

In addition to the scattering mechanisms discussed above, other mechanisms also affect the actual mobility. For example, (1) the intravalley scattering in which an electron is scattered within an energy ellipsoid (Fig. 5) and only long-wavelength phonons (acoustic phonons) are involved; and (2) the intervalley scattering in which an electron is scattered from the vicinity of one minimum to another minimum and an energetic phonon (optical phonon) is involved. For polar semiconductors such as GaAs, polar-optical-phonon scattering is significant.

Qualitatively, since mobility is controlled by scattering, it can also be related to the mean free time τ_m or mean free path λ_m by

$$\mu = \frac{q\tau_m}{m^*} = \frac{q\lambda_m}{\sqrt{3kTm^*}}. \tag{52}$$

The last term uses the relationship

$$\lambda_m = v_{th}\tau_m \tag{53}$$

where v_{th} is the thermal velocity given by

$$v_{th} = \sqrt{\frac{3kT}{m^*}}. \tag{54}$$

For multiple scattering mechanisms, the effective mean free time is derived from the individual mean free times of scattering events by

$$\frac{1}{\tau_m} = \frac{1}{\tau_{m1}} + \frac{1}{\tau_{m2}} + \ldots \tag{55}$$

It can be seen that Eqs. 51 and 55 are equivalent.

Figure 15 shows the measured mobilities of Si and GaAs versus impurity concentrations at room temperature. As the impurity concentration increases (at room temperature most shallow impurities are ionized) the mobility decreases, as predicted by Eq. 50. Also for larger m^*, μ decreases; thus for a given impurity concentration the electron mobilities for these semiconductors are larger than the hole mobilities (Appendixes F and G list the effective masses).

Figure 16 shows the temperature effect on mobility for n-type and p-type silicon samples. For lower impurity concentrations the mobility is limited by phonon scattering and it decreases with temperature as predicted by Eq. 49. The measured slopes, however, are different from –3/2 because of other scattering mechanisms. For these

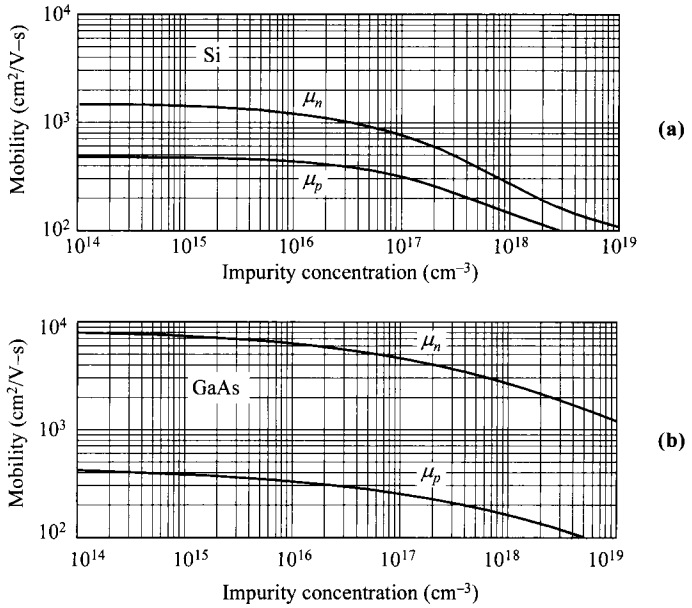


Fig. 15 Drift mobility of (a) Si (After Ref. 40.) and (b) GaAs at 300 K vs. impurity concentration (after Ref. 11).

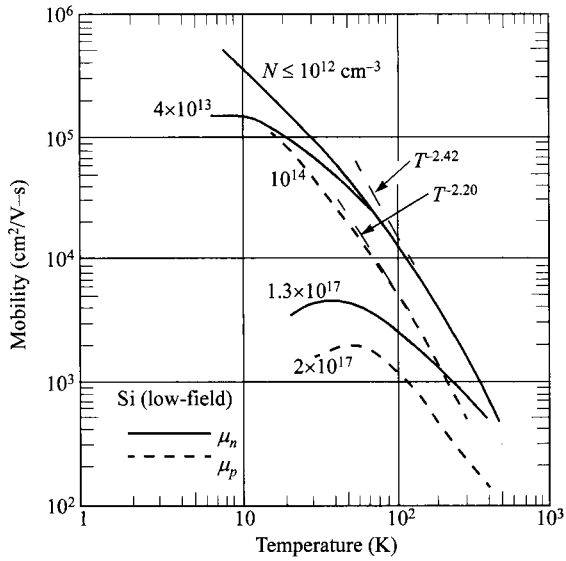


Fig. 16 Mobility of electrons and holes in Si as a function of temperature. (After Ref. 41.)

pure materials, near room temperature, the mobility varies as $T^{-2.42}$ and $T^{-2.20}$ for n- and p-type Si, respectively; and as $T^{-1.0}$ and $T^{-2.1}$ for n- and p-type GaAs (not shown), respectively.

The mobilities discussed above are the conductivity mobilities, which have been shown to be equal to the drift mobilities.[34] They are, however, different from but related to the Hall mobilities considered in the next section.

1.5.2 Resistivity and Hall Effect

For semiconductors with both electrons and holes as carriers, the drift current under an applied field is given by

$$\begin{aligned} J &= \sigma\mathscr{E} \\ &= q(\mu_n n + \mu_p p)\mathscr{E} \end{aligned} \tag{56}$$

where σ is the conductance

$$\sigma = \frac{1}{\rho} = q(\mu_n n + \mu_p p) \tag{57}$$

and ρ is the resistivity. If $n \gg p$, as in n-type semiconductors,

$$\rho = \frac{1}{q\mu_n n} \tag{58}$$

and

$$\sigma = q\mu_n n . \tag{59}$$

The most-common method for measuring resistivity is the four-point probe method (insert, Fig. 17),[42,43] A small constant current is passed through the outer two probes and the voltage is measured between the inner two probes. For a thin wafer with thickness W much smaller than either a or d, the sheet resistance $R_\square$ is given by

$$R_\square = \frac{V}{I} \cdot \mathrm{CF} \qquad \Omega/\square \tag{60}$$

where CF is the correction factor shown in Fig. 17. The resistivity is then

$$\rho = R_\square W \qquad \Omega\text{-cm}. \tag{61}$$

In the limit when $d \gg S$, where S is the probe spacing, the correction factor becomes $\pi/\ln 2$ (= 4.54).

Figure 18a shows the measured resistivity (at 300 K) as a function of the impurity concentration (n-type phosphorus and p-type boron) for silicon. Resistivity is not a linear function of concentration because mobility is not constant and usually decreases with increasing concentration. Figure 18b shows the measured resistivities for GaAs. We can obtain the impurity concentration of a semiconductor if its resistivity is known and vice versa. Note that the impurity concentration may be different from the carrier concentration because of incomplete ionization. For example, in a p-type silicon with 10^{17} cm^{-3} gallium acceptor impurities, unionized acceptors at room temperature make up about 23% (from Eq. 35, Figs. 10 and 14); in other words, the carrier concentration is only 7.7×10^{16} cm^{-3}.

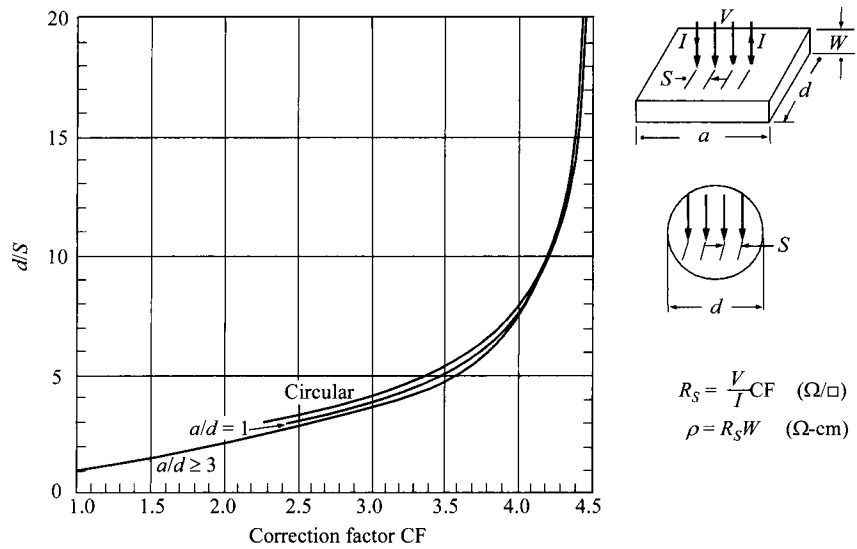


Fig. 17 Correction factor for measurement of resistivity using a four-point probe. (After Ref. 42.)

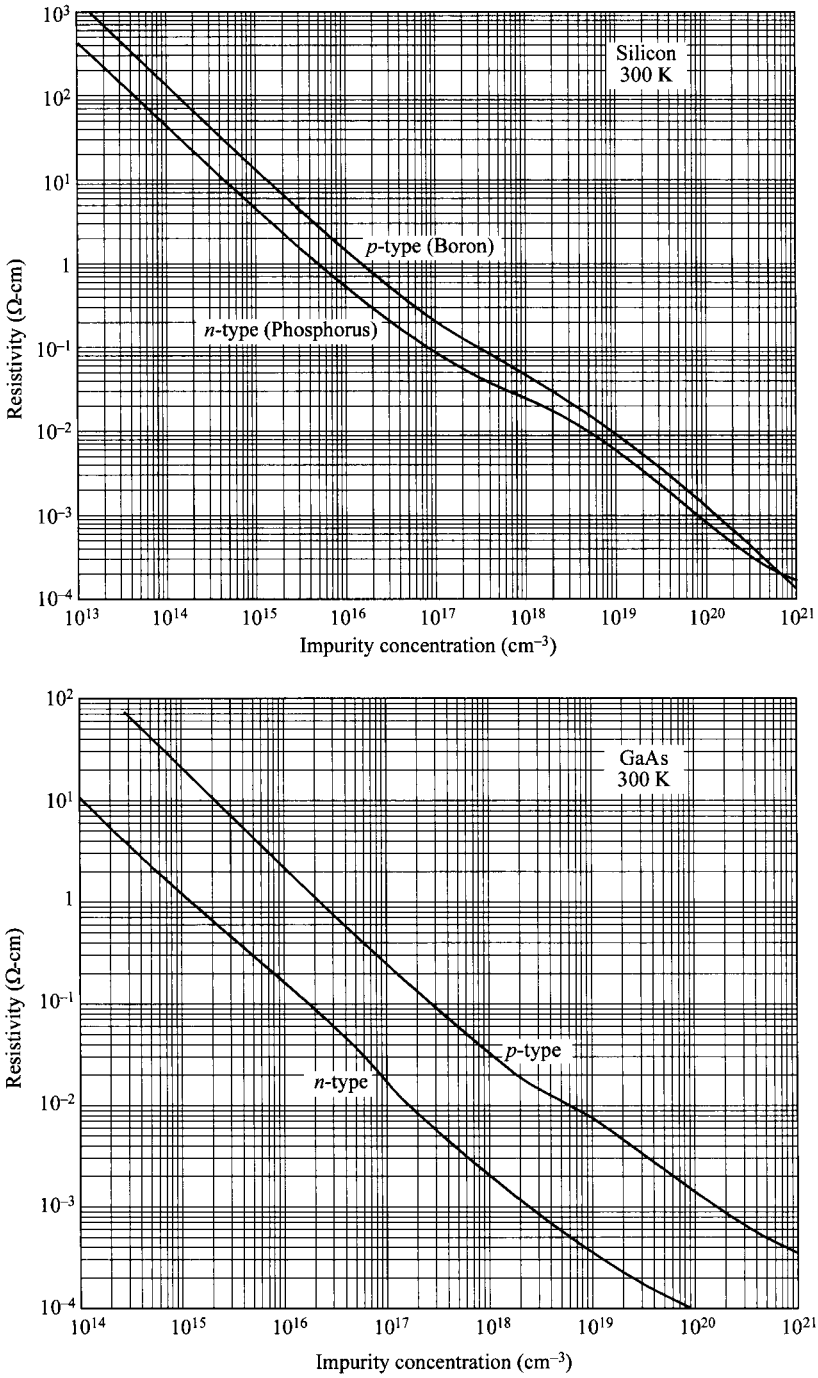


Fig. 18 Resistivity vs. impurity concentration at 300 K for (a) silicon (after Ref. 40) and (b) GaAs (after Ref. 35).

Hall Effect. Measurement of the resistivity only gives the product of the mobility and carrier concentration. To measure each parameter directly, the most-common method uses the Hall effect. The Hall effect is named after the scientist who made the discovery in 1879.[44] Even today it remains one of the most fascinating phenomena and is both fundamentally interesting and practical. Examples include the recent study of the fractional quantum Hall effect and the applications as magnetic-field sensors. The Hall effect is used in common practice to measure certain properties of semiconductors: namely, the carrier concentration (even down to a low level of 10^{12} cm^{-3}), the mobility, and the type (n or p). It is an important analytical tool since a simple conductance measurement can only give the product of concentration and mobility, and the type remains unknown.

Figure 19 shows the basic setup where an electric field is applied along the x-axis and a magnetic field is applied along the z-axis.[45] Consider a p-type sample. The Lorentz force exerts an average downward force on the holes

$$\text{Lorentz force} = q\upsilon_x \times \mathscr{B}_z, \tag{62}$$

and the downward-directed current causes a piling up of holes at the bottom side of the sample, which in turn gives rise to an electric field $\mathscr{E}_y$. Since there is no net current along the y-direction in the steady state, the electric field along the y-axis (Hall field) balances exactly the Lorentz force such that the carriers travel in a path parallel to the applied field $\mathscr{E}_x$. (For n-type material, electrons also pile up at the bottom surface, setting up a voltage of opposite polarity.)

The carrier velocity υ is related to the current density by

$$J_x = q\upsilon_x p. \tag{63}$$

Since for each carrier the Lorentz force must be equal to the force exerted by the Hall field,

$$q\mathscr{E}_y = q\upsilon_x \mathscr{B}_z, \tag{64}$$

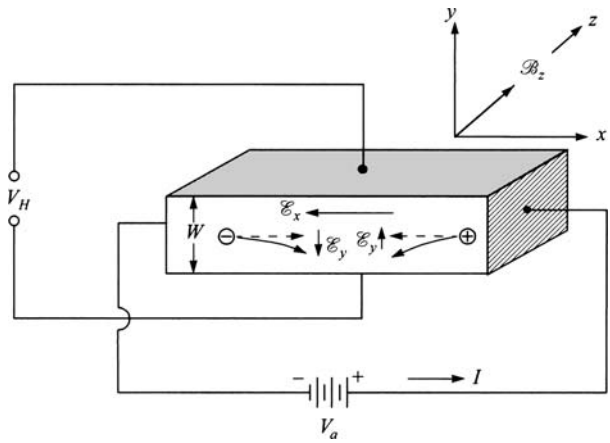


Fig. 19 Basic setup to measure carrier concentration using the Hall effect.

this Hall voltage can be measured externally and is given by

$$V_H = \mathscr{E}_y W = \frac{J_x \mathscr{B}_z W}{qp}. \tag{65}$$

When scattering is taken into account, the Hall voltage becomes

$$V_H = R_H J_x \mathscr{B}_z W \tag{66}$$

where R_H is the Hall coefficient and is given by

$$R_H = \frac{r_H}{qp} \qquad p \gg n, \tag{67a}$$

$$R_H = -\frac{r_H}{qn} \qquad n \gg p, \tag{67b}$$

with a Hall factor

$$r_H \equiv \frac{\langle \tau_m^2 \rangle}{\langle \tau_m \rangle^2}. \tag{68}$$

Thus, the carrier concentration and carrier type (electrons or holes from the polarity of the Hall voltage) can be obtained directly from the Hall measurement, provided that one type of carrier dominates and r_H is known.

Equation 67a or b also assumes conduction by a single type of carrier. A more-general solution is described by[5]

$$R_H = \frac{r_H}{q} \frac{\mu_p^2 p - \mu_n^2 n}{(\mu_p p + \mu_n n)^2}. \tag{69}$$

It can be seen in Eq. 69 that the sign of R_H and thus V_H reveals the majority type of the semiconductor sample.

