CHAPTER W11

Semiconductors

W11.1 Details of the Calculation of $n(T)$ for an n-Type Semiconductor

A general expression for n as a function of both T and N_d can be obtained as follows. After setting $N_a{}^- = 0$, multiplying each term of Eq. (11.34) of the textbook[†] by n, replacing the np product by $n_i p_i$, and rearranging the terms, the following quadratic equation can be obtained:

$$n^2 - N_d{}^+ n - n_i p_i = 0. \tag{W11.1}$$

The following substitutions are now made in this equation: from Eq. (11.27) for n, Eq. (11.28) for $n_i p_i$, and the following expression for $N_d{}^+$:

$$N_d{}^+(T) = N_d - N_d{}^{\circ}(T) = \frac{\frac{1}{2}N_d e^{\beta[E_g - E_d - \mu(T)]}}{\frac{1}{2}e^{\beta[E_g - E_d - \mu(T)]} + 1}. \tag{W11.2}$$

After setting $y = n(T)/N_c(T) = \exp[\beta(\mu(T) - E_g)]$, $w = \exp(-\beta E_d)$, and $z = \exp(-\beta E_g)$, the following equation is obtained:

$$N_c{}^2 y^2 - N_c N_d \frac{w}{(w/y) + 2} - N_c N_v z = 0. \tag{W11.3}$$

The quantities N_c and N_v are defined in Eq. (11.27).

This expression can be rearranged to yield the following cubic equation for $y(T) = n(T)/N_c(T)$:

$$y^3 + \frac{w}{2}y^2 - \left(\frac{N_d w}{2N_c} + \frac{N_v z}{N_c}\right) y - \frac{N_v w z}{2N_c} = 0. \tag{W11.4}$$

The concentration of holes will then be given by

$$p(T) = \frac{n_i(T) p_i(T)}{n(T)}, \tag{W11.5}$$

where $n(T)$ is obtained from Eq. (W11.4).

[†] The material on this home page is supplemental to *The Physics and Chemistry of Materials* by Joel I. Gersten and Frederick W. Smith. Cross-references to material herein are prefixed by a "W"; cross-references to material in the textbook appear without the "W."

In the high-temperature limit when $w \gg y$ [i.e., when $\beta(E_g - \mu(T) - E_d) \approx 2$ or greater], the following quadratic equation is obtained from Eq. (W11.3):

$$y^2 - \frac{N_d}{N_c} y - \frac{N_v}{N_c} z = 0. \tag{W11.6}$$

The appropriate solution of this equation is

$$y = \frac{N_d/N_c + \sqrt{{N_d}^2/{N_c}^2 - 4(-N_v z/N_c)}}{2}. \tag{W11.7}$$

In the $T \to 0$ K limit the terms in Eq. (W11.4) containing $z = \exp(-\beta E_g)$ can be neglected, with the following result:

$$y^2 + \frac{w}{2} y - \frac{N_d w}{2N_c} = 0. \tag{W11.8}$$

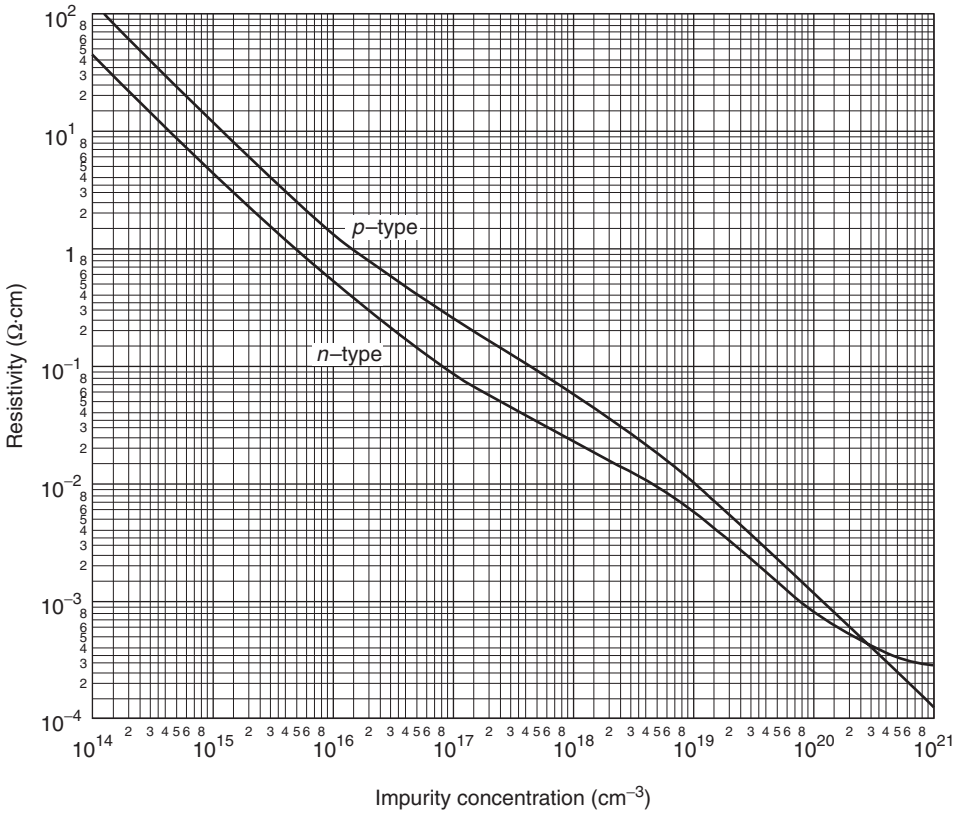

Figure W11.1. Effects of n- and p-type doping on the electrical resistivity of Si at $T = 300$ K, with ρ plotted versus the dopant concentration on a logarithmic plot. (From J. C. Irvin, *The Bell System Technical Journal*, **41**, 387 (1962). Copyright © 1962 AT&T. All rights reserved. Reprinted with permission.)

Solving this quadratic equation and also making use of the fact that $w \ll 8N_d/N_c$ yields

$$y(T) = \sqrt{\frac{N_d w}{2N_c}}. \tag{W11.9}$$

In the intermediate temperature region, where $y \ll w$, $z \ll y^2$ (i.e., $E_g > 4[E_g - \mu(T)] > 8E_d$), and $z \ll N_d w/2N_c$, Eq. (W11.4) becomes

$$\frac{w}{2}y^2 - \frac{N_d w}{2N_c}y = 0 \quad \text{or} \quad y(T) = \frac{N_d}{N_c}, \tag{W11.10}$$

which can be written as $n(T) = N_d$.

W11.2 Effects of Doping on Resistivity of Silicon

The effects of doping on the electrical resistivity of Si at $T = 300$ K are presented in Fig. W11.1, where ρ is shown plotted versus the dopant concentration N_d or N_a in a logarithmic plot. The resistivity decreases from the intrinsic value of $\rho \approx 3000\ \Omega\cdot$m with increasing N_d or N_a. Scattering from ionized dopant atoms also plays a role in causing deviations at high values of N_d or N_a from what would otherwise be straight lines with slopes of -1 on such a plot.

W11.3 Optical Absorption Edge of Silicon

The absorption edge of Si is shown in Fig. W11.2, where the absorption coefficient α determined from measurements of reflectance and transmittance at $T = 300$ K for a single-crystal Si wafer is plotted as $(\alpha\hbar\omega)^{1/2}$ versus $E = \hbar\omega$. The linear nature of this plot is in agreement with the prediction of Eq. (11.54). The onset of absorption at about 1.04 eV corresponds to $\hbar\omega = E_g - \hbar\omega_{\text{phonon}}$, while the additional absorption appearing at about 1.16 eV corresponds to $\hbar\omega = E_g + \hbar\omega_{\text{phonon}}$. These two distinct absorption

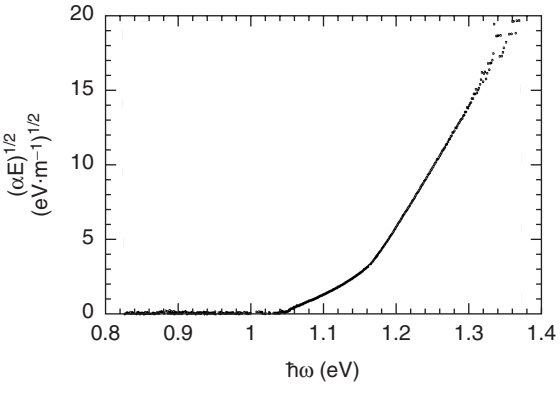

Figure W11.2. Optical absorption edge for Si at $T = 300$ K with the absorption coefficient α plotted as $(\alpha\hbar\omega)^{1/2}$ versus the photon energy $E = \hbar\omega$. The energy gap $E_g = 1.11$ eV and the energy of the phonon $\hbar\omega_{\text{phonon}} \approx 0.06$ eV participating in this indirect optical transition can be obtained in this way. (From Z. L. Akkerman, unpublished data.)

onsets which are separated from $E_g = 1.11$ eV by $\hbar\omega_{\text{phonon}} = 0.06$ eV ≈ 485 cm^{-1} are the result of the absorption and emission, respectively, of the phonon, which participates in this indirect transition. If Si were a direct-bandgap semiconductor such as GaAs, there would be only a single onset at $\hbar\omega = E_g$. In this way both E_g and the energy of the participating phonon can be obtained from straightforward optical measurements. The absorption onset associated with phonon absorption will become weaker as the temperature decreases since fewer phonons will be available, while that associated with phonon emission will be essentially independent of temperature.

W11.4 Thermoelectric Effects

The equilibrium thermal properties of semiconductors (i.e., the specific heat, thermal conductivity, and thermal expansion) are dominated by the phonon or lattice contribution except when the semiconductor is heavily doped or at high enough temperatures so that high concentrations of intrinsic electron–holes pairs are thermally excited. An important and interesting situation occurs when temperature gradients are present in a semiconductor, in which case nonuniform spatial distributions of charge carriers result and thermoelectric effects appear. Semiconductors display significant bulk thermoelectric effects, in contrast to metals where the effects are usually orders of magnitude smaller. Since the equilibrium thermal properties of materials are described in Chapters 5 and 7, only the thermoelectric power and other thermoelectric effects observed in semiconductors are discussed here. Additional discussions of the thermopower and Peltier coefficient are presented in Chapter W22.

The strong thermoelectric effects observed in semiconductors are associated with the electric fields that are induced by temperature gradients in the semiconductor, and vice versa. The connections between a temperature gradient ∇T, a voltage gradient ∇V or electric field $E = -\nabla V$, a current density $\mathbf{J}$, and a heat flux $\mathbf{J}_Q$ (W/m^2) in a material are given as follows:

$$\begin{aligned} \mathbf{J} &= \sigma(\mathbf{E} - S\nabla T) = \mathbf{J}_E + \mathbf{J}_{\nabla T}, \\ \mathbf{J}_Q &= \sigma\Pi\mathbf{E} - \kappa\nabla T. \end{aligned} \tag{W11.11}$$

Here σ and κ are the electrical and thermal conductivities, respectively. The quantity S is known as the *Seebeck coefficient*, the *thermoelectric power*, or simply the *thermopower*, and Π is the *Peltier coefficient*. While the electrical and thermal conductivities are positive quantities for both electrons and holes, it will be shown later that the thermopower S and Peltier coefficient Π are negative for electrons and positive for holes (i.e., they take on the sign of the responsible charge carrier).

The Seebeck and Peltier effects are illustrated schematically in Fig. W11.3. The thermopower S can be determined from the voltage drop ΔV resulting from a temperature difference ΔT in a semiconductor in which no net current $\mathbf{J}$ is flowing and no heat is lost through the sides. Since $\mathbf{J} = 0$ as a result of the cancellation of the electrical currents $\mathbf{J}_E$ and $\mathbf{J}_{\nabla T}$ flowing in opposite directions due to the voltage and temperature gradients, respectively, it can be seen from Eq. (W11.11) that $\mathbf{E} = S\nabla T = -\nabla V$. Therefore, S is given by

$$S = -\frac{\nabla V}{\nabla T} = -\frac{\Delta V}{\Delta T} \tag{W11.12}$$

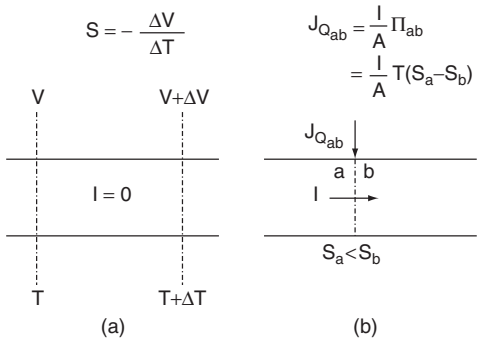

Figure W11.3. Seebeck and Peltier effects. (*a*) In the Seebeck effect a voltage difference ΔV exists in a material due to the temperature difference ΔT. The Seebeck coefficient or thermopower of the material is given by $S = -\Delta V/\Delta T$. (*b*) In the Peltier effect a flow of heat into (or out of) a junction between two materials occurs when a current I flows through the junction.

and has units of V/K. Since ΔV and ΔT have the same sign for electrons and opposite signs for holes, it follows that a measurement of the sign of S is a convenient method for determining the sign of the dominant charge carriers. The physical significance of S is that it is a measure of the tendency or ability of charge carriers to move from the hot to the cold end of a semiconductor in a thermal gradient.

The Peltier coefficient $\Pi(T)$ of a material is related to its thermopower $S(T)$ by the *Kelvin relation*:

$$\Pi(T) = TS(T). \tag{W11.13}$$

Therefore, Π has units of volts. The physical significance of the Peltier coefficient Π of a material is that the rate of transfer of heat $\mathbf{J}_{Qab}$ occurring at a junction between two materials a and b when a current is flowing through the junction from a to b is proportional to the difference $\Pi_{ab} = \Pi_a - \Pi_b$. Note that $\mathbf{J}_{Qab} < 0$ Fig. W11.3, corresponding to the flow of heat into the junction. The Peltier effect in semiconductors can be used for thermoelectric power generation or for cooling.

There is an additional thermoelectric effect, the *Thomson effect*, which corresponds to the flow of heat into or out of a material carrying an electrical current in the presence of a thermal gradient. The Thomson effect will not be described here since it usually does not play an important role in the thermoelectric applications of semiconductors.