The Hall mobility μ_H is defined as the product of the Hall coefficient and conductivity:

$$\mu_H = |R_H| \sigma. \tag{70}$$

The Hall mobility should be distinguished from the drift mobility μ_n (or μ_p) as given in Eq. 59 which does not contain the Hall factor r_H. Their relationship is given by

$$\mu_H = r_H \mu. \tag{71}$$

The parameter τ_m for the Hall factor is the mean free time between carrier collisions, which depends on the carrier energy. For example, for semiconductors with spherical constant-energy surfaces, $\tau_m \propto E^{-1/2}$ for phonon scattering and $\tau_m \propto E^{3/2}$ for ionized impurity scattering. In general,

$$\tau_m = C_1 E^{-s}, \tag{72}$$

where C_1 and s are constants. From Boltzmann distribution for nondegenerate semiconductors, the average value of the nth power of τ_m is

$$\langle \tau_m^n \rangle = \int_0^\infty \tau_m^n E^{3/2} \exp\left(-\frac{E}{kT}\right) dE \Big/ \int_0^\infty E^{3/2} \exp\left(-\frac{E}{kT}\right) dE, \tag{73}$$

so that using the general form of τ_m, we obtain

$$\langle \tau_m^2 \rangle = \frac{C_1^2 (kT)^{-2s} \Gamma(\frac{5}{2} - 2s)}{\Gamma(\frac{5}{2})} \tag{74}$$

and

$$\langle \tau_m \rangle = \frac{C_1 (kT)^{-s} \Gamma(\frac{5}{2} - s)}{\Gamma(\frac{5}{2})} \tag{75}$$

where $\Gamma(n)$ is the gamma function defined as

$$\Gamma(n) \equiv \int_0^\infty x^{n-1} e^{-x} dx. \tag{76}$$

[$\Gamma(1/2) = \sqrt{\pi}$.] From the expression above we obtain $r_H = 3\pi/8 = 1.18$ for phonon scattering and $r_H = 315\pi/512 = 1.93$ for ionized-impurity scattering. In general r_H lies in the range of 1–2. At very high magnetic fields, it approaches a value slightly below unity.

In the preceding discussion the applied magnetic field is assumed to be small enough that there is no change in the resistivity of the sample. However, under strong magnetic fields, a significant increase in the resistivity is observed, the so-called magnetoresistance effect, resulting from carriers travelling in a path that deviates from the applied electric field. For spherical-energy surfaces the ratio of the incremental resistivity to the bulk resistivity at zero magnetic field is given by[5]

$$\frac{\Delta\rho}{\rho_0} = \left\{ \left[\frac{\Gamma^2(\frac{5}{2})\Gamma(\frac{5}{2} - 3s)}{\Gamma^3(\frac{5}{2} - s)} \right] \left(\frac{\mu_n^3 n + \mu_p^3 p}{\mu_n n + \mu_p p} \right) - \left[\frac{\Gamma(\frac{5}{2})\Gamma(\frac{5}{2} - 2s)}{\Gamma^2(\frac{5}{2} - s)} \right]^2 \left(\frac{\mu_n^2 n - \mu_p^2 p}{\mu_n n + \mu_p p} \right)^2 \right\} \mathscr{B}_z^2 . \tag{77}$$

The ratio is proportional to the square of the magnetic field component perpendicular to the direction of the current flow. For $n \gg p$, $(\Delta\rho/\rho_0) \propto \mu_n^2 \mathscr{B}_z^2$. A similar result can be obtained for the case $p \gg n$.

1.5.3 High-Field Properties

In the preceding sections we considered the effect of low electric field on the transport of carriers in semiconductors. In this section we briefly consider some special effects and properties of semiconductors when the electric field is increased to moderate and high levels.

As discussed in Section 1.5.1, at low electric fields the drift velocity in a semiconductor is proportional to the field and the proportionality constant is the mobility that

is independent of the electric field. When the fields are sufficiently large, however, nonlinearities in mobility and, in some cases, saturation of drift velocity are observed. At still larger fields, impact ionization occurs. First, we consider the nonlinear mobility.

At thermal equilibrium the carriers both emit and absorb phonons and the net rate of exchange of energy is zero. The energy distribution at thermal equilibrium is Maxwellian. In the presence of an electric field the carriers acquire energy from the field and lose it to phonons by emitting more phonons than are absorbed. At moderately high fields, the most frequent scattering events involve in the emission of acoustic phonons. Thus, the carriers on average acquire more energy than they have at thermal equilibrium. As the field increases, the average energy of the carriers also increases and they acquire an effective temperature T_e that is higher than the lattice temperature T. Balancing the rate at which energy is transferred from the field to the carriers by an equal rate of energy loss to the lattice, we obtain from the rate equation, for Ge and Si (semiconductors without transferred-electron effect):[3]

$$\frac{T_e}{T} = \frac{1}{2}\left[1 + \sqrt{1 + \frac{3\pi}{8}\left(\frac{\mu_0 \mathscr{E}}{c_s}\right)^2}\right] \tag{78}$$

and

$$v_d = \mu_0 \mathscr{E} \sqrt{\frac{T}{T_e}} \tag{79}$$

where μ_0 is the low-field mobility, and c_s the velocity of sound. For moderate field strength when $\mu_0 \mathscr{E}$ is comparable to c_s, the carrier velocity v_d starts to deviate from being linearly dependent on the applied field, by a factor of $\sqrt{T/T_e}$. Finally at sufficiently high fields, carriers start to interact with optical phonons and Eq. 78 is no longer accurate. The drift velocities for Ge and Si become less and less dependent on the applied field and approach a saturation velocity

$$v_s = \sqrt{\frac{8E_p}{3\pi m_0}} \approx 10^7 \text{ cm/s} \tag{80}$$

where E_p is the optical-phonon energy (listed in Appendix G).

To eliminate the discontinuity between the regimes covered by Eqs. 78–80, a single empirical formula is often used to describe the whole range, from low-field drift velocity to velocity saturation:[46]

$$v_d = \frac{\mu_0 \mathscr{E}}{[1 + (\mu_0 \mathscr{E}/v_s)^{C_2}]^{1/C_2}}. \tag{81}$$

The constant C_2 has a value near two for electrons and one for holes, and it is a function of temperature.

The velocity-field relationship is more complicated for GaAs, and we must consider its band structure (Fig. 4). A high-mobility valley ($\mu \approx$ 4,000 to 8,000 cm^2/V-s) is located at the Brillouin zone center, and a low-mobility satellite valley ($\mu \approx$

100 cm^2/V-s) along the ⟨111⟩-axes,[47] about 0.3 eV higher in energy. The difference in mobility is due to the different electron effective masses (Eq. 52): $0.063m_0$ in the lower valley and about $0.55m_0$ in the upper valley. As the field increases, the electrons in the lower valley can be excited to the normally unoccupied upper valley, resulting in a differential negative resistance in GaAs. The intervalley transfer mechanism, called transferred-electron effect, and the velocity-field relationship are considered in more detail in Chapter 10.

Figure 20a shows the measured room-temperature drift velocities versus electric field for high-purity (low impurity concentration) Si and GaAs. For high-level impurity dopings, the drift velocity or mobility at low fields is decreased due to impurity scattering. However, the velocity at high fields is essentially independent of impurity dopings, and it approaches a saturation value.[52] For Si the saturation velocities v_s for electrons and holes are about 1×10^7 cm/s. For GaAs a wide range of negative differential mobility exists for fields above 3×10^3 V/cm, and the high-field saturation velocity approaches 6×10^6 cm/s. Figure 20b shows the temperature dependence of electron saturation velocity. As the temperature increases, the saturation velocities for both Si and GaAs decrease.

Up to now, the drift velocities discussed are for steady-state condition where carriers go through enough scattering events to get to their equilibrium values. In modern devices, the critical dimension where carriers transit across becomes smaller and smaller. When this dimension becomes comparable to or shorter than the mean free path, *ballistic transport* is said to occur before carriers start to be scattered. Figure 21 shows the drift velocity as a function of distance. Without scattering, the velocity increases with time (and distance) according to $\approx q\mathscr{E}t/m^*$. At high fields, drift velocity can attain a higher value momentarily than that at steady state, within a short space (of the order of mean free path) and time (of the order of mean free time). This phenomenon is called velocity overshoot. (In literature, confusion might arise when the peak velocity of GaAs shown in Fig. 20a—the transferred-electron effect, is also called velocity overshoot.) At low fields, the acceleration of velocity is lower and when scattering starts to occur, the attained velocity is not that high and so velocity overshoot does not occur. Note that this shape of velocity overshoot is similar to that in the transferred-electron effect but the abscissa here is distance (or time) while that in the latter is electric field.

We next consider impact ionization. When the electric field in a semiconductor is increased above a certain value, the carriers gain enough energy to excite electron-hole pairs by a process called impact ionization. The threshold energy obviously has to be larger than the bandgap. This multiplication process is characterized by an ionization rate α defined as the number of electron-hole pairs generated by a carrier per unit distance traveled (Fig. 22). So for a primary carrier of electron traveling with a velocity v_n,

$$\alpha_n = \frac{1}{n}\frac{dn}{d(tv_n)} = \frac{1}{nv_n}\frac{dn}{dt}. \tag{82}$$

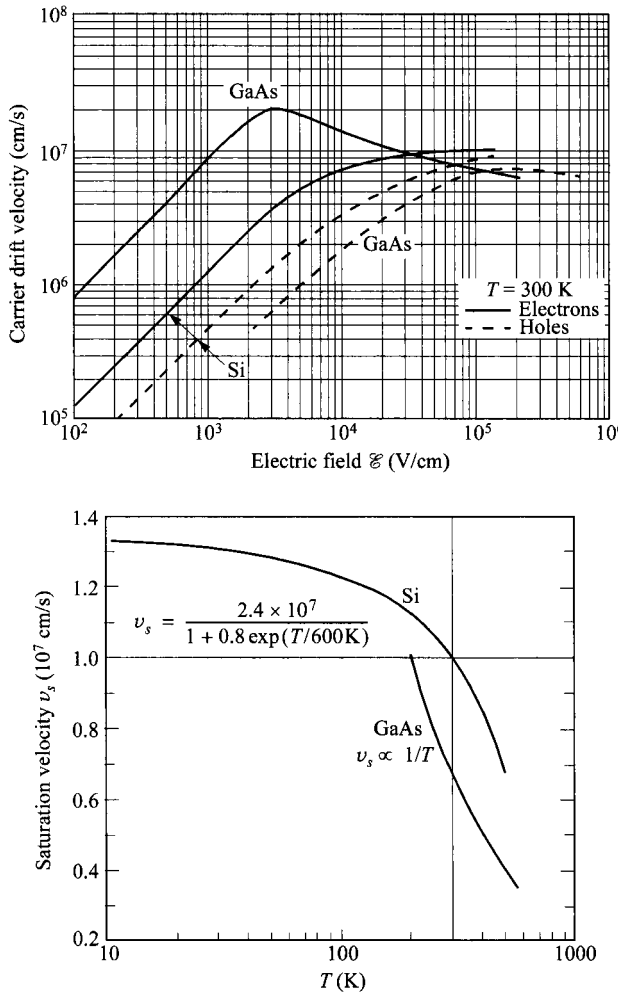


Fig. 20 (a) Measured carrier velocity versus electric field for high-purity Si and GaAs. For highly doped samples, the low-field velocities (mobilities) are lower than indicated here. In the high-field region, however, the velocity is essentially independent of dopings. (After Refs. 41, 48, 49, and 50.) (b) Saturated electron velocity versus temperature in Si and GaAs. (After Refs. 41 and 51.)

Considering both electrons and holes, the generation rate at any fixed location is given by

$$\frac{dn}{dt} = \frac{dp}{dt} = \alpha_n n v_n + \alpha_p p v_p = \frac{\alpha_n J_n}{q} + \frac{\alpha_p J_p}{q} \quad . \tag{83}$$

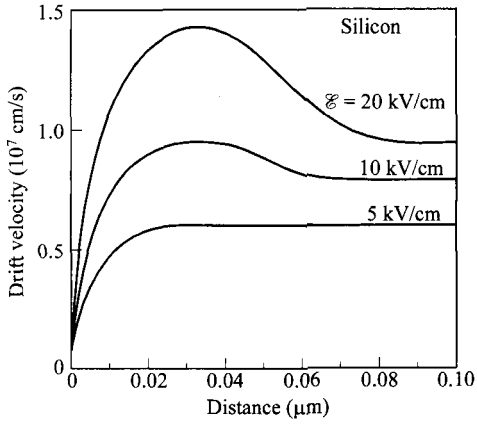


Fig. 21 Velocity overshoot in ultrashort distance. Similar behavior can be observed when the abscissa of distance is replaced with time. Example is for silicon. (After Ref. 53.)

Conversely, at any given time, the carrier density or current varies with distance and can be shown to be:

$$\frac{dJ_n}{dx} = \alpha_n J_n + \alpha_p J_p, \tag{84a}$$

$$\frac{dJ_p}{dx} = -\alpha_n J_n - \alpha_p J_p. \tag{84b}$$

The total current $(J_n + J_p)$ remains constant over distance and $dJ_n/dx = -dJ_p/dx$.

The ionization rates α_n and α_p are strongly dependent on the electric field. A physical expression for the ionization rate is given by[54]

$$\alpha(\mathscr{E}) = \frac{q\mathscr{E}}{E_I} \exp\left\{ - \frac{\mathscr{E}_I}{\mathscr{E}[1 + (\mathscr{E}/\mathscr{E}_p)] + \mathscr{E}_T} \right\} \tag{85}$$

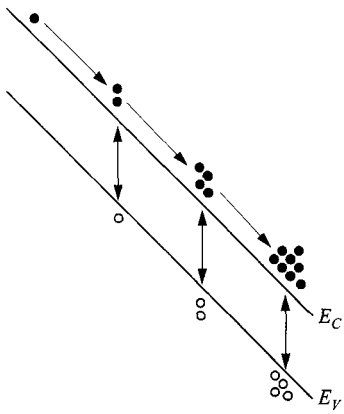


Fig. 22 Multiplication of electrons and holes from impact ionization, due to electrons (α_n) in this example ($\alpha_p = 0$).

where E_I is the high-field effective ionization threshold energy, and $\mathscr{E}_T$, $\mathscr{E}_p$, and $\mathscr{E}_I$ are threshold fields for carriers to overcome the decelerating effects of thermal, optical-phonon, and ionization scattering, respectively. For Si, the value of E_I is found to be 3.6 eV for electrons and 5.0 eV for holes. Over a limited field range, Equation 85 can be reduced to

$$\alpha(\mathscr{E}) = \frac{q\mathscr{E}}{E_I}\exp\left(-\frac{\mathscr{E}_I}{\mathscr{E}}\right), \qquad \text{if} \qquad \mathscr{E}_p > \mathscr{E} > \mathscr{E}_T\,, \tag{86}$$

or

$$\alpha(\mathscr{E}) = \frac{q\mathscr{E}}{E_I}\exp\left(-\frac{\mathscr{E}_I\mathscr{E}_p}{\mathscr{E}^2}\right), \qquad \text{if } \mathscr{E} > \mathscr{E}_p \text{ and } \mathscr{E} > \sqrt{\mathscr{E}_p\mathscr{E}_T}. \tag{87}$$

Figure 23a shows the experimental results of the ionization rates for Ge, Si, SiC, and GaN. Figure 23b shows the measured ionization rates of GaAs and a few other binary and ternary compounds. These results are obtained by using photomultiplication measurements on *p-n* junctions. Note that for certain semiconductors, such as GaAs, the ionization rate is a function of crystal orientation. There is also a general trend that the ionization rate decreases with increasing bandgap. It is for this reason that materials of higher bandgaps generally yield higher breakdown voltage. Note that Eq. 86 is applicable to most semiconductors shown in Fig. 23, except GaAs and GaP, for which Eq. 87 is applicable.

At a given electric field, the ionization rate decreases with increasing temperature. Figure 24 shows the theoretical predicted electron ionization rates in silicon as an example, together with the experimental results at three different temperatures.

1.5.4 Recombination, Generation, and Carrier Lifetimes

Whenever the thermal-equilibrium condition of a semiconductor system is disturbed (i.e., $pn \neq n_i^2$), processes exist to restore the system to equilibrium (i.e., $pn = n_i^2$). These processes are recombination when $pn > n_i^2$ and thermal generation when $pn < n_i^2$. Figure 25a illustrates the band-to-band electron-hole recombination. The energy of an electron in transition from the conduction band to the valence band is conserved by emission of a photon (radiative process) or by transfer of the energy to another free electron or hole (Auger process). The former process is the inverse of direct optical absorption, and the latter is the inverse of impact ionization.

Band-to-band transitions are more probable for direct-bandgap semiconductors which are more common among III-V compounds. For this type of transition, the recombination rate is proportional to the product of electron and hole concentrations, given by

$$R_e = R_{ec}pn. \tag{88}$$

The term R_{ec}, called the *recombination coefficient*, is related to the thermal generation rate G_{th} by

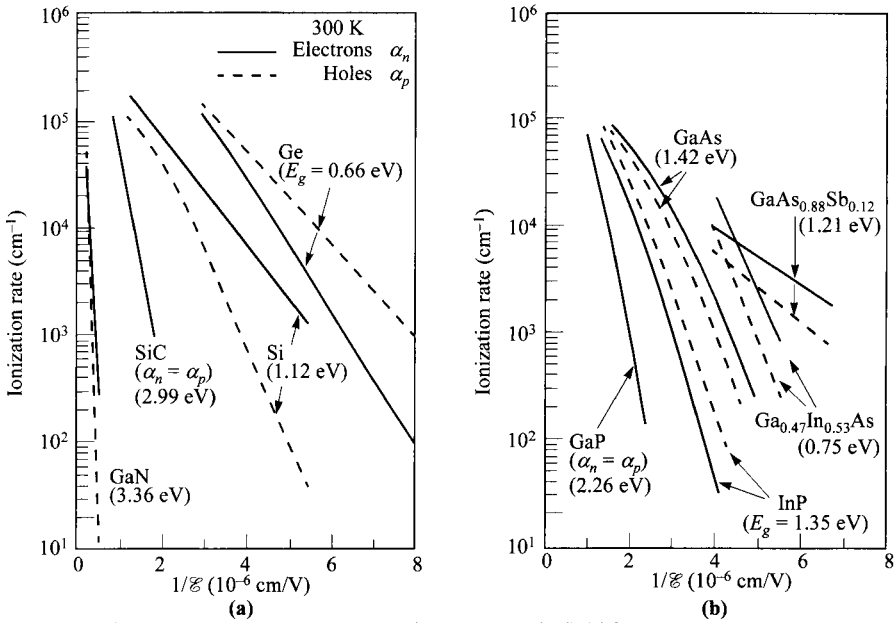


Fig. 23 Ionization rates at 300 K versus reciprocal electric field for Si, GaAs, and some IV-IV and III-V compound semiconductors. (After Refs. 55–65.)

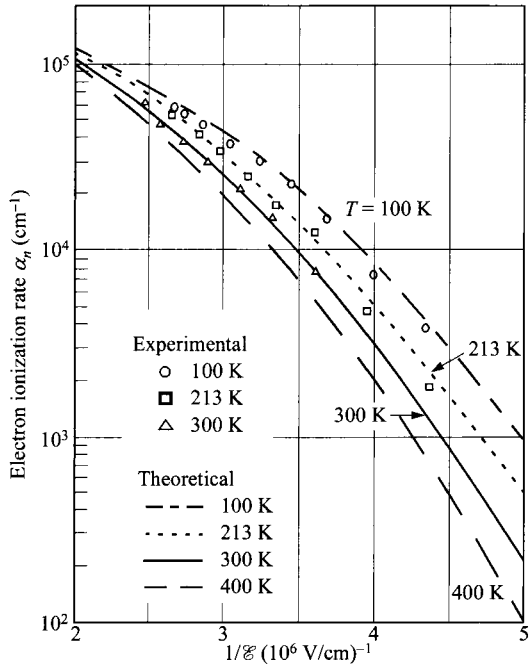


Fig. 24 Electron ionization rate versus reciprocal electric field in Si for four temperatures. (After Ref. 66.)

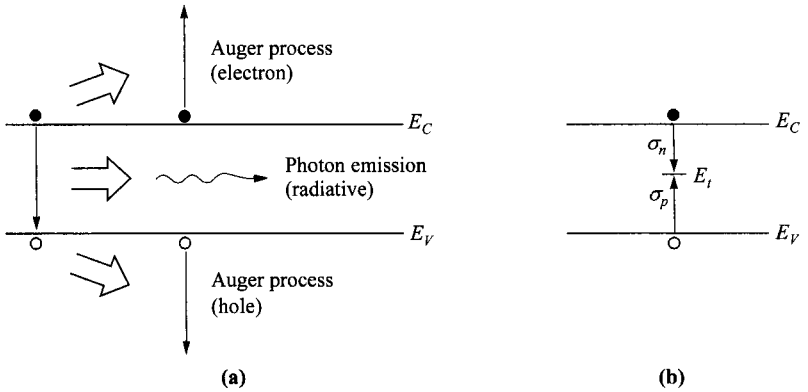


Fig. 25 Recombination processes (the reverse are generation processes). (a) Band-to-band recombination. Energy is exchanged to a radiative or Auger process. (b) Recombination through single-level traps (nonradiative).

$$R_{ec} = \frac{G_{th}}{n_i^2}. \tag{89}$$

R_{ec} is a function of temperature and is also dependent on the band structure of the semiconductor. A direct-bandgap semiconductor, being more efficient in band-to-band transitions, has a much larger R_{ec} ($\approx 10^{-10}$ cm^3/s) than an indirect-bandgap semiconductor ($\approx 10^{-15}$ cm^3/s). In thermal equilibrium, since $pn = n_i^2$, $R_e = G_{th}$ and the net transition rate $U (= R_e - G_{th})$ equals zero. Under low-level injection, defined as the case where the excess carriers $\Delta p = \Delta n$ are fewer than the majority carriers, for an n-type material $p_n = p_{no} + \Delta p$ and $n_n \approx N_D$, the net transition rate is given by

$$\begin{aligned} U &= R_e - G_{th} = R_{ec}(pn - n_i^2) \\ &\approx R_{ec}\Delta p N_D \equiv \frac{\Delta p}{\tau_p} \end{aligned} \tag{90}$$

where the carrier lifetime for holes

$$\tau_p = \frac{1}{R_{ec}N_D}, \tag{91a}$$

and in p-type material,

$$\tau_n = \frac{1}{R_{ec}N_A}. \tag{91b}$$

However, in indirect-bandgap semiconductors such as Si and Ge, the dominant transitions are indirect recombination/generation via bulk traps, of density N_t and energy E_t present within the bandgap (Fig. 25b). The single-level recombination can be described by two processes—electron capture and hole capture. The net transition rate can be described by the Shockley-Read-Hall statistics[67–69] as

$$U = \frac{\sigma_n \sigma_p \upsilon_{th} N_t (pn - n_i^2)}{\sigma_n \left[n + n_i \exp\left(\frac{E_t - E_i}{kT} \right) \right] + \sigma_p \left[p + n_i \exp\left(\frac{E_i - E_t}{kT} \right) \right]} \tag{92}$$

where σ_n and σ_p are the electron and hole capture cross sections, respectively. Without deriving this equation, some qualitative observations can be made on the final form. First, the net transition rate is proportional to $pn - n_i^2$, similar to Eq. 90, and the sign determines whether there is net recombination or generation. Second, U is maximized when $E_t = E_i$, indicating for an energy spectrum of bulk traps, only those near the mid-gap are effective recombination/generation centers. Considering only these traps, Eq. 92 is reduced to

$$U = \frac{\sigma_n \sigma_p \upsilon_{th} N_t (pn - n_i^2)}{\sigma_n (n + n_i) + \sigma_p (p + n_i)}. \tag{93}$$

Again for low-level injection in n-type semiconductors, the net recombination rate becomes

$$\begin{aligned} U &= \frac{\sigma_n \sigma_p \upsilon_{th} N_t [(p_{no} + \Delta p) n - n_i^2]}{\sigma_n n} \\ &\approx \sigma_p \upsilon_{th} N_t \Delta p \equiv \frac{\Delta p}{\tau_p} \end{aligned} \tag{94}$$

where

$$\tau_p = \frac{1}{\sigma_p \upsilon_{th} N_t}. \tag{95a}$$

Similarly for a p-type semiconductor, the electron lifetime is given by

$$\tau_n = \frac{1}{\sigma_n \upsilon_{th} N_t}. \tag{95b}$$

As expected, the lifetime arising from indirect transitions is inversely proportional to the trap density N_t, while in the previous case, the lifetime from direct transitions is inversely proportional to the doping concentration (Eqs. 91a and 91b).

For multiple-level traps, the recombination processes have gross qualitative features that are similar to those of the single-level case. However, the behavioral details are different, particularly in the high-level injection condition (i.e., where $\Delta n = \Delta p$ approaches the majority-carrier concentration), where the asymptotic lifetime is an average of the lifetimes associated with all the positively charged, negatively charged, and neutral trapping levels.

For high-level injection ($\Delta n = \Delta p > n$ and p), the carrier lifetime for band-to-band recombination becomes

$$\tau_n = \tau_p = \frac{1}{R_{ec} \Delta n}. \tag{96}$$

The lifetime resulting from traps can be derived from Eq. 93 to be

$$\tau_n = \tau_p = \frac{\sigma_n + \sigma_p}{\sigma_n \sigma_p v_{th} N_t}. \tag{97}$$

Comparing Eq. 97 to Eqs. 95a and 95b, the lifetime is actually higher with high-level injection. It is interesting to note that the lifetime due to band-to-band recombination decreases with injection level, while that due to trap recombination increases with injection level.

Equations 95a and 95b have been verified experimentally by using solid-state diffusion and high-energy radiation. Many impurities have energy levels close to the middle of the bandgap (Fig. 10). These impurities are efficient recombination centers. A typical example is gold in silicon;[70] the minority-carrier lifetime decreases linearly with the gold concentration over the range of 10^{14} to 10^{17} cm^{-3}, where τ decreases from about 2×10^{-6} s to 2×10^{-9} s. This effect is sometimes advantageous, as in some high-speed applications when a short lifetime to reduce the charge storage time is a desirable feature. Another method of shortening the minority-carrier lifetime is high-energy-particle irradiation, which causes displacement of host atoms and damage to the lattice. These, in turn, introduce energy levels in the bandgap. For example, electron irradiation in Si gives rise to an acceptor level at 0.4 eV above the valence band and a donor level at 0.36 eV below the conduction band. Also neutron irradiation creates an acceptor level at 0.56 eV; and deuteron irradiation gives rise to an interstitial state with an energy level 0.25 eV above the valence band. Similar results are obtained for Ge, GaAs, and other semiconductors. Unlike the solid-state diffusion, the radiation-induced trapping centers may be annealed out at relatively low temperatures.

When carriers are below their thermal-equilibrium values, i.e., $pn < n_i^2$, generation of carriers rather than recombination of excess carriers will occur. The generation rate can be found by starting with Eq. 93,

$$U = -\frac{\sigma_p \sigma_n v_{th} N_t n_i}{\sigma_p[1 + (p/n_i)] + \sigma_n[1 + (n/n_i)]} \equiv -\frac{n_i}{\tau_g} \tag{98}$$

where the generation carrier lifetime τ_g is equal to

$$\begin{aligned} \tau_g &= \frac{1 + (n/n_i)}{\sigma_p v_{th} N_t} + \frac{1 + (p/n_i)}{\sigma_n v_{th} N_t} \\ &= \left(1 + \frac{n}{n_i}\right)\tau_p + \left(1 + \frac{p}{n_i}\right)\tau_n. \end{aligned} \tag{99}$$

Depending on the electron and hole concentrations, the generation lifetime can be much longer than the recombination lifetime and has a minimum value of roughly twice that of the recombination lifetime, when both n and p are much smaller than n_i.

The minority-carrier lifetime τ has generally been measured using the photoconductive (PC) effect[71] or the photoelectromagnetic (PEM) effect[72]. The basic equation for the PC effect is given by

$$J_{PC} = q(\mu_n + \mu_p)\Delta n \mathscr{E}$$
$$= q(\mu_n + \mu_p)\frac{G_e}{\tau}\mathscr{E} \tag{100}$$

where J_{PC} is the incremental current density as a result of illumination with generation rate G_e, and $\mathscr{E}$ is the applied electric field along the sample. The quantity Δn is the incremental carrier density or the number of electron-hole pairs per volume created by the illumination, which equals the product of the generation rate G_e and the lifetime τ, or $\Delta n = \tau G_e$. For the PEM effect we measure the short-circuit current, which appears when a constant magnetic field $\mathscr{B}_z$ is applied perpendicular to the direction of incoming radiation. The current density is given by

$$J_{PEM} = q(\mu_n + \mu_p)\mathscr{B}_z \frac{D}{L_d}\tau G_e$$
$$= q(\mu_n + \mu_p)\mathscr{B}_z \sqrt{D\tau}\, G_e \tag{101}$$

where D and L_d [$\equiv (D\tau)^{1/2}$] are the diffusion coefficient and the diffusion length, to be discussed in the next section. Another approach to measure the carrier lifetime will be discussed in Section 1.8.2.

1.5.5 Diffusion

In the preceding section the excess carriers are uniform in space. In this section, we discuss the situations where excess carriers are introduced locally, causing a condition of nonuniform carriers. Examples are local injection of carriers from a junction, and nonuniform illumination. Whenever there exists a gradient of carrier concentration, a process of diffusion occurs by which the carriers migrate from the region of high concentration toward the region of low concentration, to drive the system toward a state of uniformity. This flow or flux of carriers, taking electrons as an example, is governed by the Fick's law,

$$\left.\frac{d\Delta n}{dt}\right|_x = -D_n \frac{d\Delta n}{dx}, \tag{102}$$

and is proportional to the concentration gradient. The proportionality constant is called the diffusion coefficient or diffusivity D_n. This flux of carriers constitutes a diffusion current, given by

$$J_n = qD_n \frac{d\Delta n}{dx}, \tag{103a}$$

and

$$J_p = -qD_p \frac{d\Delta p}{dx}. \tag{103b}$$

Physically, diffusion is due to random thermal motion of carriers as well as scattering. Because of this, we have

$$D = v_{th}\tau_m. \tag{104}$$

One also expects certain relationship between the diffusion coefficient and mobility. To derive such a relationship, we consider an *n*-type semiconductor with nonuniform doping concentration but without an external applied field. The zero net current necessitates that the drift current exactly balances the diffusion current,

$$qn\mu_n\mathscr{E} = -qD_n\frac{dn}{dx}. \tag{105}$$

In this case, the electric field is created by the nonuniform doping ($\mathscr{E} = dE_C/qdx$, and E_F is constant for equilibrium). Using Eq. 21 for n, we obtain

$$\begin{aligned}\frac{dn}{dx} &= \frac{-q\mathscr{E}}{kT}N_C\exp\left(-\frac{E_C - E_F}{kT}\right)\\ &= \frac{-q\mathscr{E}}{kT}n \quad .\end{aligned} \tag{106}$$

Substituting this into Eq. 105 will give the relationship

$$D_n = \left(\frac{kT}{q}\right)\mu_n. \tag{107a}$$

Similarly for *p*-type semiconductor, one can derive

$$D_p = \left(\frac{kT}{q}\right)\mu_p. \tag{107b}$$

These are known as the Einstein relation (valid for nondegenerate semiconductors). At 300 K, $kT/q = 0.0259$ V, and values of D are readily obtainable from the mobility results shown in Fig. 15.

Another parameter closely related to diffusion is the diffusion length,

$$L_d = \sqrt{D\tau}\,. \tag{108}$$

In common diffusion problems arising from some fixed injection source as a boundary condition, the resultant concentration profile is exponential in nature with distance, with a characteristic length of L_d. This diffusion length can also be viewed as the distance carriers can diffuse in a carrier lifetime before they are annihilated.

1.5.6 Thermionic Emission

Another current conduction mechanism is *thermionic emission*. It is a majority-carrier current and is always associated with a potential barrier. Note that the critical parameter is the barrier height, not the shape of the barrier. The most-common device is the Schottky-barrier diode or metal-semiconductor junction (see Chapter 3). Referring to Fig. 26, for the thermionic emission to be the controlling mechanism, the criterion is that collision or the drift-diffusion process within the barrier layer to be negligible. Equivalently, the barrier width has to be narrower than the mean free path, or in the case of a triangular barrier, the slope of the barrier be reasonably steep such that a drop in kT in energy is within the mean free path. In addition, after the carriers are injected over the barrier, the diffusion current in that region must not be the lim-

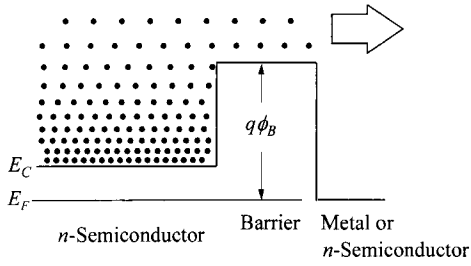


Fig. 26 Energy-band diagram showing thermionic emission of electrons over the barrier. Note that the shape of the barrier (shown as rectangular) does not matter.

iting factor. Therefore, the region behind the barrier must be another n-type semiconductor or a metal layer.

Due to Fermi-Dirac statistics, the density of electrons (for n-type substrate) decreases exponentially as a function of their energy above the conduction band edge. At any finite (nonzero) temperature, the carrier density at any finite energy is not zero. Of special interest here is the integrated number of carriers above the barrier height. This portion of the thermally generated carriers are no longer confined by the barrier so they contribute to the thermionic-emission current. The total electron current over the barrier is given by (see Chapter 3)

$$J = A^*T^2\exp\left(-\frac{q\phi_B}{kT}\right). \tag{109}$$

where ϕ_B is the barrier height, and

$$A^* \equiv \frac{4\pi q m^* k^2}{h^3} \tag{110}$$

is called the effective Richardson constant and is a function of the effective mass. The A^* can be further modified by quantum-mechanical tunneling and reflection.