In the one-dimensional case for the Seebeck effect in a semiconductor the induced electric field E_x is given by $S\,dT/dx$ and the thermopower is given by

$$S = \frac{1}{qT}\left(\frac{\langle \tau E_{e,h}\rangle}{\langle \tau\rangle} - \mu\right). \tag{W11.14}$$

In this expression $E_{e,h}$ is the kinetic energy of the charge carriers (i.e., the energy $E_e = E - E_c$ of an electron relative to the bottom of the conduction band or the energy $E_h = E_v - E$ of a hole relative to the top of the valence band). In addition, $q = \pm e$ is the charge of the dominant charge carriers. Also, the chemical potential μ is constant in space in the absence of net current flow, $\tau(E)$ is the energy-dependent scattering or momentum relaxation time for the charge carriers, and $\langle \tau\rangle$ and $\langle \tau E\rangle$ are the averages of these quantities over the appropriate distribution function.

When $\tau(E)$ obeys a power law (e.g., $\tau \propto E^r$), the thermopower for an n-type semiconductor is

$$S_n(T) = -\frac{k_B}{e}\left(\frac{E_c - \mu}{k_B T} + r + \frac{5}{2}\right), \tag{W11.15}$$

while for a p-type semiconductor,

$$S_p(T) = \frac{k_B}{e}\left(\frac{\mu - E_v}{k_B T} + r + \frac{5}{2}\right). \tag{W11.16}$$

The exponent r is equal to $-\frac{1}{2}$ for acoustic phonon scattering. The thermopowers of semiconductors are typically hundreds of times larger than those measured for metals, where, according to the free-electron model,

$$S = -\frac{\pi^2}{6}\frac{k_B}{e}\frac{k_B T}{E_F} \approx 1\ \mu\text{V/K}.$$

Physically, S is smaller in metals than in semiconductors due to the high, temperature-independent concentrations of electrons in metals. In this case only a relatively small thermoelectric voltage is required to produce the reverse current needed to balance the current induced by the temperature gradient.

The Peltier effect in a semiconductor is illustrated schematically in Fig. W11.4, where an electric field **E** is applied across the semiconductor by means of two metal contacts at its ends. As a result, the energy bands and the Fermi energy E_F slope downward from left to right. In the n-type semiconductor in which electrons flow from left to right, only the most energetic electrons in metal I are able to pass into the semiconductor over the energy barrier $E_c - \mu$ at the metal–semiconductor junction on the left. When the electrons leave the semiconductor and pass through the metal–semiconductor junction into metal II at the right, the reverse is true and they release an amount of heat equal to $(E_c - \mu + ak_BT)$ per electron. The term ak_BT represents the kinetic energy

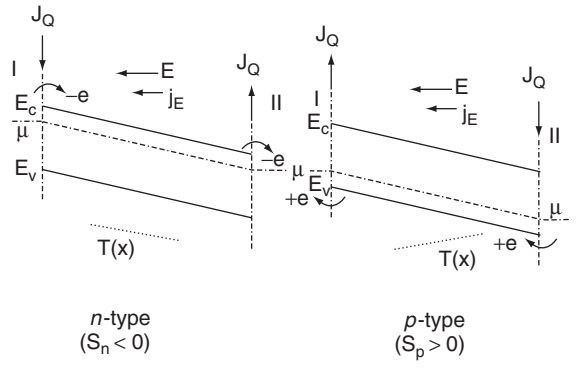

Figure W11.4. Peltier effect in a semiconductor. An electric field **E** is applied across a semiconductor, and as a result, the energy bands and the chemical potential μ slope downward from left to right. In the n-type semiconductor, electrons flow from left to right and in the p-type semiconductor holes flow from right to left. The resulting temperature gradient is also shown for each case.

transferred by the electron as it moves through the semiconductor, with $a \approx 1.5$ to 2, depending on the dominant scattering process. Therefore, the net heat flow due to electrons is from left to right through the semiconductor, with the temperature gradient in the direction shown. It follows in this case for electrons that the magnitude of the Peltier coefficient (i.e., the net energy transported by each electron divided by the charge e) is

$$\Pi_n(T) = TS_n(T) = \frac{E_c - \mu + ak_BT}{e}. \tag{W11.17}$$

This result is consistent with Eq. (W11.15). Note that the position of the chemical potential μ within the energy gap can be determined from a measurement of Π_n as $T \to 0$ K.

For the p-type semiconductor shown in Fig. W11.4, holes will flow from right to left. Since the energy of a hole increases in the downward direction on this electron energy scale, only the most energetic holes can pass into the semiconductor over the energy barrier $\mu - E_v$ at the junction on the right. In this case the net heat flow is from right to left, with the temperature gradient in the direction shown. It follows for holes that

$$\Pi_p(T) = TS_p(T) = \frac{\mu - E_v + ak_BT}{e}, \tag{W11.18}$$

which is consistent with Eq. (W11.16).

The contribution of phonons to the thermoelectric power originates in the *phonon drag effect*, the tendency of phonons diffusing from the hot to the cold end of a material to transfer momentum to the electrons, thereby "dragging" them along in the same direction. This effect becomes more noticeable at lower temperatures.

Experimental results and theoretical predictions for the Peltier coefficient Π for n- and p-type Si as functions of temperature are shown in Fig. W11.5. The Si samples

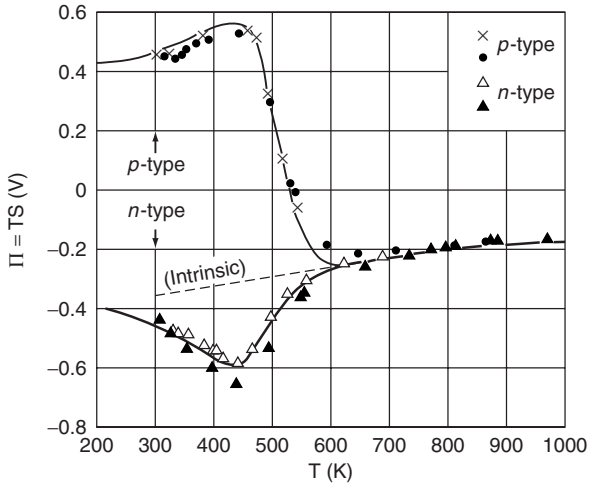

Figure W11.5. Experimental results (points) and theoretical predictions (solid lines) for the Peltier coefficient Π for n- and p-type Si are shown as functions of temperature. The Si samples show intrinsic behavior above $T \approx 600$ K. (From T. H. Geballe et al., *Phys. Rev.*, **98**, 940 (1955). Copyright © 1955 by the American Physical Society.)

show intrinsic behavior above $T \approx 600$ K. Note that plots of $e\Pi$ versus T yield as intercepts at $T = 0$ K, the quantities $-(E_c - \mu)$ and $(\mu - E_v)$ for n- and p-type semiconductors, respectively. This is a convenient way of determining the position of the chemical potential μ relative to the band edges in doped semiconductors.

W11.5 Dielectric Model for Bonding

In the dielectric model of Phillips and Van Vechten (PV) for tetrahedrally coordinated semiconductors with diamond and zincblende crystal structures the chemical bonding is considered to be the sum of covalent and ionic contributions. As discussed in Section 2.6, f_c is the fraction of covalent bonding in an A–B bond involving atoms A and B, while the ionic fraction or ionicity is $f_i = 1 - f_c$. Values of f_i obtained on the basis of the PV model are presented in Table 2.6. These values are based on the dielectric properties of these materials and differ somewhat from those proposed by Pauling, which are based on the thermochemistry of solids.

In the PV model the *average total energy gap* E_g(A–B) in, for example, a binary compound AB containing only A–B bonds is defined as the average energy separation between the bonding and antibonding energy levels associated with the orbitals involved in the A–B bond. Thus E_g is not an observable quantity and is in some sense an average energy gap between the valence and conduction bands. A spectroscopic or dielectric definition for E_g is used in the PV model rather than a thermochemical definition based on heats of formation or cohesive energies. Specifically, E_g(A-B) is defined experimentally in terms of the measured optical dielectric function by

$$\frac{\epsilon(0)}{\epsilon_o} = 1 + A_1 \left(\frac{\hbar\omega_p}{E_g}\right)^2, \tag{W11.19}$$

where

$$\omega_p^2 = \frac{ne^2}{m\epsilon_o}.$$

Here $\epsilon(0)/\epsilon_o = n^2(0)$ is the real, zero-frequency limit of the complex dielectric function $\epsilon(\omega, \mathbf{q})/\epsilon_o$, also known as the relative permittivity ϵ_r, and ω_p is the *plasma frequency*. Also, n is the concentration of valence electrons, ϵ_o the permittivity of free space, and A_1 a correction factor that is close to 1 which accounts for the possible participation of d electrons in the optical response. The bonding–antibonding energy gap E_g(A–B) differs from and is typically much larger than the optical energy gap $E_g = E_c - E_v$. Equation (W11.19) is close in form to the expression given in Eq. (8.32), which is derived from the Lorentz oscillator model for the optical dielectric function.

When the A–B bond is of a mixed ionic–covalent type, the gap E_g(A–B) is taken to be complex, with a real *covalent* or *homopolar* component E_h and an imaginary *ionic* or *heteropolar* component iC, so that

$$\begin{aligned} E_g(\text{A–B}) &= E_h + iC, \\ |E_g|^2 &= E_h^2 + C^2. \end{aligned} \tag{W11.20}$$

The definitions of E_h and C in terms of microscopic parameters associated with the A–B bond and the binary AB compound are

$$\begin{aligned} E_h(\text{A–B}) &= \frac{A_2}{d^{2.5}}, \\ C(\text{A–B}) &= 14.4b\left(\frac{z_\text{A}}{r_\text{A}} - \frac{z_\text{B}}{r_\text{B}}\right)\exp\left(-\frac{k_\text{TF}d}{2}\right). \end{aligned} \tag{W11.21}$$

where $A_2 = 39.74$ eV, the dimensionless constant $b \approx 1.5$, d is the A–B interatomic distance or bond length, and z_A and z_B are the valences and r_A and r_B the covalent radii of atoms A and B, respectively, with $d = r_\text{A} + r_\text{B}$. Here E_h and C are given in eV when r_A and r_B are in angstrom units. The exponential Thomas–Fermi screening factor, defined in Section 7.17, describes the screening of the ion cores by the valence electrons and is expressed in terms of the *Thomas–Fermi wave vector* or inverse screening length:

$$k_\text{TF} = \sqrt{\frac{3ne^2}{2\epsilon E_F}} = \sqrt{\frac{e^2\rho(E_F)}{\epsilon}}, \tag{W11.22}$$

where n is the concentration of valence electrons, E_F the Fermi energy, ϵ the permittivity of the material, and $\rho(E_F)$ the electron density of states per unit volume. Typical values of k_TF are $\approx 5 \times 10^{10}$ m^{-1}. It can be seen that C(A–B) is given by the difference between the Coulomb potentials of the two atoms A and B composing the bond.

The use of known values of d(A–A) and of E_g(A–A) determined from $\epsilon(0)$ using Eq. (W11.19) for the covalent elemental semiconductors diamond and Si allows both the exponent of d, -2.5, and the constant $A_2 = 39.74$ eV to be determined in the expression for E_h. The ionic component C(A–B) of E_g(A–B) for binary AB semiconductors can then be calculated using Eq. (W11.20) from empirical values of E_g determined from Eq. (W11.19) and values of E_h(A–B) calculated from Eq. (W11.21). It has been shown empirically that the ionic contribution $C(\text{A–B}) \propto X_\text{A} - X_\text{B}$, the difference of the electronegativities of the two atoms.

The ionicity of the A–B bond is defined in a straightforward manner by

$$f_i = \frac{C^2}{{E_g}^2}. \tag{W11.23}$$

Thus $f_i = 0$ when $C = 0$ and $f_i \to 1$ for $C \gg E_h$. The ionicities presented in Table 2.6, known as spectroscopic ionicities, have been calculated in this way using the PV model. For group III–V compounds it has been found that C is usually smaller than E_h so that $f_i < 0.5$. The bonding in these compounds is therefore predominantly covalent. The reverse is true for the group II–VI and I–VII compounds, where C is usually greater than E_h.

Values of E_h, C, $E_g(A-B)$, and f_i for several semiconductors with the diamond or zincblende crystal structures are presented in Table W11.1. Note that E_h is nearly constant for isoelectronic sequences (e.g., for Ge, GaAs, and ZnSe), where $E_h \approx$ 4.3 eV, since their NN distances d are nearly constant. The optical energy gap E_g and the average total energy gap E_g(A–B) are neither proportional to nor simply

TABLE W11.1 Values of E_h, C, $E_g(A-B)$, and f_i for Several Semiconductors

Semiconductor							
IV	III–V	II–VI	E_h (eV)	C (eV)	$E_g(A-B)$ (eV)	f_i	E_g/E_g(A–B)
C (diamond)			13.5	0	13.5	0	0.40
	BN		13.1	7.71	15.2	0.256	0.39
		BeO	11.5	13.9	18.0	0.602	0.52
3C–SiC (β-SiC)			8.27	3.85	9.12	0.177	0.25
Si			4.77	0	4.77	0	0.23
	AlP		4.72	3.14	5.67	0.307	0.43
		MgS	3.71	7.10	8.01	0.786	0.55
Ge			4.31	0	4.31	0	0.16
	GaAs		4.32	2.90	5.20	0.310	0.26
		ZnSe	4.29	5.60	7.05	0.630	0.37
Gray Sn			3.06	0	3.06	0	0.026
	InSb		3.08	2.10	3.73	0.321	0.028
		CdTe	3.08	4.90	5.79	0.717	0.25

related to each other [e.g., for the group IV elements, the ratio E_g/E_g(A–B) decreases from 0.4 for diamond to 0.026 for gray Sn].

A test of the usefulness of this definition of ionicity has been provided by correlating f_i with the crystal structures of about 70 binary group IV–IV, III–V, II–VI, and I–VII compounds. It is found that compounds with $f_i < f_{ic} = 0.785$ are all tetrahedrally coordinated and semiconducting with either the diamond, zincblende, or wurtzite crystal structures, while those with $f_i > 0.785$ are all octahedrally coordinated and insulating with the higher-density NaCl crystal structure. This is an impressive confirmation of the usefulness of the definition of ionicity provided by the PV model.