1.5.7 Tunneling

Tunneling is a quantum-mechanical phenomenon. In classical mechanics, carriers are completely confined by the potential walls. Only those carriers with excess energy higher than the barriers can escape, as in the case of thermionic emission discussed above. In quantum mechanics, an electron can be represented by its wavefunction. The wavefunction does not terminate abruptly on a wall of finite potential height and it can penetrate into and through the barrier (Fig. 27). The probability of electron tunneling through a barrier of finite height and width is thus not zero.

To calculate the tunneling probability, the wavefunction ψ has to be determined from the Schrödinger equation

$$\frac{d^2\psi}{dx^2} + \frac{2m^*}{\hbar^2}[E - U(x)]\psi = 0. \tag{111}$$

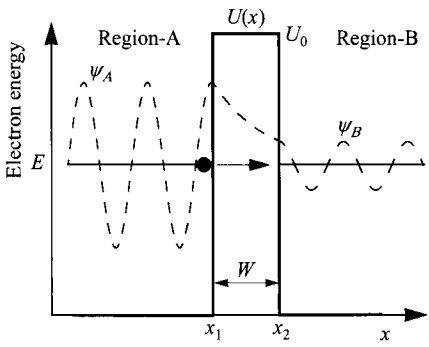


Fig. 27 Wavefunctions showing electron tunneling through a rectangular barrier.

In the case of a simple rectangular barrier of height U_0 and width W, ψ has a general form of $\exp(\pm ikx)$ where $k = \sqrt{2m^*(E - U_0)}/\hbar$. Note that for tunneling, the energy E is below the barrier U_0 so that the term within the square root is negative and k is imaginary. The solution of the wavefunctions and the tunneling probability are calculated to be

$$T_t = \frac{|\psi_B|^2}{|\psi_A|^2} = \left[1 + \frac{U_0^2 \sinh^2(|k|W)}{4E(U_0 - E)}\right]^{-1}$$
$$\approx \frac{16E(U_0 - E)}{U_0^2} \exp\left(-2\sqrt{\frac{2m^*(U_0 - E)}{\hbar^2}}\, W\right). \tag{112}$$

For more complicated barrier shapes, simplification of the Schrödinger equation is made by the WKB (Wentzel-Kramers-Brillouin) approximation if the potential $U(x)$ does not vary rapidly. The wavefunction now has a general form of $\exp\int ik(x)dx$. The tunneling probability can be calculated by

$$T_t = \frac{|\psi_B|^2}{|\psi_A|^2} \approx \exp\left\{-2\int_{x_1}^{x_2} |k(x)|\, dx\right\}$$
$$\approx \exp\left\{-2\int_{x_1}^{x_2} \sqrt{\frac{2m^*}{\hbar^2}[U(x) - E]}\, dx\right\}. \tag{113}$$

Together with known tunneling probability, the tunneling current J_t can be calculated from the product of the number of available carriers in the originating Region-A (Fig. 27), and the number of empty states in the destination Region-B,

$$J_t = \frac{qm^*}{2\pi^2\hbar^3} \int F_A N_A T_t (1 - F_B) N_B\, dE \tag{114}$$

where F_A, F_B, N_A, and N_B represent the Fermi-Dirac distributions and densities of states in the corresponding regions.

1.5.8 Space-Charge Effect

The space charge in a semiconductor is determined by both the doping concentrations and the free-carrier concentrations,

$$\rho = (p - n + N_D - N_A)q \quad . \tag{115}$$

In the neutral region of a semiconductor, $n = N_D$ and $p = N_A$, so that the space-charge density is zero. In the vicinity of a junction formed by different materials, dopant types, or doping concentrations, n and p could be smaller or larger than N_D and N_A, respectively. In the depletion approximation, n and p are assumed zero so that the space charge is equal to the majority-carrier doping level. Under bias, the carrier concentrations n and p can be increased beyond their values in equilibrium. When the injected n or p is larger than its equilibrium value as well as the doping concentration, the *space-charge effect* is said to occur. The injected carriers thus control the space charge and the electric-field profile. This results in a feedback mechanism where the field drives the current, which in turn sets up the field. The space-charge effect is more common in lightly doped materials, and it can occur outside the depletion region.

In the presence of a space-charge effect, if the current is dominated by the drift component of the injected carriers, it is called the *space-charge-limited current*. Since it is a drift current, it is given by, in the case of electron injection,

$$J = qnv. \tag{116}$$

The space charge again is determined by the injected carriers giving rise to the Poisson equation of the form

$$\frac{d^2\psi_i}{dx^2} = \frac{qn}{\varepsilon_s} \quad . \tag{117}$$

The carrier velocity v is related to the electric field by different functions, depending on the field strength. In the low-field mobility regime,

$$v = \mu\mathscr{E} \quad . \tag{118}$$

In the velocity-saturation regime, velocity v_s is independent of the field. In the limit of ultra-short sample or time scale, we have the ballistic regime where there is no scattering, and

$$v = \sqrt{\frac{2qV}{m^*}} \quad . \tag{119}$$

From Eqs. 116–119, the space-charge-limited current in the mobility regime (the Mott-Gurney law) can be solved to be (see Vol. 4 of Ref. 4)

$$J = \frac{9\varepsilon_s \mu V^2}{8L^3}, \tag{120}$$

in the velocity-saturation regime

$$J = \frac{2\varepsilon_s v_s V}{L^2}, \tag{121}$$

and in the ballistic regime (the Child-Langmuir law)

$$J = \frac{4\varepsilon_s}{9L^2}\left(\frac{2q}{m^*}\right)^{1/2} V^{3/2}. \tag{122}$$

Here L is the length of the sample in the direction of the current flow. Note that the voltage dependence is different in these regimes.

1.6 PHONON, OPTICAL, AND THERMAL PROPERTIES

In the preceding section we considered different carrier transport mechanisms in semiconductors. In this section we briefly consider other effects and properties of semiconductors that are important to the operation of semiconductor devices.

1.6.1 Phonon Spectra

Phonons are quanta of lattice vibrations, mainly resulting from the lattice thermal energy. Similar to photons and electrons, they have characteristic frequencies (or energy) and wave numbers (momentum or wavelengths). It is known that, as a demonstration in a one-dimensional lattice, with only nearest-neighbor coupling and two different masses m_1 and m_2 placed alternately, the frequencies of oscillation are given by[3]

$$\nu_{\pm} = \sqrt{\alpha_f}\left[\left(\frac{1}{m_1}+\frac{1}{m_2}\right) \pm \sqrt{\left(\frac{1}{m_1}+\frac{1}{m_2}\right)^2 - \frac{4\sin^2(k_{ph}a/2)}{m_1 m_2}}\right]^{1/2} \tag{123}$$

where α_f is the force constant of the Hooke's law, k_{ph} the phonon wave number, and a the lattice spacing. The frequency ν_- is proportional to k_{ph} near $k_{ph} = 0$. This branch is the acoustic branch, because it is the long-wavelength vibration of the lattice and the velocity ω/k is near that of sound in such a medium. The frequency ν_+ tends to be a constant $\approx [2\alpha_f(1/m_1 + 1/m_2)]^{1/2}$ as k_{ph} approaches zero. This branch, separated considerably from the acoustic mode, is the optical branch, because the frequency ν_+ is generally in the optical range. For the acoustic mode the two sublattices of the atoms with different masses move in the same direction, whereas for the optical mode they move in opposite directions.

The total number of acoustic modes is equal to the dimension times the number of atoms per cell. For a realistic three-dimensional lattice with one atom per primitive cell, such as a simple cubic, body-centered, or face-centered cubic lattice, only three acoustic modes exist. For a three-dimensional lattice with two atoms per primitive cell, such as Si and GaAs, three acoustic modes and three optical modes exist. Longitudinally polarized modes are modes with the displacement vectors of each atom along the direction of the wave vector; thus we have one longitudinal acoustic mode (LA) and one longitudinal optical mode (LO). Modes with atoms moving in the planes normal to the wave vector are called transversely polarized modes. We have two transverse acoustic modes (TA) and two transverse optical modes (TO).

Figure 28 shows the measured results for Si and GaAs in one of the crystal directions. The range of $k_{ph} = \pm\pi/a$ defines the Brillouin zone outside which the frequency-k_{ph} relationship repeats itself. Note that at small values of k_{ph}, for both LA and TA modes, the energies (or frequencies) are proportional to k_{ph}. The longitudinal optical phonon energy at $k_{ph} = 0$ is the first-order Raman scattering energy. Their values are 0.063 eV for Si and 0.035 eV for GaAs. Appendix G lists these results, together with other important properties.

1.6.2 Optical Properties

Optical measurement constitutes the most-important means of determining the band structures of semiconductors. Photon-induced electronic transitions can occur between different bands, which lead to the determination of the energy bandgap, or within a single band such as the free-carrier absorption. Optical measurements can also be used to study lattice vibrations (phonons). The optical properties of semiconductor are characterized by the complex refractive index,

$$\bar{n} = n_r - ik_e. \tag{124}$$

The real part of the refractive index n_r determines the propagation velocity (υ and wavelength λ) in the medium (assuming ambient is a vacuum having wavelength λ_0)

$$n_r = \frac{c}{\upsilon} = \frac{\lambda_0}{\lambda}. \tag{125}$$

The imaginary part k_e, called the extinction coefficient, determines the absorption coefficient

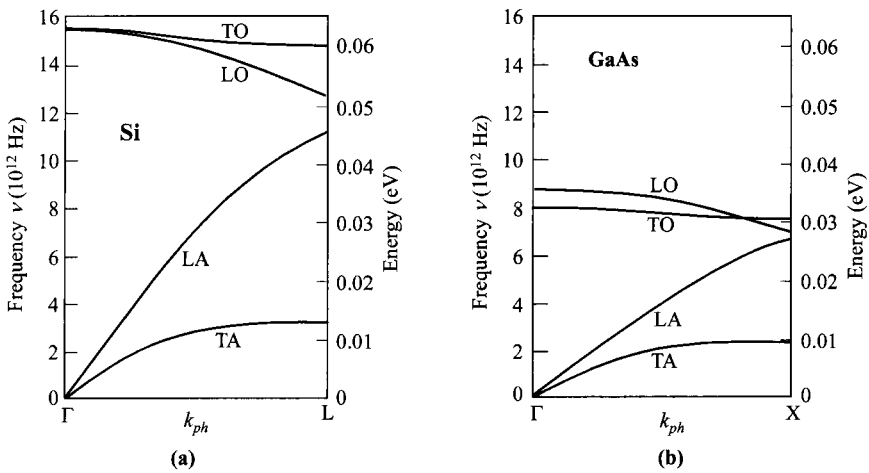


Fig. 28 Measured phonon spectra in (a) Si (After Ref. 73.) and (b) GaAs (After Ref. 74.). TO and LO stand for transverse and longitudinal optical modes, and TA and LA for transverse and longitudinal acoustic modes.

$$\alpha = \frac{4\pi k_e}{\lambda}. \tag{126}$$

In semiconductors, the absorption coefficient is a strong function of the wavelength or photon energy. Near the absorption edge, the absorption coefficient can be expressed as[5]

$$\alpha \propto (h\nu - E_g)^{\gamma} \tag{127}$$

where $h\nu$ is the photon energy and γ is a constant. There exist two types of band-to-band transitions: allowed and forbidden. (Forbidden transitions take into account the small but finite momentum of photons and are much less probable.) For direct-bandgap materials, transitions mostly occur between two bands of the same k value, as transitions (a) and (b) in Fig. 29. While allowed direct transitions can occur in all k values, forbidden direct transitions can only occur at $k \neq 0$. In the one-electron approximation, γ equals 1/2 and 3/2 for allowed and forbidden direct transitions, respectively. Note that for $k = 0$ at which the bandgap is defined, only allowed transition ($\gamma = 1/2$) occurs and thus it is used in determining the bandgap experimentally. For indirect transitions [transition (c) in Fig. 29], phonons are involved in order to conserve momentum. In these transitions, phonons (with energy E_p) are either absorbed or emitted, and the absorption coefficient is modified to

$$\alpha \propto (h\nu - E_g \pm E_p)^{\gamma}. \tag{128}$$

Here the constant γ equals 2 and 3 for allowed and forbidden indirect transitions, respectively.

In addition, increased absorption peaks and steps can be due to formation of excitons, which are bound electron-hole pairs with energy levels within the bands that move through the crystal lattice as a unit. Near the absorption edge, where the values of $(E_g - h\nu)$ become comparable with the binding energy of an exciton, the Coulomb interaction between the free electron and hole must be taken into account. The photon energy required for absorption is lowered by this binding energy. For $h\nu \gtrsim E_g$ the

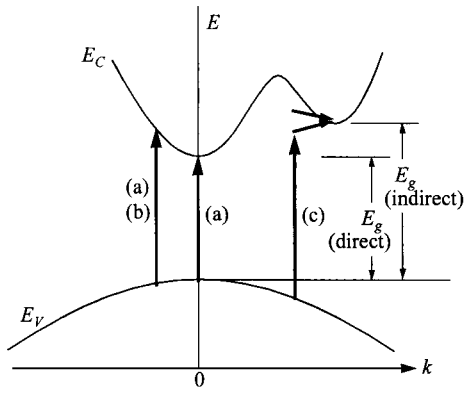


Fig. 29 Optical transitions: (a) allowed and (b) forbidden direct transitions; (c) indirect transition involving phonon emission (upper arrow) and phonon absorption (lower arrow).

absorption merges continuously into the fundamental absorption. When $h\nu \gg E_g$, higher energy bands participate in the transition processes, and complicated band structures are reflected in the absorption coefficient.

Figure 30 plots the experimental absorption coefficients α near and above the fundamental absorption edge (band-to-band transition) for Si and GaAs. The shift of the curves toward higher photon energies at lower temperature is associated with the temperature dependence of the bandgap (Fig. 6). An α of 10^4 cm^{-1} means that 63% of light will be absorbed in one micron of semiconductor.

When light passes through a semiconductor, absorption of light and generation of electron-hole pairs (G_e) occur, and the light intensity P_{op} diminishes with distance according to

$$\frac{dP_{op}(x)}{dx} = -\alpha P_{op}(x) = G_e h\nu. \tag{129}$$

Solution of the above gives an exponential decay of intensity

$$P_{op}(x) = P_0(1-R)\exp(-\alpha x) \tag{130}$$

where P_0 is the external incident light intensity and R is the reflection of the ambient-semiconductor interface at normal incidence,

$$R = \frac{(1-n_r)^2 + k_e^2}{(1+n_r)^2 + k_e^2}. \tag{131}$$

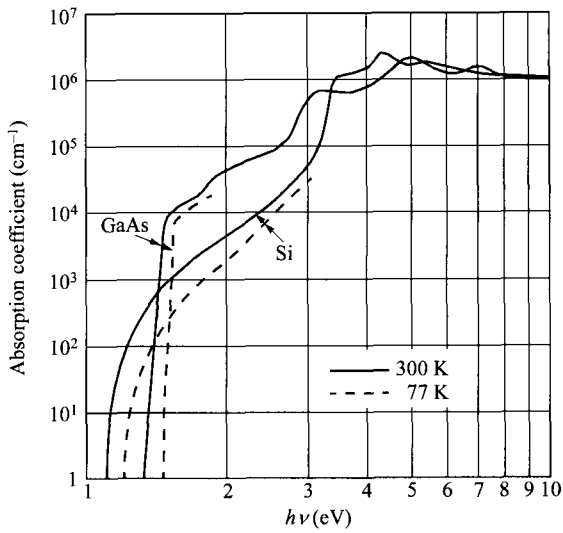


Fig. 30 Measured absorption coefficients near and above the fundamental absorption edge for Si and GaAs. (After Refs. 75–78.)

In a semiconductor sample of thickness W where the product αW is not large, multiple reflections will occur between the two interfaces. Summing up all the light components in the backward direction, the total reflection coefficient is calculated to be

$$R_\Sigma = R\left[1 + \frac{(1-R)^2 \exp(-2\alpha W)}{1 - R^2 \exp(-2\alpha W)}\right], \tag{132}$$

and the total transmission coefficient is given by

$$T_\Sigma = \frac{(1-R)^2 \exp(-\alpha W)}{1 - R^2 \exp(-2\alpha W)}. \tag{133}$$

The transmission coefficient T_Σ and the reflection coefficient R_Σ are two important quantities generally measured. By analyzing the T_Σ-λ or R_Σ-λ data at normal incidence, or by making observations on R_Σ or T_Σ for different angles of incidence, both n_r and k_e can be obtained and related to the transition energy between bands.

1.6.3 Thermal Properties

When a temperature gradient exists in a semiconductor in addition to an applied electric field, the total current density (in one dimension) is[5]

$$J = \sigma\left(\frac{1}{q}\frac{dE_F}{dx} - \mathscr{P}\frac{dT}{dx}\right) \tag{134}$$

where $\mathscr{P}$ is the thermoelectric power, so named to indicate that for an open-circuit condition the net current is zero and an electric field is generated by the temperature gradient. For a nondegenerate semiconductor with a mean free time between collisions $\tau_m \propto E^{-s}$ as discussed previously, the thermoelectric power is given by

$$\mathscr{P} = -\frac{k}{q}\left\{\frac{[\frac{5}{2} - s + \ln(N_C/n)]n\mu_n - [\frac{5}{2} - s - \ln(N_V/p)]p\mu_p}{n\mu_n + p\mu_p}\right\} \tag{135}$$

(k is Boltzmann constant). This equation indicates that the thermoelectric power is negative for n-type semiconductors and positive for p-type semiconductors, a fact often used to determine the conduction type of a semiconductor. The thermoelectric power can also be used to determine the resistivity and the position of the Fermi level relative to the band edges. At room temperature the thermoelectric power $\mathscr{P}$ of p-type silicon increases with resistivity: 1 mV/K for a 0.1 Ω-cm sample and 1.7 mV/K for a 100 Ω-cm sample. Similar results (except a change of the sign for $\mathscr{P}$) can be obtained for n-type silicon samples.

Another important thermal effect is thermal conduction. It is a diffusion type of process where the heat flow Q is driven by the temperature gradient

$$Q = -\kappa\frac{dT}{dx}. \tag{136}$$

The thermal conductivity κ has the major components of phonon (lattice) conduction κ_L and mixed free-carrier conduction κ_M of electrons and holes,

$$\kappa = \kappa_L + \kappa_M. \tag{137}$$

The lattice contribution is carried out by diffusion and scattering of phonons. These scattering events include many types, such as phonon-to-phonon, phonon-to-defects, phonon-to-carriers, boundaries and surfaces, and so on. The overall effect can be interpreted as

$$\kappa_L = \frac{1}{3} C_v v_{ph} \lambda_{ph} \tag{138}$$

where C_v is the specific heat, v_{ph} the phonon velocity, and λ_{ph} the phonon mean free path. The contribution due to mixed carriers, if $\tau_m \propto E^{-s}$ holds for both electron and hole scattering, is given by

$$\kappa_M = \frac{(\frac{5}{2} - s)k^2 \sigma T}{q^2} + \frac{k^2 \sigma T}{q^2} \frac{[5 - 2s + (E_g/kT)]^2 np\mu_n\mu_p}{(n\mu_n + p\mu_p)^2}. \tag{139}$$

Figure 31 shows the measured thermal conductivity as a function of lattice temperature for Si and GaAs. Appendix G lists the room-temperature values. The contributions of conduction carriers to the thermal conductivity are in general quite small,

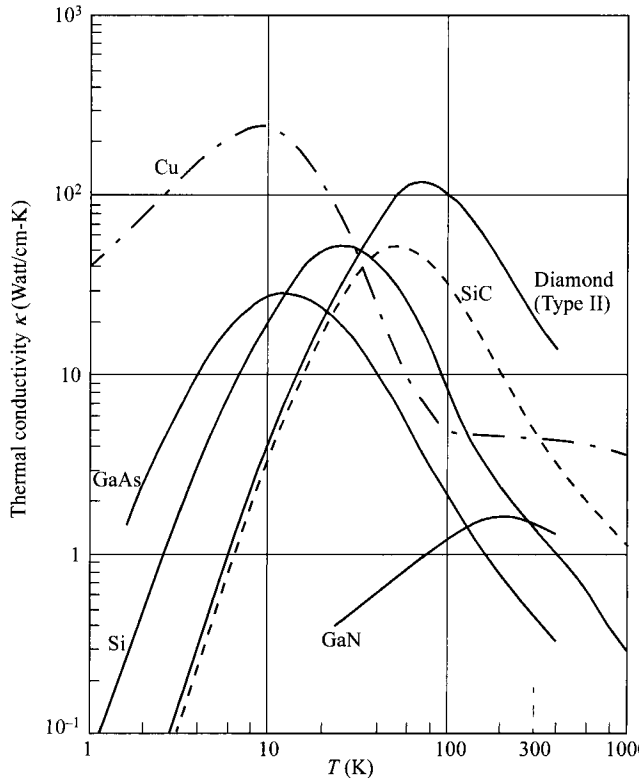


Fig. 31 Measured thermal conductivity versus temperature for pure Si, GaAs, SiC, GaN, Cu, and diamond (Type II). (After Refs. 79–83.)

so the general temperature dependence follows that of κ_L and has a inverted V-shape. At low temperatures, the specific heat has a T^3 dependence and κ goes up sharply. At high temperatures, phonon-phonon scattering dominates and λ_{ph} (and κ_L) drops according to $1/T$. Figure 31 also shows the thermal conductivities for Cu, diamond, SiC, and GaN. Copper is the most commonly used metal for thermal conduction in *p-n* junction devices; diamond has the highest room-temperature thermal conductivity known to date and is useful as the thermal sink for semiconductor lasers and IMPATT oscillators. SiC and GaN are important semiconductors for power devices.

1.7 HETEROJUNCTIONS AND NANOSTRUCTURES

A heterojunction is a junction formed between two dissimilar semiconductors. For semiconductor-device applications, the difference in energy gap provides another degree of freedom that produces many interesting phenomena. The successful applications of heterojunctions in various devices is due to the capability of epitaxy technology to grow lattice-matched semiconductor materials on top of one another with virtually no interface traps. Heterojunctions have been widely used in various device applications. The underlying physics of epitaxial heterojunction is matching of the lattice constants. This is a physical requirement in atom placement. Severe lattice mismatch will cause dislocations at the interface and results in electrical defects such as interface traps. The lattice constants of some common semiconductors are shown in Fig. 32, together with their energy gaps. A good combination for heterojunction devices is two materials of similar lattice constants but different E_g. As can be seen, GaAs/AlGaAs (or /AlAs) is a good example.

It turns out that if the lattice constants are not severely mismatched, good-quality heteroepitaxy can still be grown, provided that the epitaxial-layer thickness is small enough. The amount of lattice mismatch and the maximum allowed epitaxial layer are directly related. This can be explained with the help of Fig. 33. For a relaxed, thick heteroepitaxial layer, dislocations at the interface are inevitable due to the phys-

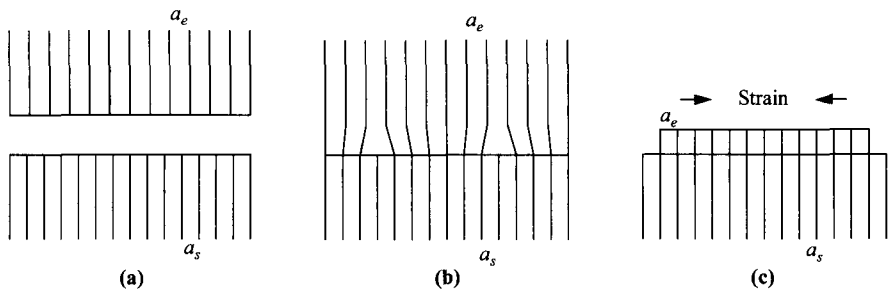


Fig. 33 Two materials with slightly mismatched lattice constants a_s and a_e. (a) In isolation. (b) Heteroepitaxy with thick, relaxed epitaxial layer having dislocations at the interface. (c) With thin, strained epitaxial layer without dislocations. Epitaxial lattice constant a_e is strained to follow that of the substrate a_s.

ical mismatch of terminating bonds at the interface. However, if the heteroepitaxial layer is thin enough, the layer can be physically strained to the degree that its lattice constant becomes the same as the substrate (Fig. 33c). When that happens, dislocations can be eliminated.

To estimate the critical thickness of this strained layer, we visualize the heteroepitaxial process from the beginning. At the start, the epitaxial layer follows the lattice of the substrate, but the strain energy builds up as the film becomes thicker. Eventually the film has built up too much strain to sustain and it transforms to a relaxed state, i.e. going from Figs. 33c to 33b. The lattice mismatch is defined as

$$\Delta \equiv \frac{|a_e - a_s|}{a_e}, \tag{140}$$

where a_e and a_s are the lattice constants of the epitaxial layer and substrate respectively. The critical thickness has been found to follow an empirical formula given by

$$t_c \approx \frac{a_e}{2\Delta} \approx \frac{a_e^2}{2|a_e - a_s|}. \tag{141}$$

A typical number for the critical thickness, from a mismatch of 2% and an a_e of 5 Å, is about 10 nm. This technique of growing strained heteroepitaxy has bough an extra degree of freedom and permits the use of a wider range of materials. It has had great impacts on expanding the applications of heterostructures, for making novel devices as well as improving their performances.

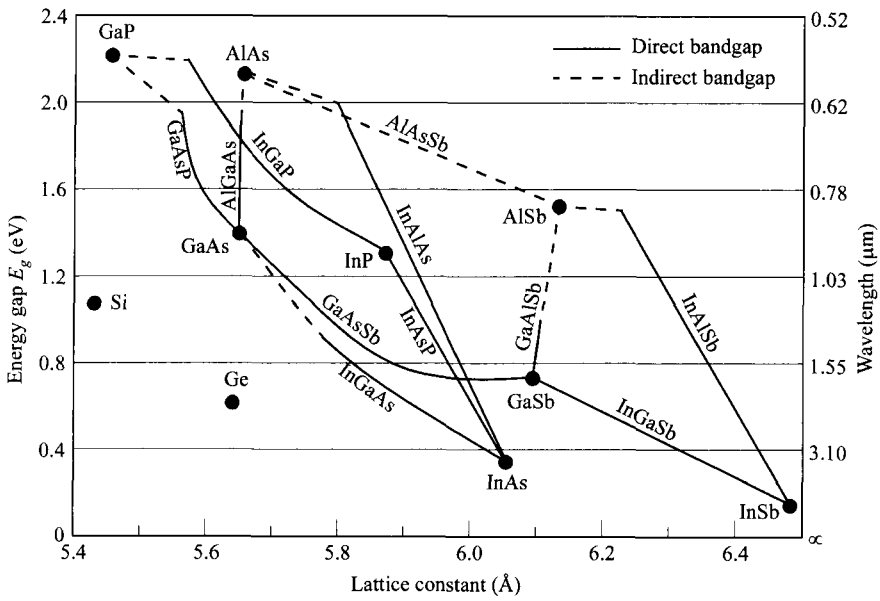


Fig. 32 Energy gap vs. lattice constant for some common elementary and binary semiconductors.

In addition to having different energy gaps, the electron affinities of these semiconductors are also different and need to be considered in device applications. This leads to different combinations of E_C and E_V alignment at the interface. According to their band alignment, heterojunctions can be classified into three groups as shown in Fig. 34: (1) Type-I or *straddling* heterojunction, (2) Type-II or *staggered* heterojunction, and (3) Type-III or *broken-gap* heterojunction. In a Type-I (straddling) heterojunction, one material has both lower E_C and higher E_V, and naturally it must have a smaller energy gap. In a Type-II (staggered) heterojunction, the locations of lower E_C and higher E_V are displaced, so electrons being collected at lower E_C and holes being collected at higher E_V are confined in different spaces. A Type-III (broken-gap) heterojunction is a special case of Type-II, but the E_C of one side is lower than the E_V of the other. The conduction band thus overlaps the valence band at the interface, hence the name *broken gap*.

Quantum Well and Superlattice. One important application of heterojunction is to use ΔE_C and ΔE_V to form barriers for carriers. A *quantum well* is formed by two heterojunctions or three layers of materials such that the middle layer has the lowest E_C for an electron well or the highest E_V for a hole well. A quantum well thus confines electrons or holes in a two-dimensional (2-D) system. When electrons are free to move in a bulk semiconductor in all directions (3-D), their energy above the conduction-band edge is continuous, given by the relationship to their momentum (Eq. 8):

$$E - E_C = \frac{\hbar^2}{2m_e^*}(k_x^2 + k_y^2 + k_z^2). \tag{142}$$

In a quantum well, carriers are confined in one direction, say in the x-coordinate such that $k_x = 0$. It will be shown that the energy within this well is no longer continuous with respect to the x-direction, but becomes quantized in subbands.

The most-important parameters for a quantum well are the well width L_x and well height ϕ_b. The energy-band diagram in Fig. 35a shows that the potential barrier is obtained from the conduction-band and valence-band offsets (ΔE_C and ΔE_V). The solution for the wavefunction of the Schrödinger equation inside the well is

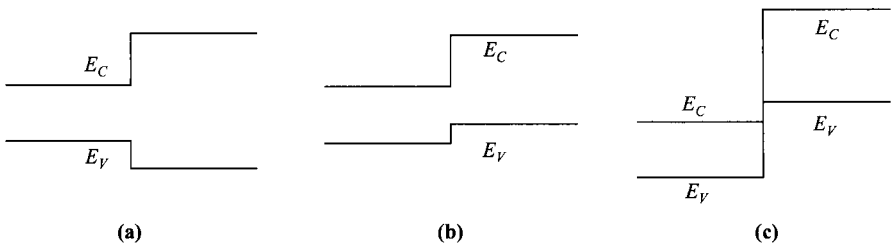


Fig. 34 Classification of heterojunctions. (a) Type-I or straddling heterojunction. (b) Type-II or staggered heterojunction. (c) Type-III or broken-gap heterojunction.

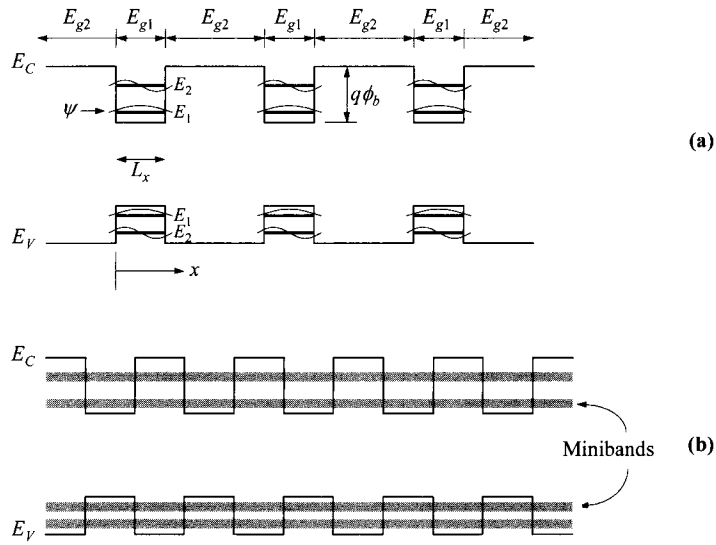


Fig. 35 Energy-band diagrams for (a) heterostructure (composition) multiple quantum wells and (b) heterostructure superlattice.

$$\psi(x) = \sin\left(\frac{i\pi x}{L_x}\right) \tag{143}$$

where i is an integer. It should be noted that at the well boundaries, ψ is truly zero only when ϕ_b is infinite. With finite ϕ_b, carriers can "leak" out (by tunneling) of the well with finite probability. This is important for the formation of a superlattice, discussed later. The pinning of nodes at the well boundaries leads to the quantization of subbands, each has a bottom energy of (with respect to the band edges)

$$E_i = \frac{\hbar^2 \pi^2 i^2}{2m^* L_x^2} \quad . \tag{144}$$

These solutions do not take into account a finite barrier height. With L_x as a variable, a quantum well can only be loosely defined. The minimum requirements should be that the quantized energy $\hbar^2\pi^2/2m^*L_x^2$ is much larger than kT, and L_x is smaller than the mean free path and the de Broglie wavelength. (Notice that the de Broglie wavelength $\lambda = h/(2m^*E)^{1/2}$ has a form similar to L_x of Eq. 144.) Also, it is interesting to note that since the continuous conduction band is now divided into subbands, carriers no longer reside on the band edges E_C or E_V but on these subbands only. In effect, the effective energy gap for interband transitions inside the quantum well becomes larger than the bulk E_g.

When quantum wells are separated from one another by thick barrier layers, there is no communication between them and this system can only be described as multiple quantum wells. However, when the barrier layers between them become thinner, to

the extent that wavefunctions start to overlap, a heterostructure (composition) *superlattice* is formed. The superlattice has two major differences from a multiple-quantum-well system: (1) the energy levels are continuous in space across the barrier, and (2) the discrete bands widen into minibands (Fig. 35b). The transition from multiple quantum wells into a superlattice is analogous to the formation of a regular lattice by pulling atoms together. The isolated atoms have discrete levels, whereas a lattice transforms these discrete levels into the continuous conduction band and valance band.