A definition of electronegativity has also been formulated in the PV model for nontransition metal elements with tetrahedral coordination. This definition differs from that of Pauling presented in Section 2.9 by including the screening of the ion cores by the valence electrons and is likely to be a more useful definition for this group of elements and crystal structures.

W11.6 Nonstandard Semiconductors

In addition to the standard semiconductors discussed in our textbook, which typically have the diamond, zincblende, wurtzite, or NaCl crystal structures, there also exist nonstandard semiconducting materials with a variety of other structures and properties, including disordered or amorphous semiconductors, oxide, organic, and magnetic semiconductors, and porous Si. Some interesting and technologically important examples of these semiconductors are next discussed briefly.

Amorphous Semiconductors. Amorphous semiconductors that lack the long-range order found in their crystalline counterparts often retain to a first approximation the short-range order corresponding to the NN local bonding configurations present in the crystal. For example, in amorphous Si (a-Si) essentially every Si atom is bonded to four NN Si atoms in a nearly tetrahedral arrangement, with bond lengths close to the crystalline value but with a significant spread of bond angles, $\approx 7^{\circ}$, centered

around the ideal value of 109.47°. As a result, a-Si and crystalline Si (c-Si) are similar in many respects, including atomic density and the fact that both are semiconductors with similar energy gaps. They differ appreciably in other important respects, including carrier mobility and ease of doping. The most important defects in a-Si correspond to broken or *dangling bonds* that are likely to be associated with voids in the material and that give rise to electronic levels lying deep within the energy gap. In addition, distorted or weak Si–Si bonds can give rise to electronic states, referred to as *tail states*, that are localized in space and that lie within the energy gap near the band edges.

The electron densities of states of c-Si, a-Si, and a-Si:H in and near the energy gap are shown schematically in Fig. W11.6. The density of states for c-Si has sharp edges at $E = E_v$ and at $E = E_c$. While the densities of states for the amorphous case are very material dependent, there exists a strong similarity between the overall shapes of the curves except in the gap region itself. The dangling-bond defect states in a-Si *pin* the Fermi energy E_F, thereby preventing its movement in the gap. These defect states thus interfere with the doping of this material and consequently with its electronic applications.

The optical dielectric functions of c-Si and a-Si are compared in Fig. W11.7a. The optical response in the crystalline and amorphous phases is qualitatively the same, especially at low energies where $\epsilon_1(0) = n^2(0)$ is essentially the same since the atomic density of the sample of a-Si is only slightly less than that of c-Si. At higher energies it can be seen that the structure in ϵ_1 and ϵ_2 observed in c-Si which is related to the existence of long-range order is absent in the amorphous material where **k** conservation is no longer required. The value of the optical energy gap E_{opt} in amorphous semiconductors such as a-Si and a-Si:H is often obtained using the *Tauc law* for band-to-band

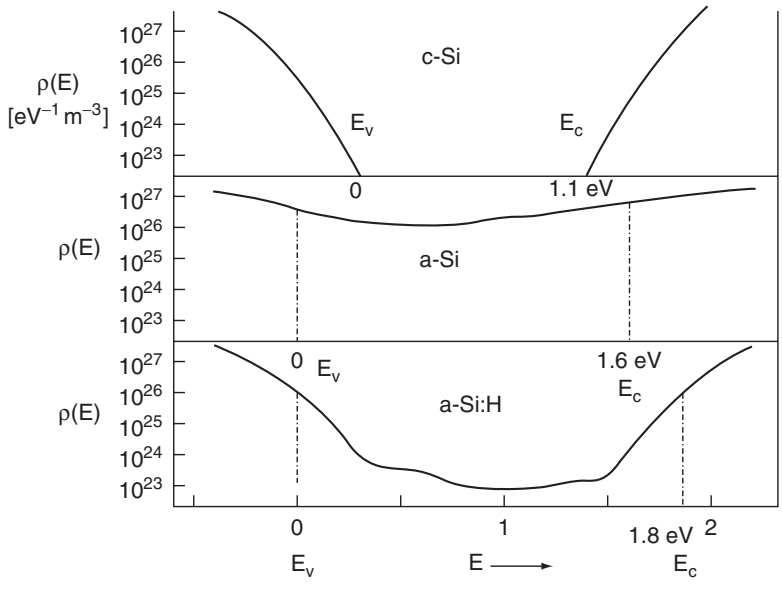

Figure W11.6. Electron densities of states in crystalline Si, a-Si, and a-Si:H in the region of the energy gap.

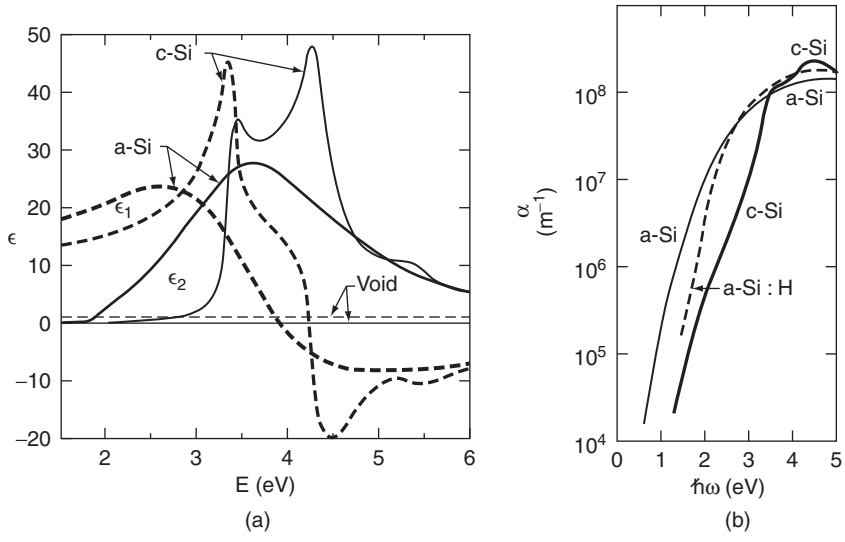

Figure W11.7. Comparison of the optical properties of crystalline and amorphous Si. (*a*) The quantities ϵ_1 (dashed lines) and ϵ_2 (solid lines) of c-Si and a-Si are plotted versus photon energy $E = \hbar\omega$. (From B. G. Bagley et al., in B. R. Appleton and G. K. Celler, eds., *Laser and Electron-Beam Interactions with Solids*, Copyright 1982, with permission from Elsevier Science). (*b*) The logarithm of the optical absorption coefficient α is plotted as a function of photon energy $\hbar\omega$ for c-Si, a-Si, and a-Si:H. (Data from E. D. Palik, *Handbook of Optical Constants of Solids*, Vol. 1, Academic Press, San Diego, Calif., 1985.)

absorption:

$$\epsilon_2(\omega) = \frac{B(\hbar\omega - E_{\text{opt}})^2}{(\hbar\omega)^2}, \tag{W11.24}$$

where B is a constant and $E_{\text{opt}} \approx E_c - E_v$. The parameter E_{opt} can therefore be obtained from a plot of $\hbar\omega\sqrt{\epsilon_2}$ versus $\hbar\omega$. Absorption at lower energies involving the tail states at either the valence- or conduction-band edges is often observed to depend exponentially on $\hbar\omega$, according to the *Urbach edge* expression:

$$\alpha(\omega) = \alpha_{\text{o}} \exp\left(\frac{\hbar\omega}{E_{\text{o}}}\right). \tag{W11.25}$$

Here E_{o} is the Urbach edge parameter and is related to the width of the tail-state regions, while α_{o} is a constant. In high-quality a-Si:H films, E_{o} can be as low as 0.05 eV.

Even though the optical energy gap is larger for a-Si, ≈ 1.6 eV, than for c-Si, light is still absorbed in a-Si for energies below 1.6 eV. In fact, as shown in Fig. W11.7*b*, both a-Si and a-Si:H have much higher absorption coefficients than c-Si in the region of the visible spectrum up to 3 eV, at which point direct transitions begin in c-Si. This is due in part to the fact that in c-Si the absorption corresponds to indirect transitions for energies below 3 eV and also to the fact that absorption in a-Si can occur below the optical gap due to transitions from localized to extended states, and vice versa. Thus films of a-Si:H in photovoltaic solar cells with thicknesses ≈ 1 μm are thick enough

to absorb most of the solar spectrum, while much thicker films of c-Si are required for the same purpose.

In a-Si and other amorphous semiconductors such as a-Ge there exist *mobility edges* located at E_v and E_c, respectively, as shown in Fig. W11.6. These mobility edges for charge carriers typically lie in the tail-state regions and divide electron states in the gap which are spatially localized from those in the energy bands that extend throughout the material. The corresponding charge-carrier mobilities μ_e and μ_h are essentially zero within the gap and are finite for $E < E_v$ and $E > E_c$ within the bands. Thermally activated conduction of charge can still occur within the localized states in the gap and at low temperatures will take place via variable-range hopping, as described in Chapter 7.

Hydrogenated amorphous Si (a-Si:H) is a particularly useful alloy in which the incorporation of H atoms leads to the removal of localized defect states from the energy gap of a-Si by forming Si–H bonds with most of the Si atoms which otherwise would have dangling bonds. The tail states associated with weak Si–Si bonds in a-Si can also be eliminated via the formation of pairs of strong Si–H bonds. The electrons occupying the strong Si–H bonds have energy levels lying within the valence band of the material, well below the band edge at E_v. In this way the concentration of electrically active defects can be reduced from $\approx 10^{26}$ eV^{-1} m^{-3} in a-Si (about one active defect per 10^3 Si atoms) to $\approx 10^{21}$ eV^{-1} m^{-3} in a-Si:H (one active defect per 10^8 Si atoms). The density of states in a-Si:H resulting from the incorporation of hydrogen is also shown in Fig. W11.6. A schematic model of a segment of the continuous random network (CRN) corresponding to the bonding in a-Si:H is shown in Fig. W11.8. Four H atoms are shown completing the Si bonds at a Si monovacancy. This is an example of the type of three-dimensional CRN structure discussed in Chapter 4. Films of a-Si:H are typically formed by plasma deposition from the vapor phase onto substrates usually held at $T \approx 250°C$.

The a-Si:H alloys can be successfully doped n- or p-type during deposition using the standard dopant atoms P and B and as a result have found important applications in photovoltaic solar cells and in the thin-film transistors (TFTs) used as switching elements in flat panel displays. These applications are described in Sections W11.8 and

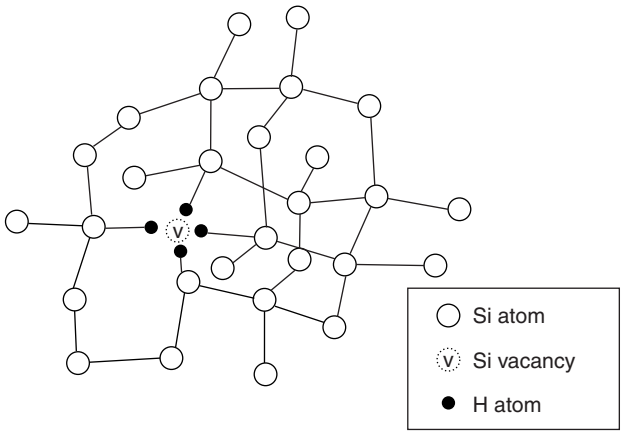

Figure W11.8. Model of a segment of the continuous random network corresponding to the bonding in a-Si:H. Four H atoms are shown completing the Si bonds at a Si monovacancy.

W11.10. The extended-state carrier mobilities in a-Si:H, $\mu_e \approx 10^{-4}$ to 10^{-3} $m^2/V{\cdot}s$ and $\mu_h \approx 3 \times 10^{-7}$ $m^2/V{\cdot}s$, are well below those found in crystalline Si, $\mu_e \approx 0.19$ $m^2/V{\cdot}s$, due to the disorder and increased scattering present in the amorphous material. The electrical conductivities attainable in a-Si:H by doping, $\sigma_n \approx 1$ Ω^{-1} m^{-1} and $\sigma_p \approx 10^{-2}$ Ω^{-1} m^{-1}, are also well below those readily attainable in c-Si, $\sigma \approx 10^4$ Ω^{-1} m^{-1}.

In amorphous alloys based on Si, C, and H, the optical gap can be varied from $E_g \approx 1.8$ eV for a-Si:H to above 3 eV for a-$Si_{0.5}C_{0.5}$:H, thus making the latter material useful as a "window" layer in photovoltaic solar cells. The attainment of even larger gaps at higher C contents is limited by the tendency in carbon-rich alloys for a mixture of tetrahedral (i.e., diamond-like) and trigonal (i.e., graphite-like) bonding of the C atoms to be present. The amorphous graphitic component of hydrogenated amorphous carbon, a-C:H, has an energy gap $E_g \approx 0.5$ eV.

Amorphous semiconducting chalcogenide-based glasses such as a-Se and a-As_2S_3 have both covalent and van der Waals components in their chemical bonding, as discussed in Section 2.2. These amorphous materials can contain molecular units such as $(Se)_8$ and therefore have networks of lower dimensionality and greater structural flexibility than a-Si and a-Ge in which the bonding is three-dimensional. A schematic model of the essentially two-dimensional CRN of a-As_2S_3 and other related materials is shown in Fig. 4.12. In these chalcogenide glasses, group V elements such as As are threefold coordinated and group VI elements such as S and Se are twofold coordinated, as in the crystalline counterparts. The highest-filled valence band in these materials typically consists of electrons occupying lone-pair orbitals on the chalcogenide atoms rather than electrons participating in chemical bonds with their NNs. These glasses are typically formed by rapid quenching from the liquid phase. Applications of amorphous chalcogenide-based glasses include their use in xerography as photoconductors, as described in Chapter 18.

Oxide Semiconductors. Some well-known oxide semiconductors include Cu_2O (cuprite), CuO, and CuO_2. Some group III–V compounds which include oxygen as the group V element are listed in Table 11.9. Semiconducting oxides such as SnO_2, In_2O_3, ITO (indium–tin oxide), Cd_2SnO_4, and ZnO can be prepared as transparent, conducting coatings and have found a wide range of applications (e.g., as transparent electrodes for photovoltaic solar cells).

Copper-based oxides such as La_2CuO_4 with $E_g \approx 2.2$ eV and with the perovskite crystal structure have received considerable attention recently due to the discovery of the high-T_c superconductivity that is observed when they become metallic through doping or alloying. For example, when La_2CuO_4 becomes p-type through the replacement of La^{3+} by Sr^{2+}, the resulting material $La_{2-x}Sr_xCuO_4$ is metallic for $x > 0.06$ and becomes superconducting at low temperatures, as described in Chapter 16.

Organic Semiconductors. Conjugated organic materials such as polymers possessing resonant π-electron bonding can be classified as semiconductors when the energy gap E_g associated with the π-electron system is in the range 1 to 3 eV. The one-dimensional polymer polyacetylene, $(CH)_n$, with alternating single and double carbon–carbon bonds, can possess very high electrical conductivities, exceeding that of copper, when suitable n-type (Na or Hg) or p-type (I) dopants are introduced. Other polymers, such as polypyrrole and polyaniline, can also exhibit high conductivities when suitably doped. A detailed description of the electronic structure and doping of

polyacetylene is presented in Chapter W14. The large nonlinear optical effects found in these materials may lead to important optoelectronic applications. Other applications include their use as photoconductors in xerography.

Semiconducting organic molecular crystals can also exhibit strong electroluminescence and photoluminescence and thus have potential applications in organic light-emitting diodes.

Magnetic Semiconductors. Wide-bandgap ZnS and CdTe and narrow-bandgap HgTe group II–VI semiconductors when alloyed with magnetic impurities such as Mn (e.g., $Zn_{1-x}Mn_xS$ with $0 \leq x \leq 0.5$) have potentially important applications based in part on the "giant" Faraday rotations and negative magnetoresistances which they can exhibit. The sp–d exchange interaction between the s and p conduction-band electrons and the d electrons of the magnetic ions leads to very large Zeeman splittings at the absorption edge and also of the free-exciton level. This sp–d interaction provides the mechanism for the Faraday rotation observed for light propagating in the direction of an applied magnetic field. The magnetic properties of these materials, known as dilute magnetic semiconductors, are discussed briefly in Chapter W17.

Porous Si. An interesting form of Si that may have useful light-emitting applications is porous Si, prepared via electrochemical etching of the surfaces of Si wafers. Porous Si is believed to be a network composed of nanometer-sized regions of crystalline Si surrounded by voids which can occupy between 50 to 90% of the volume of the material. A transmission electron micrograph of porous Si in which the Si columns are about 10 nm in diameter and the pore spaces are about 50 nm wide is shown in Fig. W11.9. Tunable room-temperature photoluminescence in porous Si has been achieved from the near-infrared to the blue-green region of the visible spectrum.

Proposals for the origins of the light emission from porous Si have focused on the quantum confinement of charge carriers in Si regions with dimensions of 2 to 3 nm. Other possible explanations are that oxidized regions with their larger bandgaps or the effects of impurities such as hydrogen can explain the emission of light. It seems clear in any case that oxygen and hydrogen play important roles in chemically passivating the surfaces of the Si nanocrystals. These surfaces would otherwise provide surface recombination sites that would quench the observed luminescence.

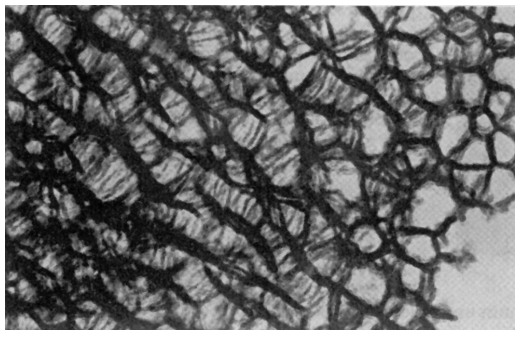

Figure W11.9. Transmission electron micrograph of porous Si in which the Si columns are about 10 nm in diameter and the pore spaces are about 50 nm wide. (Reprinted with permission of A. G. Cullis. From R. T. Collins et al., *Phys. Today*, Jan. 1997, p. 26.)

W11.7 Further Discussion of Nonequilibrium Effects and Recombination

The buildup and decay of $p_n(t)$ according to Eqs. (11.74) and (11.77), respectively, are illustrated in Fig. W11.10. Band-to-band radiative recombination can be important in highly perfect crystals of direct-bandgap semiconductors such as GaAs but is very unlikely to be important in Si, Ge, and GaP. Indirect-bandgap semiconductors have much longer recombination times (i.e., minority-carrier radiative lifetimes) than direct-bandgap materials as a result of the requirement that a phonon participate in the band-to-band recombination process. Some calculated values for minority-carrier band-to-band radiative lifetimes are given in Table W11.2. These lifetimes have been calculated using the *van Roosbroeck–Shockley relation* and are based on measured optical properties (i.e., the absorption coefficient α and index of refraction n), and on the carrier concentrations of these semiconductors. The van Roosbroeck–Shockley relation expresses a fundamental connection between the absorption and emission spectra of a semiconductor and allows calculation of the band-to-band recombination rate in terms of an integral over photon energy involving α and n. Note that the calculated intrinsic lifetimes span the range from hours for Si to microseconds for InAs.

Measured values of τ_p and τ_n in semiconductors such as Si and GaAs are often much lower than the calculated values because of enhanced recombination due to defects and

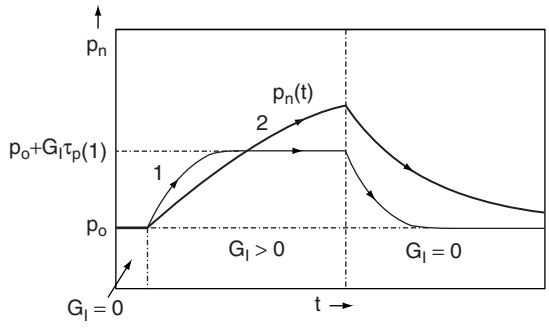

Figure W11.10. Buildup and decay of the minority-carrier hole concentration $p_n(t)$ in an n-type semiconductor under low-level carrier injection for two different minority-carrier lifetimes, with $\tau_p(1) < \tau_p(2)$.

TABLE W11.2 Calculated Minority-Carrier Band-to-Band Radiative Lifetimes at $T = 300$ K

		Lifetime	
Semiconductor	n_i (m^{-3})	Intrinsic[a]	Extrinsic[b]
Si	$\approx 8 \times 10^{15}$	4.6 h	2.5 ms
Ge	$\approx 2 \times 10^{19}$	0.61 s	0.15 ms
InAs	$\approx 2 \times 10^{21}$	15 μs	0.24 μs

[a]Lifetimes are calculated values obtained from R. N. Hall, *Proc. Inst. Electr. Eng.*, **106B**, Suppl. 17, 923 (1959).
[b]The extrinsic lifetimes correspond to carrier concentrations of 10^{23} m^{-3}.

surfaces, to be discussed later. Typical measured minority-carrier lifetimes in extrinsic Si are 1 to 100 μs, whereas in extrinsic GaAs they are 1 to 50 ns.

Minority-carrier recombination times can be on the order of picoseconds in amorphous semiconductors, due to the strong disorder and very high concentrations of defects. Amorphous semiconductors can therefore be very "fast" materials with regard to the speed of their response to external carrier excitation. The recombination times τ_p and τ_n in crystalline semiconductors are typically much longer than the average collision times $\langle\tau\rangle \approx 10^{-13}$ to 10^{-12} s.

Electron–hole recombination in the indirect-bandgap semiconductors Si, Ge, and GaP is much more likely to occur via the participation of defects and surfaces. These two extrinsic recombination mechanisms are discussed next.

Defect-Mediated Recombination. Defects such as metallic impurities and dislocations disturb the periodic potential of the lattice and as a result introduce energy levels deep within the energy gap of the semiconductor, often near midgap, as shown in Fig. 11.22 for Si. The recombination rate will then be enhanced when electrons in the conduction band fall first into the empty defect levels and then fall further into empty levels in the valence band. The defect-mediated recombination rate is proportional to the concentration of defects that have empty energy levels in the energy gap. These defects with deep levels in the gap are often referred to as *recombination centers* or *traps*. The carrier wavefunctions associated with traps are highly localized. While band-to-band recombination can be expected to be the dominant recombination process at high temperatures when n, p, and their product $n\,p$ are all large due to thermal generation, defect-mediated recombination will often be the dominant recombination mechanism at lower temperatures.

The case of defect levels with two charge states, neutral (unoccupied) and negative (occupied by a single electron), has been treated in detail by Hall and by Shockley and Read.[†] Only a brief outline is presented here. The key idea is that empty defect levels near midgap will greatly increase the rate of recombination of electrons and holes due to the fact that such transitions are enhanced when the energy involved is smaller (e.g., $\approx E_g/2$) than the energy E_g for band-to-band recombination.

The possible transitions involving electrons and holes resulting from a defect level at the energy E_t in the gap are presented in Fig. W11.11. Transitions 1 and 2 correspond to the *capture* by the defect of an electron from the conduction band and of a hole from the valence band, respectively, with transitions $1 + 2$ together resulting in the *recombination* of an electron with a hole. Transitions 3 and 4 correspond to the *emission* by the defect of a hole into the valence band and of an electron into the conduction band, respectively, with transitions $3 + 4$ together resulting in the *creation* of an electron–hole pair. These defect levels are also effective in deactivating donors and acceptors in semiconductors through the capture of the donor electrons and acceptor holes.

When the rates of the individual transitions 1 to 4 are considered along with the probabilities of occupation of the levels, the following results are obtained for the steady-state emission probabilities of electrons and holes from the levels [for details, see Grove (1967)].

[†] R. N. Hall, *Phys. Rev.*, **87**, 387 (1952); W. Shockley and W. T. Read, *Phys. Rev.*, **87**, 835 (1952).

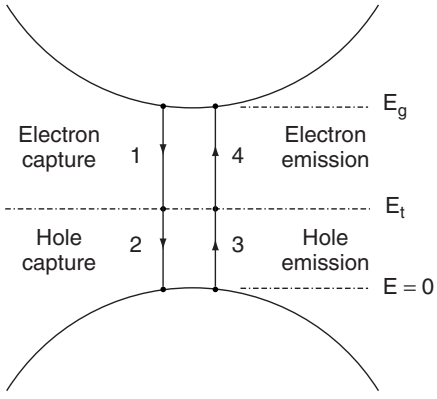

Figure W11.11. Possible transitions involving electrons and holes and resulting from a defect level at the energy E_t in the gap. 1, Capture of an electron; 2, capture of a hole; 3, emission of a hole; 4, emission of an electron.

Absence of Carrier Injection ($G_I = 0$). The total emission rates for holes and electrons, transitions 3 and 4, respectively, will be proportional to the following rates:
Transition 3:

$$\text{hole emission rate} \qquad e_p = v_{p\text{th}}\sigma_p N_v \exp\left(-\frac{E_t}{k_B T}\right) \tag{W11.26}$$

Transition 4:

$$\text{electron emission rate} \qquad e_n = v_{n\text{th}}\sigma_n N_c \exp\left(-\frac{E_g - E_t}{k_B T}\right) \tag{W11.27}$$

Here $v_{p\text{th}} = \sqrt{3k_B T/m_h^*}$ and $v_{n\text{th}} = \sqrt{3k_B T/m_e^*}$ are the thermal velocities, σ_p and σ_n are the capture cross sections ($\approx 10^{-19}$ m^2), and N_v and N_c are the effective densities of states defined in Eq. (11.27), all for holes and electrons, respectively. The rates of transitions 1 to 4 will also be proportional to the concentration of recombination centers N_t and to the probabilities expressed in terms of the Fermi–Dirac distribution function that the final state is empty.

Low-Level Carrier Injection ($G_I > 0$). Net recombination rate due to defects (assuming that $\sigma_n = \sigma_p = \sigma$):

$$U = R - G_T = \frac{\sigma(v_{n\text{th}} v_{p\text{th}})^{1/2} N_t (pn - n_i^2)}{n + p + 2n_i \cosh[(2E_t - E_g)/2k_B T]}. \tag{W11.28}$$

Here the carrier concentrations n and p depend on the injection rate G_I, and N_t is the density of defects whose energy levels lie in the gap at an energy E_t. The recombination rate U has its maximum value for a given G_I when $E_t = E_g/2$ (i.e., when the hyperbolic cosine term in the denominator has its minimum value of unity). Thus recombination centers or traps are most effective when their energy levels are located at midgap.

In an n-type semiconductor the defect energy levels at E_t will ordinarily be occupied by electrons since $n \gg p$. These electrons can be thought of as originating directly from the donor levels. As a result, the effective donor concentration will be reduced to $N_d - N_t$ in an n-type semiconductor containing a concentration N_t of recombination centers. This phenomenon, which can also occur in p-type semiconductors, is known as *majority-carrier removal* and leads to an increase of the resistivity of the semiconductor.

The lifetime for the minority-carrier holes in an n-type semiconductor containing recombination centers and under low-level injection is determined by their rate of capture by these centers. The capture lifetime can be shown to be given by

$$\tau_p = \frac{1}{\sigma_p v_{p\text{th}} N_t}. \tag{W11.29}$$

A similar equation for τ_n is valid for electrons in a p-type semiconductor but with σ_p and $v_{p\text{th}}$ replaced by σ_n and $v_{n\text{th}}$. As soon as a hole is captured by a recombination center in an n-type semiconductor (transition 2 in Fig. W11.11), an electron will be captured essentially immediately by the center (transition 1) due to the high concentration of electrons in the conduction band. Thus the rate-limiting step for electron–hole recombination in a semiconductor containing recombination centers will be the capture by the center of minority carriers. As a result, the minority-carrier lifetime is an important parameter in semiconductor devices.

The minority-carrier lifetimes τ_p or τ_n can be determined experimentally from the decay of the photoconductivity associated with photogenerated carriers. This lifetime is typically much longer than $\langle \tau \rangle$, the average elastic scattering time, which determines the mobility of the charge carriers. The minority-carrier lifetimes τ_p or τ_n can be determined reliably only for low levels of illumination or injection.