Another approach to form quantum wells and superlattices is by spatial variation in doping,[84] where the potential barriers are formed by space-charge fields (Fig. 36a). The barrier shape in this case is parabolic rather than rectangular. There are two interesting features in this doping (or *n-i-p-i*) multiple-quantum-well structure. First, the conduction-band minimum and the valence-band maximum are displaced from each other, meaning that electrons and holes accumulate at different locations. This leads to minimal electron-hole recombination and very long carrier lifetime, many orders of magnitude higher than that of the regular material. This is similar to a Type-II heterojunction. Second, the effective energy gap, which is now between the first quantized levels for the electrons and holes, is reduced from the fundamental material. This tunable effective energy gap enables light emission and absorption of longer wavelengths. This structure is unique in that it has an indirect energy gap in "real space", as opposed to *k*-space. When the doping quantum wells are close together, a doping (*n-i-p-i*) superlattice is again formed (Fig. 36b).

Quantum Wire and Quantum Dot. The physical dimensions of a semiconductor have significant implications on the electronic properties, as these dimensions are

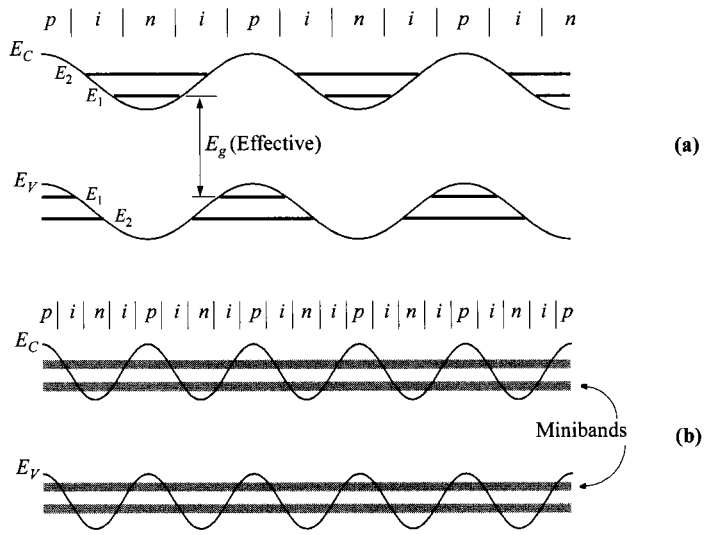


Fig. 36 Energy-band diagrams for (a) doping (*n-i-p-i*) multiple quantum wells, and (b) doping superlattice.

reduced to the order of the de Broglie wavelength. The confinement of carriers can be further extended to one- and zero-dimension, resulting in what are known as *quantum wire* and *quantum dot*. One of the major effects is on the density of states $N(E)$. Depending on the degree of confinement, $N(E)$ has very different shapes as a function of energy. The qualitative shapes of $N(E)$ for bulk semiconductor, quantum well, quantum wire, and quantum dot are shown in Fig. 37. For a 3-D system, the density of states has been given earlier (Eq. 14) and is repeated here

$$N(E) = \frac{m^* \sqrt{2m^* E}}{\pi^2 \hbar^3}. \tag{145}$$

The density of states in a 2-D system (quantum well) has a step function of

$$N(E) = \frac{m^* i}{\pi \hbar^2 L_x}. \tag{146}$$

The density of states in a 1-D system (quantum wire) has an inverse energy relationship of

$$N(E) = \frac{\sqrt{2m^*}}{\pi \hbar L_x L_y} \sum_{i,j} (E - E_{i,j})^{-1/2}, \tag{147}$$

where

$$E_{i,j} = \frac{\hbar^2 \pi^2}{2m^*} \left(\frac{i^2}{L_x^2} + \frac{j^2}{L_y^2} \right). \tag{148}$$

The density of states in a 0-D system (quantum dot) is continuous and independent of energy,

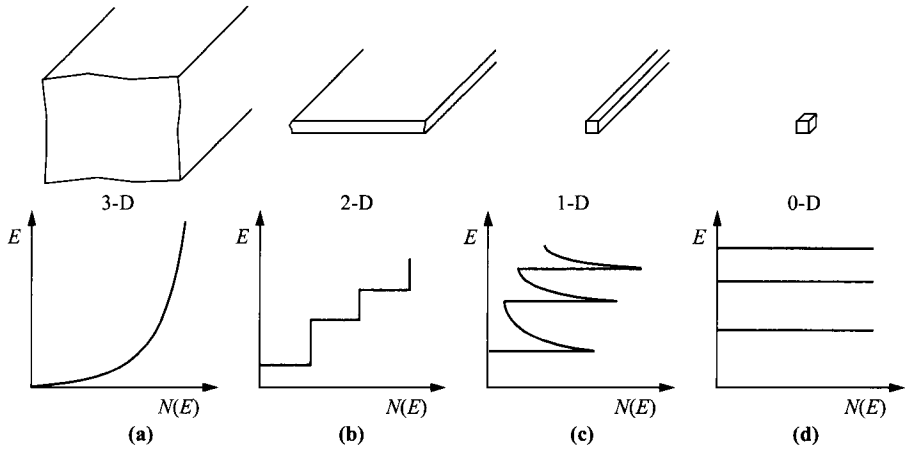


Fig. 37 Density of states $N(E)$ for (a) bulk semiconductor (3-D), (b) quantum well (2-D), (c) quantum wire (1-D), and (d) quantum dot (0-D).

$$N(E) = \frac{2}{L_x L_y L_z} \sum_{i,j,k} \delta(E - E_{i,j,k}), \tag{149}$$

where

$$E_{i,j,k} = \frac{\hbar^2 \pi^2}{2m^*}\left(\frac{i^2}{L_x^2} + \frac{j^2}{L_y^2} + \frac{k^2}{L_z^2}\right). \tag{150}$$

Since the carrier concentration and its distribution in energy is given by the density of states multiplied by the Fermi-Dirac distribution, these density-of-state functions are important for the device operation as their physical dimensions are scaled to near the de Broglie wavelength (≈ 20 nm).

1.8 BASIC EQUATIONS AND EXAMPLES

1.8.1 Basic Equations

The basic equations for semiconductor-device operation describe the static and dynamic behavior of carriers in semiconductors under external influences, such as applied field or optical excitation, that cause deviation from the thermal-equilibrium conditions.[36] The basic equations can be classified in three groups; electrostatic equations, current-density equations, and continuity equations.

Electrostatic Equations. There are two important equations relating charge to electric field (= $\mathscr{D}/\varepsilon_s$ where $\mathscr{D}$ is electric displacement). The first is from one of Maxwell equations,

$$\nabla \cdot \mathscr{D} = \rho(x, y, z), \tag{151}$$

also known as Gauss' law or Poisson equation. For a one-dimensional problem, this reduces to a more useful form of

$$\frac{d^2 \psi_i}{dx^2} = -\frac{d\mathscr{E}}{dx} = -\frac{\rho}{\varepsilon_s} = \frac{q(n - p + N_A - N_D)}{\varepsilon_s} \tag{152}$$

($\psi_i \equiv -E_i/q$). This is commonly used, for example, to determine the potential and field distribution caused by a charge density ρ within the depletion layer. The second equation deals with charge density along an interface, instead of bulk charge. The boundary conditions across an interface of charge sheet Q is given by

$$\mathscr{E}_1(0^-)\varepsilon_1 = \mathscr{E}_2(0^+)\varepsilon_2 - Q. \tag{153}$$

Current-Density Equations. The most-common current conduction consists of the drift component, caused by the electric field, and the diffusion component, caused by the carrier-concentration gradient. The current-density equations are:

$$\boldsymbol{J}_n = q\mu_n n\mathscr{E} + qD_n \nabla n, \tag{154a}$$

$$\boldsymbol{J}_p = q\mu_p p\mathscr{E} - qD_p \nabla p, \tag{154b}$$

$$\boldsymbol{J}_{\text{cond}} = \boldsymbol{J}_n + \boldsymbol{J}_p, \tag{155}$$

where $\boldsymbol{J}_n$ and $\boldsymbol{J}_p$ are the electron and hole current densities, respectively. The values of the electron and hole mobilities (μ_n and μ_p) have been given in Section 1.5.1. For nondegenerate semiconductors the carrier diffusion constants (D_n and D_p) and the mobilities are given by the Einstein relation [$D_n = (kT/q)\mu_n$, etc.].

For a one-dimensional case, Eqs. 154a and 154b reduce to

$$J_n = q\mu_n n\mathscr{E} + qD_n\frac{dn}{dx} = q\mu_n\left(n\mathscr{E} + \frac{kT}{q}\frac{dn}{dx}\right) = \mu_n n\frac{dE_{Fn}}{dx}, \tag{156a}$$

$$J_p = q\mu_p p\mathscr{E} - qD_p\frac{dp}{dx} = q\mu_p\left(p\mathscr{E} - \frac{kT}{q}\frac{dp}{dx}\right) = \mu_p p\frac{dE_{Fp}}{dx} \tag{156b}$$

where E_{Fn} and E_{Fp} are quasi Fermi levels for electrons and holes, respectively. These equations are valid for low electric fields. At sufficiently high fields the term $\mu_n\mathscr{E}$ or $\mu_p\mathscr{E}$ should be replaced by the saturation velocity v_s (and the last equalities about E_{Fn} and E_{Fp} do not hold any more). These equations do not include the effect from an externally applied magnetic field where the magneto-resistive effect reduces the current.

Continuity Equations. While the above current-density equations are for steady-state conditions, the continuity equations deal with time-dependent phenomena such as low-level injection, generation and recombination. Qualitatively, the net change of carrier concentration is the difference between generation and recombination, plus the net current flowing in and out of the region of interest. The continuity equations are:

$$\frac{\partial n}{\partial t} = G_n - U_n + \frac{1}{q}\nabla\cdot\boldsymbol{J}_n, \tag{157a}$$

$$\frac{\partial p}{\partial t} = G_p - U_p - \frac{1}{q}\nabla\cdot\boldsymbol{J}_p \tag{157b}$$

where G_n and G_p are the electron and hole generation rate (cm^{-3}-s^{-1}), respectively, caused by external influences such as the optical excitation with photons or impact ionization under large electric fields. The recombination rates, $U_n = \Delta n/\tau_n$ and $U_p = \Delta p/\tau_p$, have been discussed in Section 1.5.4.

For the one-dimensional case under a low-injection condition, Eqs. 157a and 157b reduce to

$$\frac{\partial n_p}{\partial t} = G_n - \frac{n_p - n_{po}}{\tau_n} + n_p\mu_n\frac{\partial\mathscr{E}}{\partial x} + \mu_n\mathscr{E}\frac{\partial n_p}{\partial x} + D_n\frac{\partial^2 n_p}{\partial x^2} \tag{158a}$$

$$\frac{\partial p_n}{\partial t} = G_p - \frac{p_n - p_{no}}{\tau_p} - p_n\mu_p\frac{\partial\mathscr{E}}{\partial x} - \mu_p\mathscr{E}\frac{\partial p_n}{\partial x} + D_p\frac{\partial^2 p_n}{\partial x^2}. \tag{158b}$$

1.8.2 Examples

In this section, we demonstrate the use of the continuity equations for studying the time dependence and space dependence of excess carriers. Excess carriers can be

created by optical excitation or injection from a nearby junction. In these examples we use optical excitation for simplicity.

Decay of Excess Carriers with Time. Consider an n-type sample, as shown in Fig. 38a, that is illuminated with light in which the electron-hole pairs are generated uniformly throughout the sample with a uniform generation rate G_p. In this example the sample thickness is much smaller than $1/\alpha$, and the space dependence is absent here. The boundary conditions are $\mathscr{E} = \partial\mathscr{E}/\partial x = 0$ and $\partial p_n/\partial x = 0$. We have from Eq. 158b:

$$\frac{dp_n}{dt} = G_p - \frac{p_n - p_{no}}{\tau_p}. \tag{159}$$

At steady state, $\partial p_n/\partial t = 0$ and

$$p_n - p_{no} = \tau_p G_p = \text{constant}. \tag{160}$$

If at an arbitrary time, say $t = 0$, the light is suddenly turned off, the differential equation is now

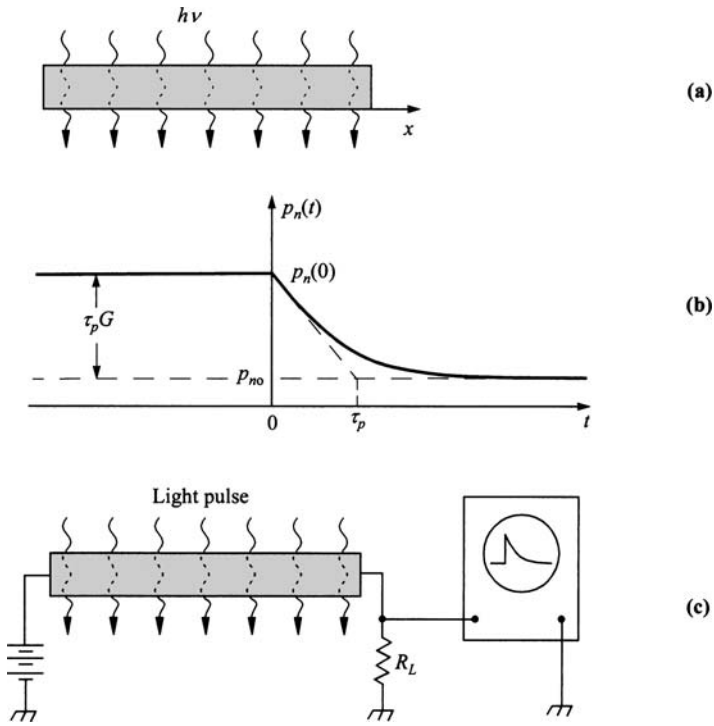


Fig. 38 Decay of photo-excited carriers. (a) n-type sample under constant illumination. (b) Decay of minority carriers (holes) with time. (c) Schematic experimental setup to measure minority carrier lifetime. (After Ref. 71.)

$$\frac{dp_n}{dt} = -\frac{p_n - p_{no}}{\tau_p}. \tag{161}$$

With the boundary conditions $p_n(t = 0) = p_{no} + \tau_p G_p$, as given in Eq. 160, and $p_n(\infty) = p_{no}$, the solution is

$$p_n(t) = p_{no} + \tau_p G_p \exp\left(-\frac{t}{\tau_p}\right). \tag{162}$$

Figure 38b shows the variation of p_n with time.

The example above presents the main idea of the Stevenson-Keyes method for measuring minority-carrier lifetime.[71] Figure 38c shows a schematic setup. The excess carriers generated uniformly throughout the sample by the light pulses cause a momentary increase in the conductivity and current. During the periods when the light pulses are off, the decay of this photoconductivity can be observed on an oscilloscope which monitors the voltage drop across a resistor load R_L, and is a measure of the lifetime.

Decay of Excess Carriers with Distance. Figure 39a shows another simple example where excess carriers are injected from one side (e.g., by high-energy photons that create electron-hole pairs at the surface only). Referring to Fig. 30, note that for $h\nu = 3.5$ eV, the absorption coefficient is about 10^6 cm^{-1}, in other words, the light intensity decreases by a factor of e in a distance of 10 nm.

At steady state there is a concentration gradient near the surface. The differential equation for an n-type sample without bias is, from Eq. 158b,

$$\frac{\partial p_n}{\partial t} = 0 = -\frac{p_n - p_{no}}{\tau_p} + D_p \frac{\partial^2 p_n}{\partial x^2}. \tag{163}$$

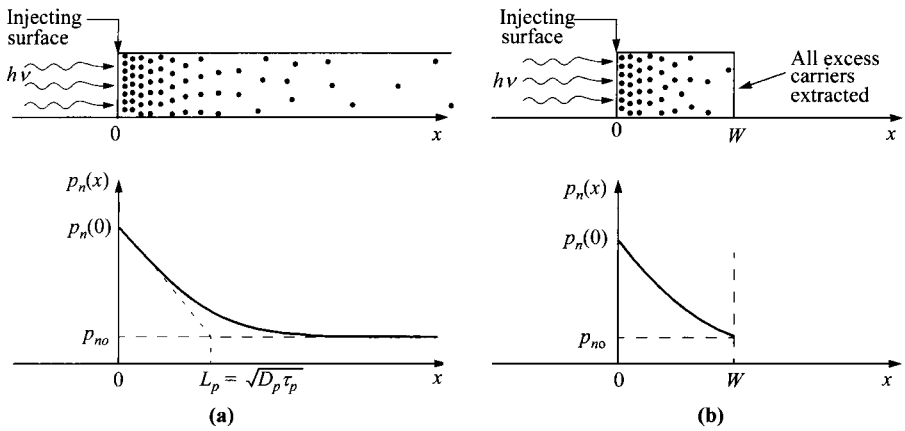


Fig. 39 Steady-state carrier injection from one side. (a) Semiinfinite sample. (b) Sample with length W.