Surface Recombination. The recombination rates of electrons and holes can be enhanced at the surface of a semiconductor due to the presence of *surface states* (i.e., electron energy levels lying deep within the energy gap which result from distortions near the surface of the bulk periodic lattice potential). These levels in the energy gap can arise from broken or reconstructed chemical bonds at the surface of the semiconductor, as described in Chapter 19. When surface recombination is important, the electron and hole concentrations will vary spatially and both will be depressed near the surface of the semiconductor due to the enhanced recombination occurring there.

The recombination rate per unit area of surface for holes in an n-type semiconductor under low-level injection is usually taken to be proportional to $(p_n - p_0)$ and of the form

$$R_{\text{surface}} = s_p(p_n - p_0), \tag{W11.30}$$

where s_p is the *surface recombination velocity* and has units of m/s. This velocity can be shown to be given by

$$s_p = \sigma_p v_{p\text{th}} N_{ts}, \tag{W11.31}$$

where N_{ts} is the concentration of recombination centers per unit area at the surface. Typical values of s_p for Si surfaces are ≈ 1 m/s but can be as high as 10^3 m/s. The value of s_p for Si can be reduced to 10^{-2} to 10^{-1} m/s when the Si surface is oxidized. The

removal of these centers by passivation of the surface (e.g., by growing or depositing a surface film of a-SiO_2) is an important step in the fabrication of semiconductor devices (see Chapter W21). The spatial dependence $p(x)$ of the hole concentration near the surface due to recombination can be obtained by solving the continuity equation (11.65) with the incorporation of an appropriate hole diffusion term. In addition, the effect of a space-charge region near the surface on the recombination rate can be determined. For details of these calculations, see Grove (1967).

The total minority-carrier recombination rate in a semiconductor is given by

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}}, \tag{W11.32}$$

where τ_r and τ_{nr} are the *radiative* and *nonradiative* lifetimes, respectively. Another useful expression for $1/\tau_p$ in an n-type semiconductor when all three types of recombination are important is

$$\frac{1}{\tau_p} = k_1 n_0 + \sigma_p v_{p\text{th}} N_t + \frac{\sigma_p v_{p\text{th}} N_{ts}}{d_s}, \tag{W11.33}$$

where Eqs. (11.72), (W11.29), and (W11.31) have been used. Here d_s is the width of the region near the surface where surface recombination is effective.

W11.8 Transistors

The relative suitability of semiconductors for given types of applications is often evaluated on the basis of relevant *figures of merit* (FOMs) which are specific functions of the properties of the semiconductors. For example, the *Johnson* FOM for the power capacity of high-frequency devices is $\text{JM} = (E_c v_{\text{sat}}/\pi)^2$, the *Keyes* FOM for the thermal dissipation capacity of high-frequency devices is $\text{KM} = \kappa\sqrt{v_{\text{sat}}/\epsilon}$, and the *Baliga* FOM for power-loss minimization at high frequencies is $\text{BHFM} = \mu E_c^2$. In these expressions E_c is the critical electric field for breakdown, v_{sat} the saturated carrier drift velocity, κ the thermal conductivity, ϵ the permittivity, and μ the carrier mobility. Figures of merit for various semiconductors, normalized to 1 for Si, are presented in Table W11.3.

TABLE W11.3 Figures of Merit for Various Semiconductors

Semiconductor	E_g (eV)	JM $(E_c v_{\text{sat}}/\pi)^2$	KM $(\kappa\sqrt{v_{\text{sat}}/\epsilon})$	BHFM (μE_c^2)
Si	1.11	1.0	1.0	1.0
InP	1.27	13	0.72	6.6
GaAs	1.42	11	0.45	16
GaP	2.24	37	0.73	38
3C-SiC (β-SiC)	2.3	110	5.8	12
4H-SiC	3.27	410	5.1	34
C (diamond)	5.4	6220	32	850

Source: Data from T. P. Chow and R. Tyagi, *IEEE Trans. Electron Devices*, **41**, 1481 (1994).

The entries in Table W11.3 indicate that the semiconductors listed with wider bandgaps than Si offer in many cases potential order-of-magnitude improvements in performance in high-power, high-frequency electronic applications. This is to be expected since E_c is observed to increase with increasing E_g.

Transistors are semiconductor electronic devices with at least three electrodes, as shown in Fig. W11.12 for the case of an *npn* bipolar junction transistor. The term *bipolar* refers to the fact that both electrons and holes flow within the device in response to applied voltages. Other transistor structures in which only electrons or holes respond to applied voltages include *field-effect transistors* (FETs) such as the junction FET and the *metal–oxide–semiconductor* FET (MOSFET). A wide variety of structures are employed for transistors, depending on the application (e.g., amplification or switching involving high frequency, high power, high speed, etc.). Only a brief outline of transistor action and the most important transistor structures will be presented here.

Bipolar Junction Transistor. A Si bipolar junction transistor consists physically of three distinct regions of Si with different types and levels of doping and separated by *p-n* junctions of opposite polarity in series with each other. These three regions can either be embedded in a single piece of Si or can consist of layers of Si grown epitaxially on a Si substrate. The latter configuration is found in planar device technology, as described in Chapter W21. The two possible types of bipolar junction transistors are *npn* and *pnp*. The physical principles of operation are the same in each type, but with electrons and holes switching roles, and so on. When the *npn* junction transistor is connected to an external circuit as shown in Fig. W11.13, the left-hand side is the *n*-type *emitter*, the central region is the *p*-type *base*, and the right-hand side is the *n*-type *collector*. The built-in electric fields in the *n-p* and *p-n* junctions are in opposite directions, as shown in Fig. W11.12. The electron energy bands at zero bias are shown for the case when all three regions are nondegenerate, but with the emitter more heavily doped (i.e., n^+) than the base or the collector.

The operation of the *npn* transistor consists of forward biasing of the emitter–base *n-p* junction and a stronger reverse biasing of the base–collector *p-n* junction, as shown in Fig. W11.13. The electron energy bands are also shown for the *npn* transistor when biased as described above. Electrons are injected from the emitter into the base where

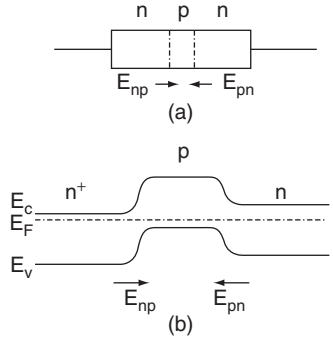

Figure W11.12. An *npn* bipolar junction transistor: (*a*) directions of the built-in electric fields at the two junctions; (*b*) electron energy bands across the transistor at zero bias.

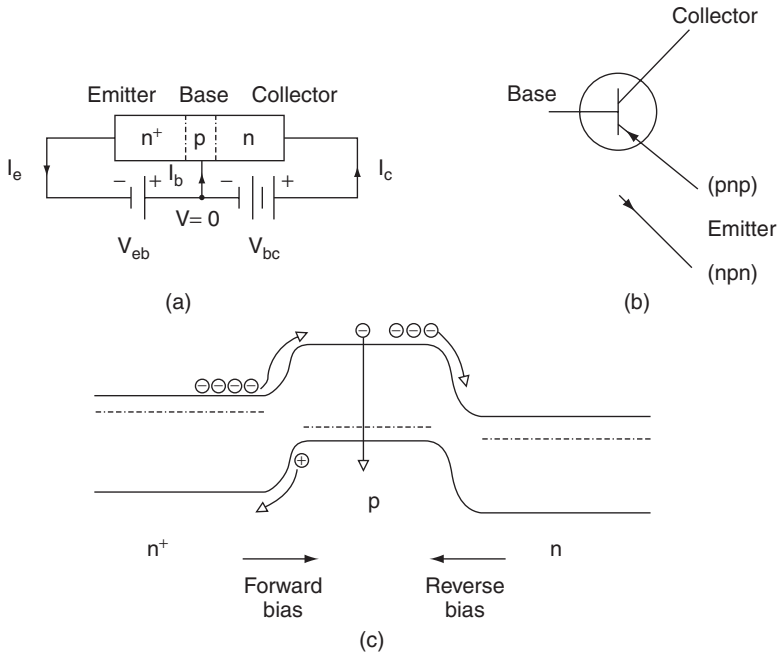

Figure W11.13. Operation of an *npn* transistor. (*a*) The emitter-base *n*-*p* junction is forward biased, while the base–collector *p*-*n* junction is given a stronger reverse bias. The directions of the three resulting currents I_e, I_b, and I_c for the emitter, base, and collector are shown. (*b*) Symbol used for an *npn* junction transistor in a circuit diagram. The arrow on the emitter indicates the direction of the conventional electric current. The direction of this arrow would be reversed for a *pnp* junction transistor. (*c*) Electron energy bands for the biased *npn* transistor.

they diffuse rapidly across the narrow base region whose thickness is less than the electron diffusion length $L_e = \sqrt{D_e \tau_n}$. The electrons that cross the *p*-type base region without recombining with the majority-carrier holes are then swept across the reverse-biased base–collector *n*-*p* junction by its built-in electric field into the collector. The motions of the electrons are shown on the energy-band diagram for the junction, with the smaller hole current from base to emitter also indicated.

The directions of the three resulting currents I_e, I_b, and I_c for the emitter, base, and collector are shown in Fig. W11.13*a*. The emitter current is given by

$$I_e = I_b + I_c = (1 + \beta) I_b, \tag{W11.34}$$

where $\beta = I_c / I_b$ is the *current gain* of the transistor. For alternating currents the small-signal current gain of the transistor is dI_c / dI_b. The ratio of the collector current to the emitter current is given by

$$\frac{I_c}{I_e} = \frac{\beta}{1 + \beta} \lesssim 1. \tag{W11.35}$$

Since most of the electrons injected from the emitter are able to travel across both the base and the base–collector junction into the collector without recombining with

holes, it follows that I_c is almost as large as I_e and that the base current is usually much smaller than either I_e or I_c. Therefore, the current gain defined by Eq. (W11.34) can be $\beta \approx 100$ to 1000. A very thin base with a high diffusion coefficient and a very long lifetime for minority carriers is required for high current gains in bipolar junction transistors. Defect-free Si with its indirect bandgap, and hence very long minority-carrier lifetimes, is clearly an excellent choice for this type of transistor.

A simplified circuit illustrating the use of an *npn* transistor as an amplifier of a small ac voltage $v(t)$ is shown in Fig. W11.14. The dc voltage sources V_{eb} and V_{bc} provide the biasing of the two *p-n* junctions and the source of the input signal $v(t)$ is placed in the base circuit. Kirchhoff's loop rule applied to the emitter–base circuit can be written as

$$V_{bc} + v(t) = V_b - V_e - I_e R_e. \tag{W11.36}$$

Since the emitter–base junction is forward-biased, the voltage drop $V_b - V_e$ across the *n-p* junction will in general be much smaller than the other terms in this equation. Therefore, Eq. (W11.35) can be rewritten with the help of Eq. (W11.36) as

$$I_c = -\frac{\beta}{1+\beta}\frac{V_{bc} + v(t)}{R_e} \approx \frac{V_{bc} + v(t)}{R_e}. \tag{W11.37}$$

The additional output voltage $\Delta V_c(t)$ appearing across the resistor R_c in the collector circuit and due to the input voltage $v(t)$ is equal to $[I_c(v) - I_c(v = 0)]R_c$. The *voltage gain* of this transistor can therefore be shown to be

$$G = \frac{\Delta V_c}{|v|} = \frac{R_c}{R_e}. \tag{W11.38}$$

Thus a small ac voltage in the base circuit can result in a much larger voltage in the collector circuit. Typical voltage gains of junction transistors are ≈ 100. In addition to being used as an amplifier, transistors can also function as switches. In this case, by controlling the base current I_b using the base voltage, the much larger collector current I_c can be switched from a very high value to a very low value.

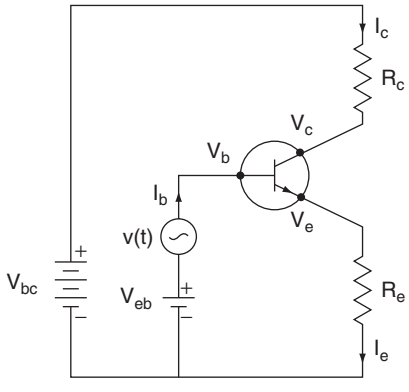

Figure W11.14. Simplified circuit illustrating the use of an *npn* transistor as an amplifier of a small ac voltage $v(t)$. The dc voltage sources V_{bc} and V_{eb} provide the biasing of the two junctions and the source of the input signal $v(t)$ appears in the base circuit.

The intrinsic switching speed of the *npn* junction transistor described here is limited by the time it takes the minority-carrier electrons to travel across the base region of thickness d. Since the distance traveled by a diffusing electron in time t is given by $d = \sqrt{Dt}$, where D is the electron's diffusivity, the electron transit time or *switching time* of the transistor is

$$t_{\text{tr}} \cong \frac{d^2}{D} = \frac{ed^2}{\mu_e k_B T}. \tag{W11.39}$$

Here μ_e is the mobility of the minority-carrier electrons, and the Einstein relation given for D in Eq. (11.81) has been used. To achieve high switching speeds and operation at high frequencies (i.e., a rapid response of the transistor to changes in applied signals), it is important to make the base region as thin as possible and also to fabricate the transistor from a semiconductor with as high a mobility as possible. With $D \approx 5 \times 10^{-3}$ m^2/s for Si and $d \approx 1$ μm, the value of t_{tr} is $\approx 2 \times 10^{-10}$ s, while for GaAs, values of t_{tr} can be as low as 4×10^{-11} s for the same value of d due to its much higher diffusivity $D \approx 0.023$ m^2/s. When the transit time t_{tr} is shorter than the minority-carrier lifetime τ, the minority carriers can travel across the base *ballistically* (i.e., without being scattered). Ballistic propagation of charge carriers can occur in a device as its dimensions shrink in size and, as a result, the usual concepts of average scattering time $\langle\tau\rangle$ and mobility $\mu = e\langle\tau\rangle/m_c^*$ play much less important roles in limiting the drift velocities of the carriers and operation of the device. Under these conditions very high device speeds can be achieved.

Transistor action in a bipolar *npn* junction transistor thus corresponds to the injection of minority-carrier electrons across the forward-biased emitter–base n-p junction into the p-type base region. These electrons diffuse across the base and then drift and diffuse in the accelerating electric field of the reverse-biased base–collector p-n junction, where they then appear as collector current. The base current I_b, which limits the current gain $\beta = I_c/I_b$, corresponds to the back injection of holes from the base to the emitter across the emitter–base n-p junction. The analysis of the operation of a transistor must take into account the exact spatial distributions of dopants in the emitter, base, and collector regions and must include the possible effects of high-level injection.

A type of bipolar transistor that provides better gain and higher-frequency operation than the bipolar junction transistor just discussed is the *heterojunction* bipolar transistor (HBT). In an *npn* HBT the emitter is an n-type semiconductor with a wider bandgap than the base and collector semiconductors. The electron energy-band diagram for an HBT shown in Fig. W11.15 indicates that a potential barrier exists in the valence band which hinders the back injection of holes from the p-type base into the emitter, thereby limiting the current I_b flowing in the base circuit and increasing the current gain $\beta = I_c/I_b$. Due to the very fast, ballistic transport across the base, in contrast to the slower diffusive transport that is ordinarily observed in bipolar junction transistors, HBTs have been developed into the fastest devices of this kind and are used in microwave applications and wireless communication devices.

In one successful HBT structure composed of group III–V semiconductors, InP with $E_g = 1.27$ eV is grown epitaxially on a lattice-matched $In_{0.53}Ga_{0.47}As$ alloy with $E_g \approx 0.8$ eV. Electrons from the InP emitter reach the heavily doped p^+-type $In_{0.53}Ga_{0.47}As$ base region with excess kinetic energy and travel essentially ballistically to the collector. The high cutoff frequency of 165 GHz and average electron

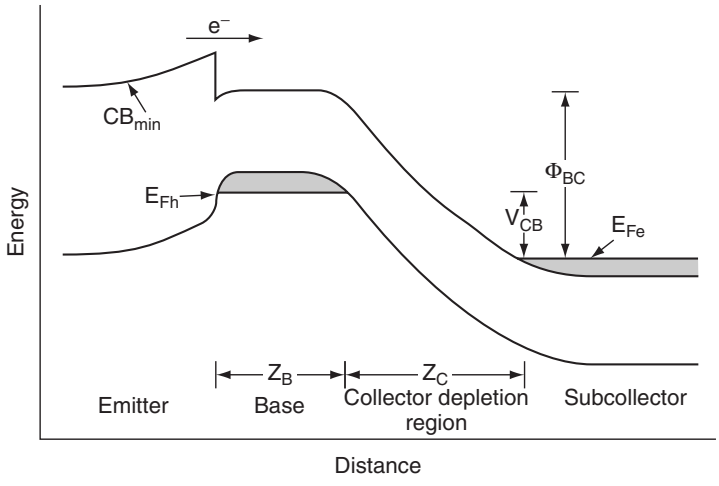

Figure W11.15. Electron energy-band diagram for a heterojunction bipolar transistor (HBT). In the *npn* HBT shown here the emitter has a wider bandgap than the base and collector semiconductors. A potential barrier exists in the valence band that hinders the back injection of holes from the *p*-type base into the emitter. (From A. F. J. Levi et al., *Phys. Today*, Feb. 1990, p. 61. Copyright © 1990 by the American Institute of Physics.)

velocity of 4×10^5 m/s measured at $T = 300$ K in the active region correspond to a total delay of less than 1 ps in the active region between the emitter and the bulk of the collector. The extreme process control ideally required for the fabrication of such HBT devices is indicated by the need to maintain an atomically flat interface between the InP emitter and the base and to restrict the width of the emitter–base doping profile to about 5 nm. Molecular beam epitaxy, described in Chapter W21, is capable of achieving the control needed in the deposition process. Nevertheless, due to the extreme deposition control needed and due to the lack of a reliable native oxide, these group III–V-based devices are unlikely to replace Si technology, despite their outstanding characteristics.

Another material demonstrating impressive performance and high speed in HBT structures is alloys of SiGe grown heteroepitaxially on Si substrates. The lower-bandgap *p*-type SiGe base region in Si–SiGe HBTs allows carriers to travel much faster across the base and thus operation at higher frequencies.

A class of transistors whose operation involves only majority carriers is known as *field-effect transistors* (FETs). These devices are simpler than bipolar junction transistors and correspond in practice to a resistor whose resistance is controlled by an applied voltage and the resulting electric field in the semiconductor. They therefore operate on a completely different physical mechanism than the bipolar junction transistors. Instead of having an emitter, collector, and base, FETs consist of a *source* and a *drain* for electrons and a *gate* that is used either to control or create a conducting channel in the semiconductor. FETs can be viewed as electronic switches that are in either an "on" or an "off" state. As a result, an FET corresponds in a real sense to a single bit (i.e., a binary unit of information). The junction field-effect transistor is discussed briefly next. The metal–oxide–semiconductor FET (MOSFET) is described in Chapter 11.

Junction Field-Effect Transistor. The configuration of a junction FET in a rectangular bar of n-type Si is shown schematically in Fig. W11.16. The two metallic electrodes at the ends of the bar are the source and drain and the conducting channel in the n-type Si between them is controlled by the two p^+-type gates at the center of the bar. The bar of Si acts as a resistor whose resistance R is controlled by the reverse-bias gate voltage V_g. As V_g is increased, the depletion regions at the two reverse-biased p^+-n junctions widen and effectively restrict the cross-sectional area of the path or conducting channel of the majority-carrier electrons as they flow from source to drain. The conductance $G = 1/R$ of the Si bar is therefore controlled by the gate voltage V_g. The junction FET is "on" when the channel is open and conducting and is "off" when it is closed and nonconducting. The speed of the junction FET is controlled by the transit time of the majority carriers through the channel and so is inversely proportional to the gate length.

Current–voltage characteristics of a junction FET are also presented in Fig. W11.16 in the form of the source-to-drain current I_d versus the source-to-drain voltage V_d for a series of gate voltages V_g. For a given V_g, the current I_d is observed to increase linearly and then to saturate. The analysis of the current response of a junction FET is complicated by the fact that the widths of the two depletion regions on opposite sides of the bar are not constant along the channel. As shown in Fig. W11.16, the width will be greater near the drain, where the voltage V_d adds its contribution to the reverse biasing of the two p^+-n junctions. The conducting channel will be "pinched" (i.e., will decrease in cross-sectional area to a small value) when the two depletion regions are very close to each other near the drain electrode. The current I_d does not in fact go to zero due to this "pinching" effect but instead, saturates, as observed. As the channel shrinks in cross section, the electric field lines are squeezed into a smaller area. As a result, the electric field in the channel increases and current continues to flow. In this case, Ohm's law will no longer be valid when the electric field reaches a value where the mobility of the majority carriers starts to decrease due to inelastic scattering effects associated with "hot" carriers, as described in the discussion of high-field effects in Section 11.7.

The rapid increase in drain current I_d that is observed to occur in Fig. W11.16 as either V_g and/or V_d increase in magnitude is just the junction breakdown which occurs when the p^+-n junctions are strongly reverse-biased. It can be seen that both V_g and V_d contribute to the breakdown of the junction FET.

In the junction FET the gate voltage effectively controls the resistance R or conductance G of the p-type Si region and so controls the flow of current through the device. The *transconductance* of the transistor is defined by

$$g_m = \frac{\partial I_d}{\partial V_g}. \tag{W11.40}$$

Here g_m expresses the degree of amplification and control of the source-to-drain current I_d by the gate voltage V_g and is one of the most important characteristics of the transistor.

Other Types of Transistors. An intrinsic problem in semiconductor devices is that the doping procedure which provides the majority carriers can also lead to a decrease in the carrier mobility at high doping levels, as illustrated in Fig. 11.15. This

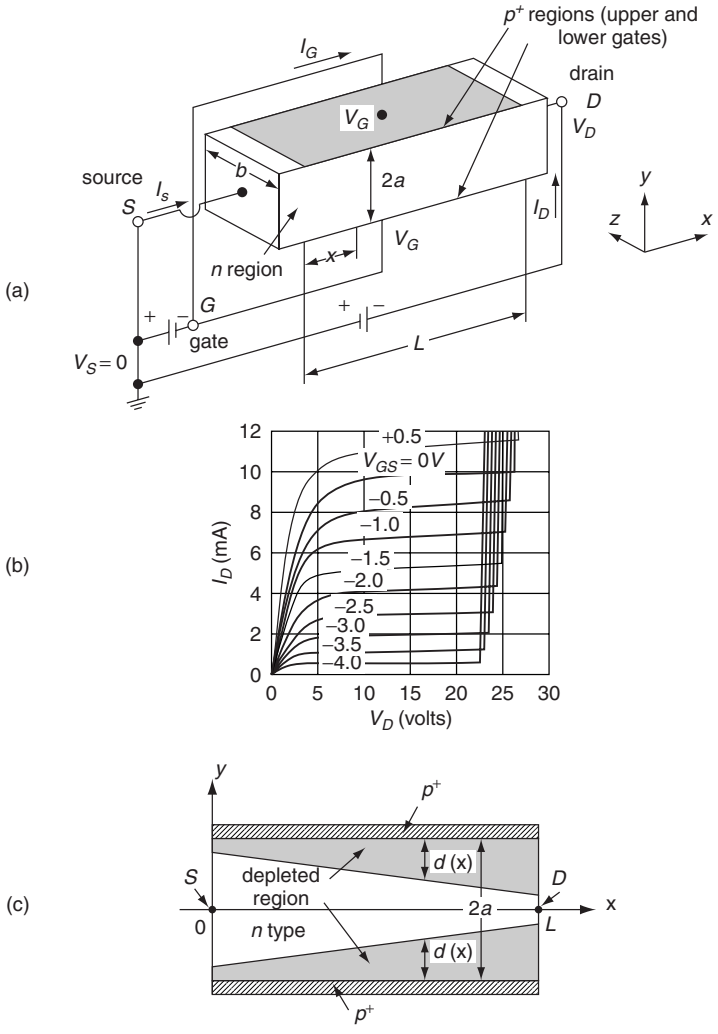

Figure W11.16. Properties of a junction FET. (*a*) Configuration of a junction FET in a rectangular bar of n-type Si. The two metallic electrodes at the ends of the bar are the source and drain, and the conducting channel between them is controlled by the p-type gates at the center of the bar. (*b*) Current–voltage characteristics of the 2N3278 junction FET in the form of the source-to-drain current I_d versus the source-to-drain voltage V_d for a series of gate voltages V_g. (*c*) The width of the depletion regions is greater near the drain electrode, where the drain voltage V_d adds its contribution to the reverse biasing of the two p^+-n junctions. (From B. Sapoval and C. Hermann, *Physics of Semiconductors*, Springer-Verlag, New York, 1993.)

decrease occurs because the ionized donor and acceptor ions act as charged scattering centers, and this additional scattering leads to a decrease in the average scattering or momentum relaxation time $\langle\tau\rangle$. A procedure that can minimize this effect makes use of heterostructures or superlattices and is known as *modulation doping*. Modulation doping involves introduction of the dopant atoms into a wider-bandgap layer (e.g., $Al_xGa_{1-x}As$ with E_g up to 2.2 eV) and the subsequent transfer of the carriers across

the interface to lower-lying energy levels in an adjacent layer with a narrower bandgap (e.g., GaAs with $E_g = 1.42$ eV). The carriers are thereby spatially separated from the charged scattering centers, as shown in Fig. W11.17. Much higher carrier mobilities, up to 150 $m^2/V{\cdot}s$ in GaAs at $T \approx 4.2$ K, can be achieved using modulation doping than are ordinarily attainable using normal doping procedures. Very fast electronic devices which can be fabricated using modulation doping and in which the charge carriers move ballistically include MODFETs (i.e., *modulation-doped* FETs) and HEMTs (i.e., *high-electron-mobility transistors*).

In applications related to information technology, such as displays and photocopiers, where larger, rather than smaller, physical dimensions are needed, it is advantageous to be able to deposit large areas of semiconducting thin films which can then be processed into devices such as *thin-film transistors* (i.e., TFTs). A semiconducting material that is useful for many of these applications is hydrogenated amorphous Si, a-Si:H, that can be deposited over large areas onto a wide variety of substrates via plasma deposition techniques and that can be successfully doped *n*- and *p*-type during the deposition process, as discussed in Chapter W21.

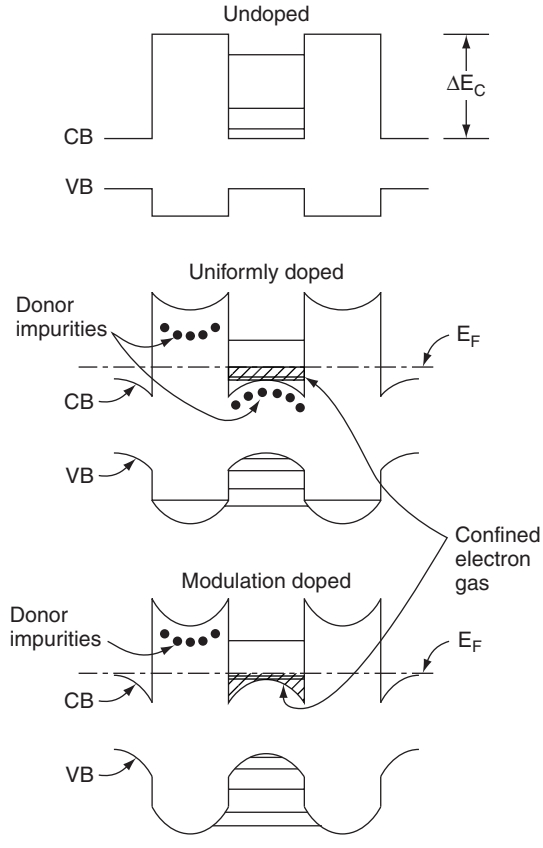

Figure W11.17. Modulation doping in GaAs-$Al_xGa_{1-x}As$ superlattices. The carriers are spatially separated from the charged scattering centers associated with the dopant impurity ions. (From R. Dingle et al., *Appl. Phys. Lett.*, **33**, 665 (1978). Copyright © 1978 by the American Institute of Physics.)