The boundary conditions are $p_n(x = 0)$ = constant, depending on the injection level, and $p_n(\infty) = p_{no}$. The solution of $p_n(x)$ is

$$p_n(x) = p_{no} + [p_n(0) - p_{no}]\exp\left(-\frac{x}{L_p}\right) \tag{164}$$

where the diffusion length is $L_p = (D_p\tau_p)^{1/2}$ (Fig. 39a). The maximum values of L_p and L_n are of the order of 1 cm in silicon, but only of the order of 10^{-2} cm in gallium arsenide.

Of special interest is the case where the second boundary condition is changed so that all excess carriers at the back surface ($x = W$) are extracted or $p_n(W) = p_{no}$, then we obtain from Eq. 163 a new solution,

$$p_n(x) = p_{no} + [p_n(0) - p_{no}]\left\{\frac{\sinh[(W-x)/L_p]}{\sinh(W/L_p)}\right\}. \tag{165}$$

This result is shown in Fig. 39b. The current density at $x = W$ is given by Eq. 156b:

$$J_P = -qD_p\frac{dp}{dx}\bigg|_W = \frac{qD_p[p_n(0) - p_{no}]}{L_p\sinh(W/L_p)}. \tag{166}$$

It will be shown later that Eq. 166 is related to the current gain in bipolar transistors (Chapter 5).

Decay of Excess Carriers with Time and Distance. When localized light pulses generate excess carriers in a semiconductor (Fig. 40a), the transport equation after the pulse without bias is given by Eq. 158b by setting $G_p = \mathscr{E} = \partial\mathscr{E}/\partial x = 0$:

$$\frac{\partial p_n}{\partial t} = -\frac{p_n - p_{no}}{\tau_p} + D_p\frac{\partial^2 p_n}{\partial x^2}. \tag{167}$$

The solution is given by

$$p_n(x, t) = \frac{N'}{\sqrt{4\pi D_p t}}\exp\left(-\frac{x^2}{4D_p t} - \frac{t}{\tau_p}\right) + p_{no} \tag{168}$$

where N' is the number of electrons or holes generated initially per unit area. Figure 40b shows this solution as the carriers diffuse away from the point of injection, and they also recombine (area under curve is decreased).

If an electric field is applied along the sample, the solution is in the same form but with x replaced by $(x - \mu_p\mathscr{E}t)$ (Fig. 40c); thus the whole *package* of excess carrier moves toward the negative end of the sample with a drift velocity $\mu_p\mathscr{E}$. At the same time, the carriers diffuse outward and recombine as in the field-free case.

The example above is similar to the celebrated Haynes-Shockley experiment for the measurement of carrier drift mobility in semiconductors.[85] With known sample length, applied field, and the time delay between the applied signals (bias on and light off) and the detected signal at the sample end (both displayed on the oscilloscope), the drift mobility $\mu = x/\mathscr{E}t$ can be calculated.

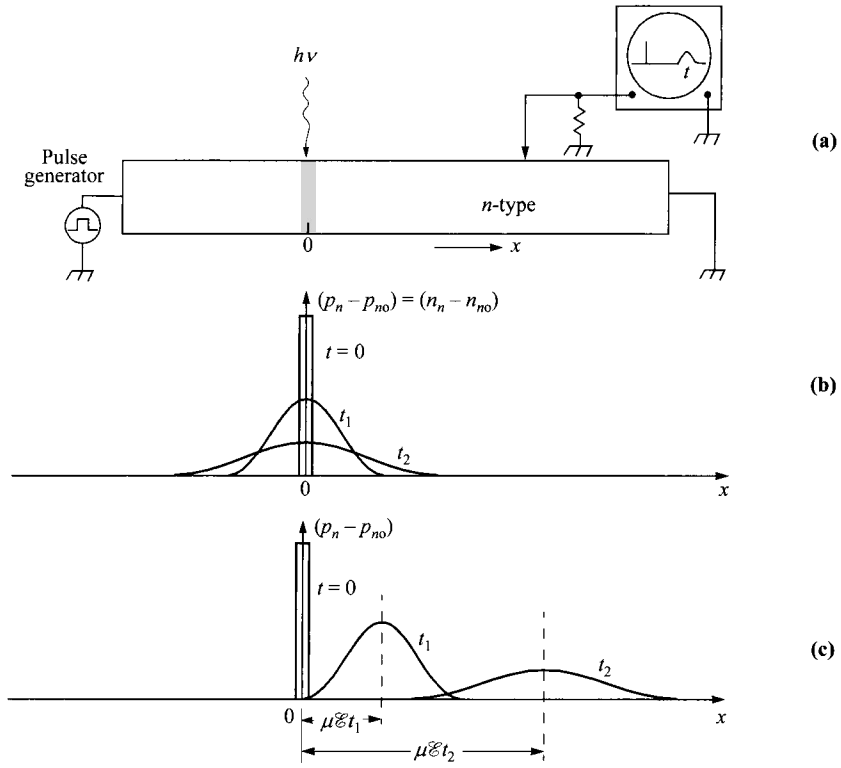


Fig. 40 Transient and steady-state carrier diffusion after a localized light pulse. (a) Experimental setup. (b) Without applied field. (c) With applied field.

Surface Recombination. When surface recombination is introduced at one end of a semiconductor sample (Fig. 41), the boundary condition at $x = 0$ is governed by

$$qD_p \left.\frac{dp_n}{dx}\right|_{x=0} = qS_p[p_n(0) - p_{no}] \tag{169}$$

which states that the minority carriers that reach the surface recombine there. The constant S_p with units cm/s is defined as the surface recombination velocity for holes. The boundary condition at $x = \infty$ is given by Eq. 160. The differential equation, without bias and at steady state, is

$$0 = G_p - \frac{p_n - p_{no}}{\tau_p} + D_p \frac{d^2 p_n}{dx^2}. \tag{170}$$

The solution of the equation subject to the boundary conditions above is

$$p_n(x) = p_{no} + \tau_p G_p \left[1 - \frac{\tau_p S_p \exp(-x/L_p)}{L_p + \tau_p S_p}\right] \tag{171}$$

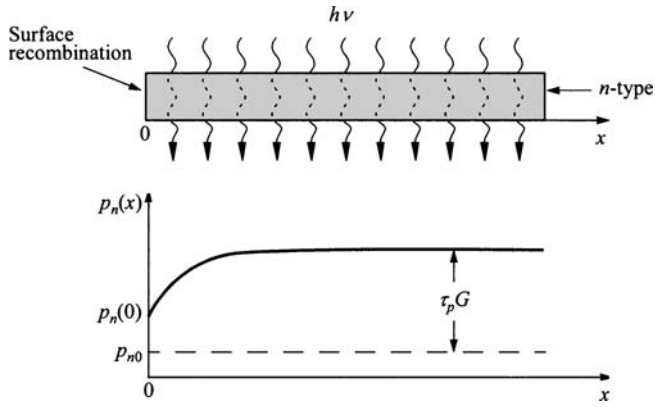


Fig. 41 Surface recombination at $x = 0$. The minority-carrier distribution near the surface is affected by the surface recombination velocity.

which is plotted in Fig. 41 for a finite S_p. When $S_p \to 0$, then $p_n(x) \to p_{no} + \tau_p G_p$, which was obtained previously (Eq. 160). When $S_p \to \infty$, then $p_n(x) \to p_{no} + \tau_p G_p[1 - \exp(-x/L_p)]$, and the minority carrier density at the surface approaches its thermal equilibrium value p_{no}. Analogous to the low-injection bulk-recombination process, in which the reciprocal of the minority-carrier lifetime $(1/\tau)$ is equal to $\sigma_p \upsilon_{th} N_t$ (Eq. 95a), the surface recombination velocity is given by

$$S_p = \sigma_p \upsilon_{th} N'_{st} \tag{172}$$

where N'_{st} is the number of surface trapping centers per unit area at the boundary region.

REFERENCES

1. W. C. Dunlap, *An Introduction to Semiconductors*, Wiley, New York, 1957.
2. O. Madelung, *Physics of III-V Compounds*, Wiley, New York, 1964.
3. J. L. Moll, *Physics of Semiconductors*, McGraw-Hill, New York, 1964.
4. T. S. Moss, Ed., *Handbook on Semiconductors*, Vols. 1–4, North-Holland, Amsterdam, 1980.
5. R. A. Smith, *Semiconductors*, 2nd Ed., Cambridge University Press, London, 1979.
6. K. W. Böer, *Survey of Semiconductor Physics*, Van Nostrand Reinhold, New York, 1990.
7. K. Seeger, *Semiconductor Physics*, 7th Ed., Springer-Verlag, Berlin, 1999.
8. S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, New Jersey, 1989.
9. R. K. Willardson and A. C. Beer, Eds., *Semiconductors and Semimetals*, Vol. 2, *Physics of III-V Compounds*, Academic, New York, 1966.

10. W. B. Pearson, *Handbook of Lattice Spacings and Structure of Metals and Alloys*, Pergamon, New York, 1967.
11. H. C. Casey, Jr. and M. B. Panish, *Heterostructure Lasers*, Academic, New York, 1978.
12. See, for example, C. Kittel, *Introduction to Solid State Physics*, 7th Ed., Wiley, New York, 1996.
13. L. Brillouin, *Wave Propagation in Periodic Structures*, 2nd Ed., Dover, New York, 1963.
14. J. M. Ziman, *Principles of the Theory of Solids*, Cambridge University Press, London, 1964.
15. M. L. Cohen, "Pseudopotential Calculations for II-VI Compounds," in D. G. Thomas, Ed., *II-VI Semiconducting Compounds*, W. A. Benjamin, New York, 1967, p. 462.
16. C. Kittel, *Quantum Theory of Solids*, Wiley, New York, 1963.
17. L. C. Allen, "Interpolation Scheme for Energy Bands in Solids," *Phys. Rev.*, **98**, 993 (1955).
18. F. Herman, "The Electronic Energy Band Structure of Silicon and Germanium," *Proc. IRE*, **43**, 1703 (1955).
19. J. C. Phillips, "Energy-Band Interpolation Scheme Based on a Pseudopotential," *Phys. Rev.*, **112**, 685 (1958).
20. M. L. Cohen and J. R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors*, 2nd Ed., Springer-Verlag, Berlin, 1988.
21. J. M. Ziman, *Electrons and Phonons*, Clarendon, Oxford, 1960.
22. C. D. Thurmond, "The Standard Thermodynamic Function of the Formation of Electrons and Holes in Ge, Si, GaAs and GaP," *J. Electrochem. Soc.*, **122**, 1133 (1975).
23. V. Alex, S. Finkbeiner, and J. Weber, "Temperature Dependence of the Indirect Energy Gap in Crystalline Silicon," *J. Appl. Phys.*, **79**, 6943 (1996).
24. W. Paul and D. M. Warschauer, Eds., *Solids under Pressure*, McGraw-Hill, New York, 1963.
25. R. S. Ohl, "Light-Sensitive Electric Device," U.S. Patent 2,402,662. Filed May 27, 1941. Granted June 25, 1946.
26. M. Riordan and L. Hoddeson, "The Origins of the *pn* Junction", *IEEE Spectrum*, **34**-6, 46 (1997).
27. J. S. Blackmore, "Carrier Concentrations and Fermi Levels in Semiconductors," *Electron. Commun.*, **29**, 131 (1952).
28. W. B. Joyce and R. W. Dixon, "Analytic Approximations for the Fermi Energy of an Ideal Fermi Gas," *Appl. Phys. Lett.*, **31**, 354 (1977).
29. O. Madelung, Ed., *Semiconductors–Basic Data*, 2nd Ed., Springer-Verlag, Berlin, 1996.
30. R. N. Hall and J. H. Racette, "Diffusion and Solubility of Copper in Extrinsic and Intrinsic Germanium, Silicon, and Gallium Arsenide," *J. Appl. Phys.*, **35**, 379 (1964).
31. A. G. Milnes, *Deep Impurities in Semiconductors*, Wiley, New York, 1973.
32. J. Hermanson and J. C. Phillips, "Pseudopotential Theory of Exciton and Impurity States," *Phys. Rev.*, **150**, 652 (1966).
33. J. Callaway and A. J. Hughes, "Localized Defects in Semiconductors," *Phys. Rev.*, **156**, 860 (1967).
34. E. M. Conwell, "Properties of Silicon and Germanium, Part II," *Proc. IRE*, **46**, 1281 (1958).

35. S. M. Sze and J. C. Irvin, "Resistivity, Mobility, and Impurity Levels in GaAs, Ge, and Si at 300 K," *Solid-State Electron.*, **11**, 599 (1968).
36. W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, Princeton, New Jersey, 1950.
37. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
38. J. Bardeen and W. Shockley, "Deformation Potentials and Mobilities in Nonpolar Crystals," *Phys. Rev.*, **80**, 72 (1950).
39. E. Conwell and V. F. Weisskopf, "Theory of Impurity Scattering in Semiconductors," *Phys. Rev.*, **77**, 388 (1950).
40. C. Bulucea, "Recalculation of Irvin's Resistivity Curves for Diffused Layers in Silicon Using Updated Bulk Resistivity Data," *Solid-State Electron.*, **36**, 489 (1993).
41. C. Jacoboni, C. Canali, G. Ottaviani, and A. A. Quaranta, "A Review of Some Charge Transport Properties of Silicon," *Solid-State Electron.*, **20**, 77 (1977).
42. W. E. Beadle, J. C. C. Tsai, and R. D. Plummer, Eds., *Quick Reference Manual for Silicon Integrated Circuit Technology*, Wiley, New York, 1985.
43. F. M. Smits, "Measurement of Sheet Resistivities with the Four-Point Probe," *Bell Syst. Tech. J.*, **37**, 711 (1958).
44. E. H. Hall, "On a New Action of the Magnet on Electric Currents," *Am. J. Math.*, **2**, 287 (1879).
45. L. J. Van der Pauw, "A Method of Measuring Specific Resistivity and Hall Effect of Disc or Arbitrary Shape," *Philips Res. Rep.*, **13**, 1 (Feb. 1958).
46. D. M. Caughey and R. E. Thomas, "Carrier Mobilities in Silicon Empirically Related to Doping and Field," *Proc. IEEE*, **55**, 2192 (1967).
47. D. E. Aspnes, "GaAs Lower Conduction-Band Minima: Ordering and Properties," *Phys. Rev.*, **B14**, 5331 (1976).
48. P. Smith, M. Inoue, and J. Frey, "Electron Velocity in Si and GaAs at Very High Electric Fields," *Appl. Phys. Lett.*, **37**, 797 (1980).
49. J. G. Ruch and G. S. Kino, "Measurement of the Velocity-Field Characteristics of Gallium Arsenide," *Appl. Phys. Lett.*, **10**, 40 (1967).
50. K. Brennan and K. Hess, "Theory of High-Field Transport of Holes in GaAs and InP," *Phys. Rev. B*, **29**, 5581 (1984).
51. B. Kramer and A. Mircea, "Determination of Saturated Electron Velocity in GaAs," *Appl. Phys. Lett.*, **26**, 623 (1975).
52. K. K. Thornber, "Relation of Drift Velocity to Low-Field Mobility and High Field Saturation Velocity," *J. Appl. Phys.*, **51**, 2127 (1980).
53. J. G. Ruch, "Electron Dynamics in Short Channel Field-Effect Transistors," *IEEE Trans. Electron Devices*, **ED-19**, 652 (1972).
54. K. K. Thornber, "Applications of Scaling to Problems in High-Field Electronic Transport," *J. Appl. Phys.*, **52**, 279 (1981).
55. R. A. Logan and S. M. Sze, "Avalanche Multiplication in Ge and GaAs *p-n* Junctions," Proc. Int. Conf. Phys. Semicond., Kyoto, and *J. Phys. Soc. Jpn. Suppl.*, **21**, 434 (1966).
56. W. N. Grant, "Electron and Hole Ionization Rates in Epitaxial Silicon at High Electric Fields," *Solid-State Electron.*, **16**, 1189 (1973).