Although a-Si:H is inferior to c-Si in its electronic properties (e.g., a-Si:H possesses an electron mobility $\mu_e \approx 10^{-4}$ m^2/V·s compared to $\mu_e = 0.19$ m^2/V·s for c-Si), these properties are sufficient for applications in field-effect TFTs (or thin-film FETs), which act as the switches which, for example, control the state of the pixels in large-area liquid-crystal displays. A common configuration of an a-Si:H field-effect TFT is shown in Fig. W11.18, along with its source-to-drain current I_d versus gate voltage V_g transfer characteristic, which is similar to that of a conventional MOSFET. At the transition from the "on" to the "off" state, the source-to-drain resistance R_d increases by about six orders of magnitude. Other large-area applications of a-Si:H films in photovoltaic solar cells are discussed in Section W11.10. Polycrystalline Si has a higher mobility than a-Si:H and thus can operate at higher frequencies in TFTs.

Another material with significant potential for electronic device applications is SiC. SiC is considered to be a nearly ideal semiconductor for high-power, high-frequency transistors because of its high breakdown field of 3.8×10^8 V/m, high saturated electron drift velocity of 2×10^5 m/s, and high thermal conductivity of 490 W/m·K. Its wide bandgaps of 3.0 and 3.2 eV in the hexagonal 6H– and 4H–SiC forms, respectively, allow SiC FETs to provide high radio-frequency (RF) output power at high temperatures. In addition, SiC has the important advantage over most group III–V and II–VI semiconductors in that its native oxide is SiO_2, the same oxide that provides passivation for Si.

A SiC metal–semiconductor field-effect transistor (MESFET) is shown schematically in Fig. W11.19. The gate configuration in the MESFET consists of a rectifying metal–semiconductor Schottky barrier at the surface of a doped epitaxial layer of SiC that is grown on either a high-resistivity substrate or a lightly doped substrate of the opposite conductivity type. When used in RF applications, an RF voltage that is

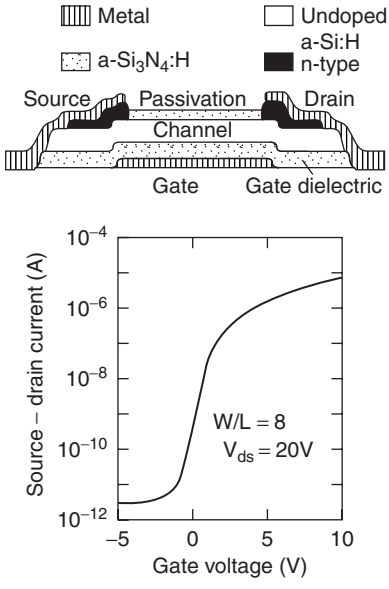

Figure W11.18. Common configuration of an a-Si:H field-effect TFT, along with its source-to-drain current I_d versus gate voltage V_g transfer characteristic. (From R. A. Street, *Mater. Res. Soc. Bull.*, **17**(11), 71 (1992).)

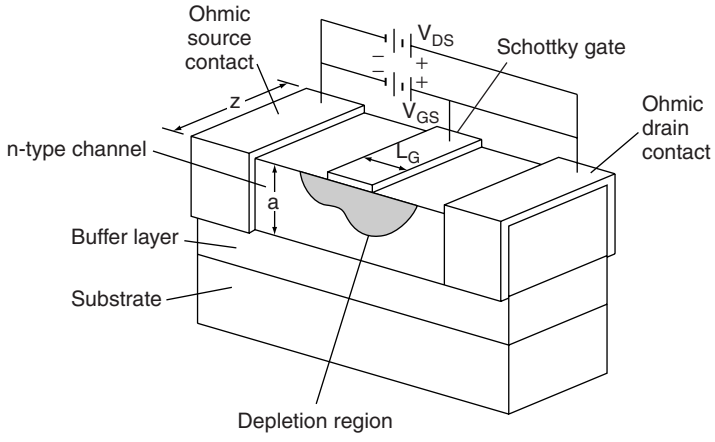

Figure W11.19. SiC metal–semiconductor field-effect transistor (MESFET). The gate configuration in the MESFET consists of a rectifying metal–semiconductor Schottky barrier at the surface of a doped, epitaxial layer of SiC. (From K. Moore et al., *Mater. Res. Soc. Bull.*, **23**(3), 51 (1997).)

superimposed on the dc gate voltage V_g modulates the source-to-drain current in the conducting channel, thereby providing RF gain. The SiC MESFET can provide significantly higher operating frequencies and higher output power densities than either Si RF power FETs or GaAs MESFETs.

W11.9 Quantum Hall Effect

The study of the electrical properties of the two-dimensional electron gas (2DEG) has yielded some remarkable and unexpected results. In the experiment[†] that led to the discovery of the quantum Hall effect, a high-mobility silicon MOSFET was used to create the 2DEG, and its electrical properties were studied at low temperatures, $T \approx 1.5$ K, and high magnetic fields, $B \approx 15$ T. More recent studies utilize the GaAs–AlGaAs heterostructure to create the 2DEG. Consider the geometry shown in Fig. W11.20, in which a magnetic induction **B** is imposed perpendicular to the 2DEG, which lies in the xy plane. The longitudinal resistivity, $\rho_{xx} = (V_L/I)(w/L)$, and Hall resistivity, $\rho_{xy} = V_H/I$, are measured in two dimensions, where w is the width and L is the length of the 2DEG, respectively. The electrons are in the ground quantum state of a potential well in the z direction, perpendicular to the plane of motion. The spatial extent of the wavefunction in the z direction is small compared with w and L.

Prior to the experiments, the a priori expectations for the behavior of these resistivities as a function of **B** were simple. If N is the number of electrons per unit area in the 2DEG, then, in analogy with the discussion in Section 7.3, it was expected that $\rho_{xy} = B/Ne$ (i.e., the Hall resistivity should be proportional to the magnetic field and inversely proportional to the number of electrons per unit area, N). The naive Drude expectation for ρ_{xx} was that it shows no magnetoresistance. However, Shubnikov and

[†] K. von Klitzing, G. Dorda, and M. Pepper, *Phys. Rev. Lett.*, **45**, 494 (1980).

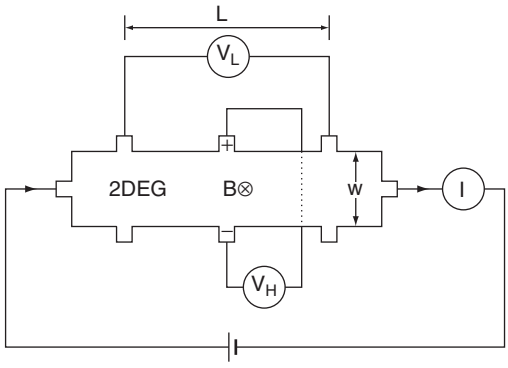

Figure W11.20. Geometry of the measurement of the quantum Hall effect for the two-dimensional electron gas.

de Haas† had found oscillatory structure in the magnetoresistivity of three-dimensional conductors as a function of $1/B$. The period of this structure is given by a formula derived by Onsager, $\Delta(1/B) = 2\pi e/\hbar A_F$, where A_F is the area of the equatorial plane of the Fermi sphere in k space with the magnetic field along the polar axis. The physical origin involves Landau levels (discussed in Appendix W11A) crossing the Fermi level as the magnetic field is varied. Similar oscillations were expected in two-dimensional conductors. In place of a Fermi sphere there would be a Fermi circle in the $(k_x k_y)$ plane.

A sketch of the experimental data for the integer quantum Hall effect (IQHE) is presented in Fig. W11.21. A steplike structure with exceedingly flat plateaus is found

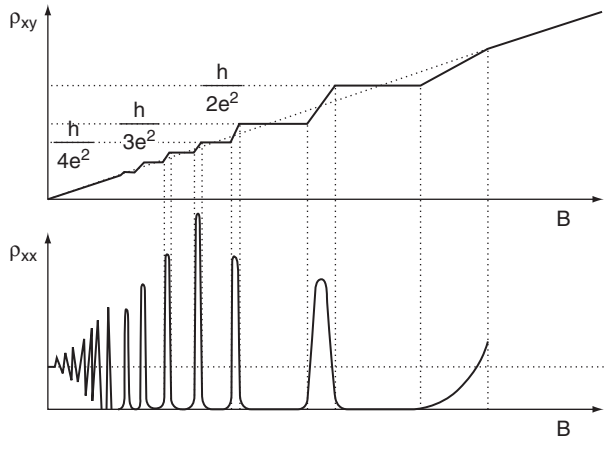

Figure W11.21. Experimental results for the Hall resistivity ρ_{xy} and magnetoresistivity ρ_{xx} for the two-dimensional electron gas. (Reprinted with permission of H. Iken. Adapted from B. I. Halperin, The quantized Hall effect, *Sci. Am.*, Apr., 1986, p 52.)

† W. J. de Haas, J. W. Blom, and L. W. Schubnikow, *Physica* **2**, 907 (1935).

for ρ_{xy} as a function of B. The flatness is better than 1 part in 10^7. The resistivity for the nth step is $\rho_{xy} = h/ne^2 = 25{,}812.8056\ \Omega/n$, where $n = 1, 2, 3, \ldots$, and is now used as the standard of resistance. In addition, ρ_{xx} consists of a series of spikelike features as a function of B. The location of the spikes coincides with the places where the transitions between the plateaus occur. In between the spikes it is found that the longitudinal resistivity vanishes.

In the absence of a magnetic field, the density of states (number of states per unit energy per unit area) for a free-electron gas in two dimensions is predicted to be constant (see Table 11.5). Thus, for a parabolic conduction band,

$$\rho(E) = \frac{1}{A}\sum_{\mathbf{k},m_s} \delta(E_k - E) = \int \frac{2d^2k}{(2\pi)^2}\delta\left(\frac{\hbar^2k^2}{2m_e^*} - E\right) = \frac{m_e^*}{\pi\hbar^2}\Theta(E), \tag{W11.41}$$

where m_e^* is the effective mass of the electron and $\Theta(E)$ is the unit step function. The Fermi energy is obtained by evaluating

$$N = \int dE\rho(E)\Theta(E_F - E) = \frac{m_e^*E_F}{\pi\hbar^2}. \tag{W11.42}$$

The radius of the Fermi circle is given by $k_F = \sqrt{2\pi N}$.

In the presence of a magnetic field, the density of states is radically transformed. The spectrum degenerates into a series of equally spaced discrete lines called *Landau levels*. The states are labeled by three quantum numbers: a nonnegative integer n, a continuous variable k_x, and a spin-projection quantum number m_s. Details are presented in Appendix W11A. The energies of the Landau levels are given by the formula $E_{nk_xm_s} = (n + \frac{1}{2})\hbar\omega_c + g\mu_B Bm_s$, where $\omega_c = eB/m_e^*$ is the cyclotron frequency of the electron in the magnetic field. Note that the energy does not depend on k_x. The energy formula includes the Zeeman splitting of the spin states. The density of states becomes

$$\rho(E) = \frac{1}{A}\sum_{nk_xm_s} \delta(E - E_{nk_xm_s}) = D\sum_{m_s}\sum_{n=0}^{\infty}\delta\left(E - \left(n + \frac{1}{2}\right)\hbar\omega_c - g\mu_B Bm_s\right). \tag{W11.43}$$

A sketch of the density of states is presented in Fig. W11.22. Figure W11.22*a* corresponds to the case where there is no magnetic field. Figure W11.22*b* shows the formation of Landau levels when the magnetic field is introduced but when there is no disorder. The degeneracy per unit area of each Landau level, D, is readily evaluated by taking the limit $\omega_c \to 0$ and converting the right-hand sum to an integral over n. The result may then be compared with Eq. (W11.41) to give $D = m_e^*\omega_c/2\pi\hbar = eB/h$. The filling factor is defined by $\nu = N/D$. For $\nu = 1$ the first Landau level (with $n = 0$ and $m_s = -\frac{1}{2}$) is filled, for $\nu = 2$ the second Landau level (with $n = 0$ and $m_s = \frac{1}{2}$) is also filled, and so on for higher values of n. Each plateau in ρ_{xy} is found to be associated with an integer value of ν (i.e., $\rho_{xy} = h/\nu e^2$). The filling of the Landau levels may be controlled by either varying B or N. The areal density N may be changed by varying the gate voltage in a MOSFET or by applying the appropriate voltages to a heterostructure.

The boundaries of the 2DEG in a magnetic field act as one-dimensional conductors. In the interior of a two-dimensional conductor the electrons are believed to be localized

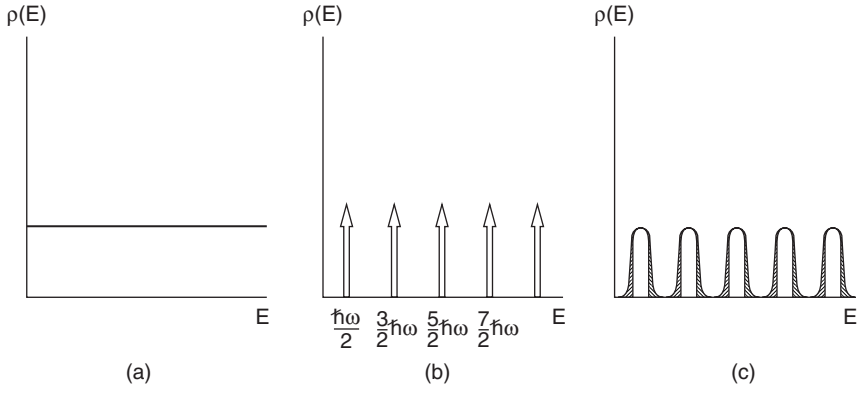

Figure W11.22. Density of states for a two-dimensional electron gas: (*a*) in the absence of a magnetic field; (*b*) in the presence of a magnetic field, but with no disorder; (*c*) in the presence of a magnetic field and with disorder. The smaller Zeeman spin splitting of the Landau levels is not shown.

by scattering from the random impurities. On the edges, however, the electrons collide with the confining potential walls and the cyclotron orbits consist of a series of circular arcs that circumscribe the 2DEG. Electrons in such edge states are not backscattered and carry current. Recalling the mechanism responsible for weak localization discussed in Section W7.5, it is observed that the edge states cannot become localized. As a result, edge states are delocalized over the entire circumference of the 2DEG. Phase coherence is maintained around the circumference. If one were to follow an electron once around the 2DEG, Eq. (W11A. 5) implies that its wavefunction accumulates a phase shift of amount

$$\delta\phi = \frac{e}{\hbar}\oint \mathbf{A}\cdot d\mathbf{l} = \frac{e}{\hbar}\int \mathbf{B}\cdot\hat{n}\, dS = \frac{e\Phi}{\hbar}, \tag{W11.44}$$

where $\mathbf{A}$ is the vector potential, dS an area element, and Φ the magnetic flux through the sample. Uniqueness of the wavefunction requires that $\delta\phi = 2\pi N_F$, where N_F is an integer. Thus $\Phi = N_F\Phi_0$, where $\Phi_0 = h/e = 4.1357 \times 10^{-15}$ Wb is the quantum of flux. Each Landau level contributes an edge state that circumscribes the 2DEG. Eventually, as the Hall electric field builds up due to charge accumulation on the edges, the cyclotron orbits of the edge states will straighten out into linear trajectories parallel to the edges.