57. G. H. Glover, "Charge Multiplication in Au-SiC (6H) Schottky Junction," *J. Appl. Phys.*, **46**, 4842 (1975).
58. T. P. Pearsall, F. Capasso, R. E. Nahory, M. A. Pollack, and J. R. Chelikowsky, "The Band Structure Dependence of Impact Ionization by Hot Carriers in Semiconductors GaAs," *Solid-State Electron.*, **21**, 297 (1978).
59. I. Umebu, A. N. M. M. Choudhury, and P. N. Robson, "Ionization Coefficients Measured in Abrupt InP Junction," *Appl. Phys. Lett.*, **36**, 302 (1980).
60. R. A. Logan and H. G. White, "Charge Multiplication in GaP *p-n* Junctions," *J. Appl. Phys.*, **36**, 3945 (1965).
61. T. P. Pearsall, "Impact Ionization Rates for Electrons and Holes in $Ga_{0.47}In_{0.53}As$," *Appl. Phys. Lett.*, **36**, 218 (1980).
62. T. P. Pearsall, R. E. Nahory, and M. A. Pollack, "Impact Ionization Rates for Electrons and Holes in $GaAs_{1-x}Sb_x$ Alloys," *Appl. Phys. Lett.*, **28**, 403 (1976).
63. L. W. Cook, G. E. Bulman, and G. E. Stillman, "Electron and Hole Impact Ionization Coefficients in InP Determined by Photomultiplication Measurements," *Appl. Phys. Lett.*, **40**, 589 (1982).
64. I. H. Oguzman, E. Bellotti, K. F. Brennan, J. Kolnik, R. Wang, and P. P. Ruden, "Theory of Hole Initiated Impact Ionization in Bulk Zincblende and Wurtzite GaN," *J. Appl. Phys.*, **81**, 7827 (1997).
65. M. R. Brozel and G. E. Stillman, Eds., *Properties of Gallium Arsenide*, 3rd Ed., INSPEC, London, 1996.
66. C. R. Crowell and S. M. Sze, "Temperature Dependence of Avalanche Multiplication in Semiconductors," *Appl. Phys. Lett.*, **9**, 242 (1966).
67. C. T. Sah, R. N. Noyce, and W. Shockley, "Carrier Generation and Recombination in *p-n* Junction and *p-n* Junction Characteristics," *Proc. IRE*, **45**, 1228 (1957).
68. R. N. Hall, "Electron-Hole Recombination in Germanium," *Phys. Rev.*, **87**, 387 (1952).
69. W. Shockley and W. T. Read, "Statistics of the Recombination of Holes and Electrons," *Phys. Rev.*, **87**, 835 (1952).
70. W. M. Bullis, "Properties of Gold in Silicon," *Solid-State Electron.*, **9**, 143 (1966).
71. D. T. Stevenson and R. J. Keyes, "Measurement of Carrier Lifetime in Germanium and Silicon," *J. Appl. Phys.*, **26**, 190 (1955).
72. W. W. Gartner, "Spectral Distribution of the Photomagnetic Electric Effect," *Phys. Rev.*, **105**, 823 (1957).
73. S. Wei and M. Y. Chou, "Phonon Dispersions of Silicon and Germanium from First-Principles Calculations," *Phys. Rev. B*, **50**, 2221 (1994).
74. C. Patel, T. J. Parker, H. Jamshidi, and W. F. Sherman, "Phonon Frequencies in GaAs," *Phys. Stat. Sol.* (b), **122**, 461 (1984).
75. W. C. Dash and R. Newman, "Intrinsic Optical Absorption in Single-Crystal Germanium and Silicon at 77°K and 300°K," *Phys. Rev.*, **99**, 1151 (1955).
76. H. R. Philipp and E. A. Taft, "Optical Constants of Silicon in the Region 1 to 10 eV," *Phys. Rev. Lett.*, **8**, 13 (1962).
77. D. E. Hill, "Infrared Transmission and Fluorescence of Doped Gallium Arsenide," *Phys. Rev.*, **133**, A866 (1964).

78. H. C. Casey, Jr., D. D. Sell, and K. W. Wecht, "Concentration Dependence of the Absorption Coefficient for *n*- and *p*-type GaAs between 1.3 and 1.6 eV," *J. Appl. Phys.*, **46**, 250 (1975).

79. C. Y. Ho, R. W. Powell, and P. E. Liley, *Thermal Conductivity of the Elements—A Comprehensive Review*, Am. Chem. Soc. and Am. Inst. Phys., New York, 1975.

80. M. G. Holland, "Phonon Scattering in Semiconductors from Thermal Conductivity Studies," *Phys. Rev.*, **134**, A471 (1964).

81. B. H. Armstrong, "Thermal Conductivity in SiO_2", in S. T. Pantelides, Ed., *The Physics of SiO_2 and Its Interfaces*, Pergamon, New York, 1978.

82. G. A. Slack, "Thermal Conductivity of Pure and Impure Silicon, Silicon Carbide, and Diamond," *J. Appl. Phys.*, **35**, 3460 (1964).

83. E. K. Sichel and J. I. Pankove, "Thermal Conductivity of GaN, 25–360 K," *J. Phys. Chem. Solids*, **38**, 330 (1977).

84. G. H. Dohler, "Doping Superlattices—Historical Overview", in P. Bhattacharya, Ed., *III-V Quantum Wells and Superlattices*, INSPEC, London, 1996.

85. J. R. Haynes and W. Shockley, "The Mobility and Life of Injected Holes and Electrons in Germanium," *Phys. Rev.*, **81**, 835 (1951).

PROBLEMS

1. (a) Find the maximum fraction of a conventional unit-cell volume which can be filled by identical hard spheres in a diamond lattice.
(b) Find the number of atoms per square centimeter in silicon in (111) plane at 300 K.

2. Calculate the tetrahedral bond angle, i.e., the angle between any pair of the four bonds. (*Hint*: Represent the 4 bonds as vectors of equal lengths. What must be the sum of the 4 vectors equal? Take components of this vector equation along the direction of one of these vectors.)

3. For a face centered cubic, the volume of a conventional unit cell is a^3. Find the volume of a fcc primitive unit cell with three basis vectors: (0,0,0→a/2,0,a/2), (0,0,0→a/2,a/2,0), and (0,0,0→0,a/2,a/2).

4. (a) Derive an expression for the bond length d in the diamond lattice in terms of the lattice constant a.
(b) In a silicon crystal, if a plane has intercepts at 10.86 Å, 16.29 Å, and 21.72 Å along the three Cartesian coordinates, find the Miller indices of the plane.

5. Show (a) that each vector of the reciprocal lattice is normal to a set of planes in the direct lattice, and (b) the volume of a unit cell of the reciprocal lattice is inversely proportional to the volume of a unit cell of the direct lattice.

6. Show that the reciprocal lattice of a body-centered cubic (bcc) lattice with a lattice constant a is a face-centered cubic (fcc) lattice with the side of the cubic cell to be $4\pi/a$.
[*Hint*: Use a symmetric set of vectors for bcc:

$$\boldsymbol{a} = \frac{a}{2}(\boldsymbol{y}+\boldsymbol{z}-\boldsymbol{x}), \quad \boldsymbol{b} = \frac{a}{2}(\boldsymbol{z}+\boldsymbol{x}-\boldsymbol{y}), \quad \boldsymbol{c} = \frac{a}{2}(\boldsymbol{x}+\boldsymbol{y}-\boldsymbol{z})$$

where a is the lattice constant of a conventional primitive cell, and $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ are unity vectors of a Cartesian coordinate. For fcc;

$$a = \frac{a}{2}(y+z), \quad b = \frac{a}{2}(z+x), \quad c = \frac{a}{2}(x+y).]$$

7. Near the conduction band minima the energy can be expressed as

$$E = \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_x^*} + \frac{k_y^2}{m_y^*} + \frac{k_z^2}{m_z^*}\right).$$

In Si there are six cigar-shaped minima along [100]. If the ratio of the axes of constant energy ellipsoid is 5:1, find the ratio of longitudinal effective mass m_l^* to the transverse effective mass m_t^*.

8. In the conduction band of a semiconductor, it has a lower valley at the center of the Brillouin zone, and six upper valleys at the zone boundary along [100]. If the effective mass for the lower valley is $0.1m_0$ and that for the upper valleys is $1.0m_0$, find the ratio of the effective density of states in the upper valleys to that in the lower valley.

9. Derive the density of states in the conduction band as given by Eq. 14.
(*Hint*: The wavelength λ of a standing wave is related to the length of the semiconductor L by $L/\lambda = n_x$ where n_x is an integer. The wavelength can be expressed by de Broglie hypothesis $\lambda = h/p_x$. Consider a three-dimensional cube of side L.)

10. Calculate the average kinetic energy of electrons in the conduction band of an n-type nondegenerate semiconductor. The density of states is given by Eq. 14.

11. Show that

$$N_D^+ = N_D\left[1 + 2\exp\left(\frac{E_F - E_D}{kT}\right)\right]^{-1}.$$

[*Hint*: The probability of occupancy is

$$F(E) = \left[1 + \frac{h}{g}\exp\left(\frac{E - E_F}{kT}\right)\right]^{-1}$$

where h is the number of electrons that can physically occupy the level E, and g is the number of electrons that can be accepted by the level, also called the ground-state degeneracy of the donor impurity level ($g = 2$).]

12. If a silicon sample is doped with 10^{16} phosphorous impurities/cm^3, find the ionized donor density at 77 K. Assume that the ionization energy for phosphorous donor impurities and the electron effective mass are independent of temperature. (*Hint*: First select a N_D^+ value to calculate the Fermi level, then find the corresponding N_D^+. If they don't agree, select another N_D^+ value and repeat the process until a consistent N_D^+ is obtained.)

13. Using graphic method to determine the Fermi level for a boron-doped silicon sample with an impurity concentration of 10^{15} cm^{-3} at 300 K (note $n_i = 9.65\times10^9$ cm^{-3}).

14. The Fermi-Dirac distribution function is given by $F(E) = \dfrac{1}{1 + \exp[(E - E_F)/kT]}$. The differentiation of $F(E)$ with respect to energy is $F'(E)$. Find the width of $F'(E)$, i.e., $2\left[E(\text{at } F'_{\max}) - E\left(\text{at } \frac{1}{2}F'_{\max}\right)\right]$ where $|F'_{\max}|$ is the maximum value of $F'(E)$.

15. Find the position of the Fermi level with respect to the bottom of the conduction band ($E_C - E_F$) for a silicon sample at 300 K, which is doped with 2×10^{10} cm^{-3} fully ionized donors.

16. Gold in Si has two energy levels in the bandgap: $E_C - E_A = 0.54$ eV, $E_D - E_V = 0.29$ eV. Assume the third level $E_D - E_V = 0.35$ eV is inactive. (a) What will be the state of charge of the gold levels in Si doped with high concentration of boron atoms? Why? (b) What is the effect of gold on electron and hole concentrations?

17. From Fig. 13, evaluate and determine what kind of impurity atoms is used to dope the Si sample?

18. For an n-type silicon sample doped with 2.86×10^{16} cm^{-3} phosphorous atoms, find the ratio of the neutral to ionized donors at 300 K. $(E_C - E_D) = 0.045$ eV.

19. (a) Assume the mobility ratio $\mu_n/\mu_p \equiv b$ in Si is a constant independent of impurity concentration. Find the maximum resistivity ρ_m in terms of the intrinsic resistivity ρ_i at 300 K. If $b = 3$ and the hole mobility of intrinsic Si is 450 cm^2/V-s, calculate ρ_i and ρ_m.
(b) Find the electron and hole concentration, mobility, and resistivity of a GaAs sample at 300 K with 5×10^{15} zinc atoms/cm^3, 10^{17} sulfur atoms/cm^3, and 10^{17} carbon atoms/cm^3.

20. The Gamma Function is defined as $\Gamma(n) \equiv \int_0^\infty x^{n-1}\exp(-x)dx$.
(a) Find $\Gamma(1/2)$, and (b) show that $\Gamma(n) = (n-1)\Gamma(n-1)$.

21. Consider a compensated n-type silicon at $T = 300$ K, with a conductivity of $\sigma = 16$ S/cm and an acceptor doping concentration of 10^{17} cm^{-3}. Determine the donor concentration and the electron mobility. (A compensated semiconductor is one that contains both donor and acceptor impurity atoms in the same region.)

22. Find the resistivity at 300 K for a silicon sample doped with 1.0×10^{14} cm^{-3} of phosphorous atoms, 8.5×10^{12} cm^{-3} of arsenic atoms, and 1.2×10^{13} cm^{-3} of boron atoms. Assume that the impurities are completely ionized and the mobilities are $\mu_n = 1500$ cm^2/V-s, $\mu_p = 500$ cm^2/V-s, independent of impurity concentrations.

23. A semiconductor has a resistivity of 1.0 Ω-cm, and a Hall coefficient of –1250 cm^2/Coul. Calculate the carrier density and mobility, assuming that only one type of carrier is present and the mean free time is proportional to the carrier energy, i.e., $\tau \propto E$.

24. Derive the recombination rate for indirect recombination as given by Eq. 92.
(*Hint*: Refer to Fig. 25b, the capture rate of an electron by a recombination center is proportional to $R_e \propto nN_t(1-F)$ where n is the density of electrons in the conduction band, N_t is the density of recombination centers, F is the Fermi distribution, and $N_t(1-F)$ is the density of unoccupied recombination centers available for electron capture.)

25. The recombination rate is given by Eq. 92. Under low injection condition, U can be expressed as $(p_n - p_{no})/\tau_r$, where τ_r is the recombination lifetime. If $\sigma_n = \sigma_p = \sigma_o$, $n_{no} = 10^{15}$ cm^{-3}, and $\tau_{ro} \equiv (\upsilon_{th}\sigma_o N_t)^{-1}$, find the values of $(E_t - E_i)$ at which the recombination lifetime τ_r becomes $2\tau_{ro}$.

26. For single-level recombination with identical electron and hole capture cross sections, find the number of trap centers per unit volume per generation rate under the condition of complete depletion of carriers. Assume that the trap centers are located at mid bandgap, $\sigma = 2\times10^{-16}$ cm^2, and $\upsilon_{th} = 10^7$ cm/s.

27. In a region of semiconductor which is completely depleted of carriers (i.e., $n \ll n_i$, $p \ll n_i$), electron-hole pairs are generated by alternate emission of electrons and of holes by the centers. Derive the average time that takes place between such emission process (assume

$\sigma_n = \sigma_p = \sigma$); also find the average time for $\sigma = 2\times10^{-16}$ cm^2, $\upsilon_{th} = 10^7$ cm/s, and $E_t = E_i$. ($T = 300$ K).

28. For a single-level recombination process, find the average time that takes place between each recombination process in a region of a silicon sample where $n = p = 10^{13}$ cm^{-3}, $\sigma_n = \sigma_p = 2\times10^{-16}$ cm^2, $\upsilon_{th} = 10^7$ cm/s, $N_t = 10^{16}$ cm^{-3}, and $(E_t - E_i) = 5kT$.

29. (a). Derive Eq. 123.
(*Hint*: Assume a linear chain of atoms and the atoms interact only with nearest neighbors. The even-numbered atoms have mass m_1 and the odd-numbered atoms have mass m_2.)
(b) For a silicon crystal with $m_1 = m_2$ and $\sqrt{\alpha_f/m_1} = 7.63\times10^{12}$ Hz, find the optical phonon energy at the boundary of the Brillouin zone. The force constant is α_f.

30. Assume $Ga_{0.5}In_{0.5}As$ is lattice matched with InP substrate at 500°C. When the sample is cooled to 27°C, find the lattice mismatch between the layers.

31. Find the ratio of the conduction-band discontinuity of the heterojunction $Al_{0.4}Ga_{0.6}As$/GaAs to the $Al_{0.4}Ga_{0.6}As$ bandgap.

32. In a Haynes-Shockley experiment, the maximum amplitudes of the minority carriers at $t_1 = 25$ μs and $t_2 = 100$ μs differ by a factor of 10. Find the minority carrier lifetime.

33. From the expression which describes the drift and diffusion of carriers in the Haynes-Shockley experiment, find the half-width of the pulse of carriers at $t = 1$ s. Assume the diffusion coefficient is 10 cm^2/s.

34. Excess carriers are injected on one surface ($x = 0$) of a thin slide of n-type (3×10^{17} cm^{-3}) silicon with length $W = 0.05$ mm and extracted at the opposite surface where $p_n(W) = p_{no}$. If the carrier lifetime is 50 μs, find the portion of injected current which reaches the opposite surface by diffusion.

35. A GaAs n-type sample with $N_D = 5\times10^{15}$ cm^{-3} is illuminated. The uniformly absorbed light creates 10^{17} electron-hole pairs/cm^3-s. The lifetime τ_p is 10^{-7} s, $L_p = 1.93\times10^{-3}$ cm, the surface recombination velocity S_p is 10^5 cm/s. Find the number of holes recombining at the surface per unit surface area in unit time.

36. An n-type semiconductor has excess carrier holes 10^{14} cm^{-3}, a minority carrier lifetime 10^{-6} s in the bulk material, and a minority carrier lifetime 10^{-7} s at the surface. Assume zero applied electric field and let $D_p = 10$ cm^2/s. Determine the steady-state excess carrier concentration as a function of distance from the surface ($x = 0$) of the semiconductor.