States with noninteger ν are compressible. If N/D is not an integer, one may imagine compressing the electrons into a smaller area A' so that N' will be the new electron density in that area. Because of the high degeneracy of the Landau level, this may be done without a cost in energy until N'/D reaches the next-larger integer value. To compress the electron gas further requires populating the next-higher Landau level, which involves elevating the electronic energies. Therefore, states with integer ν are incompressible.

The zero longitudinal resistivity of the 2DEG for integer ν may be a consequence of the incompressibility of the filled Landau levels. If all the electrons flow as an incompressible fluid across the 2DEG sheet, there is considerable inertia associated with this flow. Furthermore, the fluid interacts simultaneously with many scattering

centers, some attractive and some repulsive. Consequently, as the fluid moves along, there is no net change in the potential energy of the system and no net scattering.

It is worth examining the condition $\nu = N/D$ in light of the condition for quantized flux. Suppose that ν is an integer. Let there be a total of N_e conduction electrons in the 2DEG. Then

$$\nu = \frac{N}{D} = \frac{N_e h}{e\Phi} = \frac{N_e}{N_F}. \tag{W11.45}$$

Thus associated with each flux quantum are ν electrons.

For an electron to be able to pass through the sheet without being deflected by the magnetic field, the magnetic force must be equal in magnitude, but opposite in direction, to the Hall electric force (i.e., $evB = eE_H$). The Hall electric field ($E_H = V_H/w$) is due to charge that accumulates along the edges of the sample. Thus

$$V_H = wvB = \frac{v}{L}\Phi = \frac{v}{L}N_F\Phi_0 = \frac{N_F v h}{eL}. \tag{W11.46}$$

The current carried by the 2DEG is given by

$$I = Nvew = \frac{N_e v e}{L}. \tag{W11.47}$$

The Hall resistivity is therefore given by

$$\rho_{xy} = \frac{V_H}{I} = \frac{N_F h}{N_e e^2} = \frac{h}{\nu e^2}. \tag{W11.48}$$

It is believed that the plateaus in the Hall resistivity coincide with regions where the Fermi level resides in localized states between the Landau levels. The localized states are a consequence of disorder. When there is disorder present, the density of states no longer consists of a series of uniformly spaced delta functions. Rather, each delta function is spread out into a broadened peak due to the local potential fluctuations set up by the scattering centers. The states associated with the region near the centers of the peaks are extended throughout the 2DEG, while those in the wings of the peak are localized. This is illustrated in Fig. W11.22*c*, where the shaded regions correspond to localized states and the unshaded regions correspond to extended states. The area under each peak is D. As the magnetic field is varied and ω_c changes, the Landau levels move relative to the fixed Fermi level. When the Fermi level resides in the localized states these states do not contribute to the conductivity. As long as no new extended states are added while the localized states sweep past the Fermi level, ρ_{xy} remains constant. When B increases and E_F enters a band of extended states, a charge transfer occurs across the 2DEG which causes ρ_{xy} to increase. Laughlin[†] has presented a general argument based on gauge transformations showing how this happens.

The conductivity tensor is the inverse of the resistivity tensor. Thus, in the plateau regions the Hall conductivity is $\sigma_{xy} = -\rho_{xy}/(\rho_{xx}\rho_{yy} - \rho_{xy}\rho_{yx}) \to 1/\rho_{yx}$, since $\rho_{xx} = 0$. Thus $|\sigma_{xy}| = \nu e^2/h$. This is expected from the Landauer theory of conduction. The

[†] R. B. Laughlin, *Phys. Rev. B*, **23**, 5632 (1981).

current is carried by the edge states, with each Landau level contributing an edge state. Note that both edges of the 2DEG can conduct through each edge state.

Further investigations of the quantum Hall effect at higher magnetic fields for the lowest Landau level[†] have revealed additional plateaus in the Hall resistivity at fractional values of ν. The phenomenon is called the *fractional quantum Hall effect* (FQHE). If ν is expressed as the rational fraction $\nu = p/q$, only odd values of q are found. For the case $p = 1$, this is equivalent to saying that each electron is associated with an odd number, q, of flux quanta.

The system of electrons that exhibits the FQHE is highly correlated, meaning that the size of the electron–electron interaction is larger than the kinetic energy of the electron. Instead of describing the physics in terms of bare electrons, one introduces quasiparticles. One such description involves the use of what are called *composite fermions*.[‡] In this picture each electron is described as a charged particle attached to a flux quantum. It may further become attached to an additional even number of flux quanta. In such a description the composite fermion may be shown to obey Fermi–Dirac statistics. The FQHE is then obtained as an IQHE for the composite fermions.

In another description of the quasiparticles[§] it is useful to think of the fractionization of charge. For example, in the case where $\nu = \frac{1}{3}$, the quasiparticles are regarded as having charge $e^* = e/3$. This does not mean that the actual physical charge of the electron has been subdivided but that the wavefunction of a physical electron is such that the electron is as likely to be found in three different positions. These positions may, however, independently undergo dynamical evolution and may even change abruptly due to tunneling. Experiments on quantum shot noise[¶] have, in fact, shown that the current in the FQHE is carried by fractional charges $e/3$. More recent shot-noise experiments have shown that the $\nu = \frac{1}{5}$ FQHE involves carriers with charge $e/5$.

W11.10 Photovoltaic Solar Cells

The *photovoltaic effect* in a semiconductor can occur when light with energy $\hbar\omega > E_g$ is incident in or near the depletion region of a p-n junction. The electron–hole pairs that are generated within a diffusion length of the depletion region can be separated spatially and accelerated by the electric field in the depletion region. They can thus contribute to the drift current through the junction. This additional photo-induced drift current (i.e., *photocurrent*) of electrons and holes upsets the balance between the drift and diffusion currents that exists for $V_{\text{ext}} = 0$ when the junction is in the dark. The photocurrent flows from the n- to the p-type side of the junction (i.e., it has the same direction as the net current that flows through the junction under reverse-bias conditions when $V_{ext} < 0$). The total current density that flows through an illuminated junction when a photo-induced voltage (i.e., a *photovoltage*) V is present is given by

$$J(V, G_I) = J(G_I) - J(V) = J(G_I) - J_s[\exp(eV/k_BT) - 1], \tag{W11.49}$$

[†] D. C. Tsui, H. L. Stormer, and A. C. Gossard, *Phys. Rev. Lett.*, **48**, 1559 (1982).
[‡] J. K. Jain, *Phys. Rev. Lett.*, **63**, 199 (1989).
[§] R. B. Laughlin, *Phys. Rev. Lett.*, **50**, 1395 (1983).
[¶] R. de Picciotto et al., *Nature*, **389**, 162 (1997).

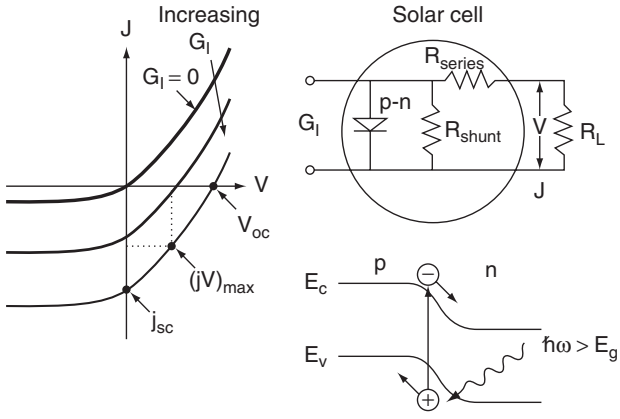

Figure W11.23. Predicted current–voltage characteristics for a photovoltaic solar cell in the form of a *p-n* junction, both in the dark ($G_I = 0$) and illuminated ($G_I > 0$), shown schematically when the solar cell is connected to an external circuit. The generation rate of photo-excited electron–hole pairs is given by G_I. Also shown are the processes giving rise to the photo-induced current.

where G_I is the rate of generation or injection of carriers due to the incident light and $J(V)$ is the voltage-dependent junction current given by Eq. (11.103).

Current–voltage characteristics predicted by Eq. (W11.49) are shown schematically in Fig. W11.23 for a *p-n* junction connected to an external circuit, both in the dark ($G_I = 0$) and when illuminated ($G_I > 0$). Also shown are the equivalent circuit of the *solar cell* comprised of the *p-n* junction with series and shunt resistances and, in addition, the processes giving rise to the photo-induced current. The useful current that can be derived from the photovoltaic effect and which can deliver electrical power to an external circuit corresponds to the branch of the *J-V* curve in the fourth quadrant where $V > 0$ and $J < 0$. The voltage V_{oc} is the *open-circuit voltage* that appears across the *p-n* junction when $J(G_I, V) = 0$ (i.e., when no current flows). This voltage can be obtained from Eq. (W11.49) and is given by

$$V_{oc} = \frac{k_B T}{e} \ln\left[\frac{J(G_I)}{J_s} + 1\right]. \tag{W11.50}$$

The *short-circuit current density* at $V = 0$ is $J_{sc} = J(G_I)$. Note that V_{oc} corresponds to a forward-bias voltage and has a maximum value for a given semiconductor equal to the built-in voltage V_B of the *p-n* junction, as defined in Eq. (11.94). The magnitude of the short-circuit current density J_{sc} will be proportional to the integrated flux of absorbed photons and to the effective quantum efficiency η_{eff} of the device (i.e., the fraction of absorbed photons that generate electron–holes pairs, which are then collected and contribute to the photocurrent). Note that V_{oc} and J_{sc} change in opposite ways as the energy gap of the semiconductor is varied. The voltage V_{oc} increases with increasing E_g, while J_{sc}, being proportional to number of carriers excited across the bandgap, decreases with increasing E_g.

The optimal operating point of the *p-n* junction solar cell is in the fourth quadrant, as shown. At this point the product JV has its maximum value $(JV)_{max}$ (i.e., the

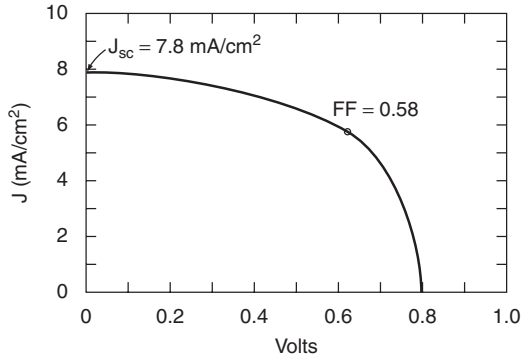

Figure W11.24. Typical J–V curve for an a-Si:H Schottky-barrier solar cell under illumination of 650 W/m^2. (From M. H. Brodsky, ed., *Amorphous Semiconductors*, 2nd ed., Springer-Verlag, New York, 1985.)

inscribed rectangle has the maximum possible area). The *fill factor* (FF) of the solar cell is defined to be FF $= (JV)_{\max}/J_{\rm sc}V_{\rm oc}$, and a value as close to 1 as possible is the goal. For a typical crystalline Si solar cell it is found that $V_{\rm oc} \approx 0.58V$, $J_{\rm sc} \approx 350$ A/m^2, and FF ≈ 0.8. A typical J-V curve for an a-Si:H Schottky barrier solar cell under illumination of 650 W/m^2 is shown in Fig. W11.24.

The efficiency of a photovoltaic solar cell in converting the incident spectrum of solar radiation at Earth's surface to useful electrical energy depends on a variety of factors, one of the most important of which is the energy gap E_g of the semiconductor. There are, however, two conflicting requirements with regard to the choice of E_g. To absorb as much of the incident light as possible, E_g should be small. In this case essentially all of the solar spectrum with $\hbar\omega > E_g$ could be absorbed, depending on the reflectance R of the front surface of the cell, and so on. Most of the photo-generated electrons and holes would, however, be excited deep within their respective energy bands with considerable kinetic energies (i.e., their energies relative to the appropriate band edge would be a significant fraction of $\hbar\omega$). As a result, these charge carriers would lose most of their kinetic energy nonradiatively via the process of phonon emission as they relax to the nearest band edge. Only the relatively small fraction $E_g/\hbar\omega$ of each photon's energy would be available to provide useful electrical energy to an external circuit.

An alternative solution would involve the use of a semiconductor with a high energy gap so that a greater fraction of the energy of each absorbed photon could be converted to useful electrical energy. Although this is true, the obvious drawback is that many fewer photons would be absorbed and thus available to contribute to the photo-induced current. From a consideration of both effects, it has been calculated that the optimum energy gap for collecting energy at Earth's surface in a *single-color solar cell* (i.e., a solar cell fabricated from a single semiconductor) would be $E_g \approx 1.4$ eV, which is close to the energy gap of GaAs. In this case the maximum possible efficiency of the solar cell would be $\approx 26\%$.

For crystalline Si with $E_g = 1.11$ eV, the maximum possible efficiency is $\approx 20\%$. It has been possible so far to fabricate Si solar cells with efficiencies of $\approx 15\%$. An alternative to crystalline Si is a-Si:H since a-Si:H films with thicknesses of 1 μm are sufficient to absorb most of the solar spectrum. Even though its energy gap $E_g \approx 1.8$ eV is relatively high, a-Si:H is a direct-bandgap semiconductor due to the breakdown of

selection rules involving conservation of wave vector **k** for optical absorption. As a result, a-Si:H has higher optical absorption than c-Si (see Fig. W11.7b). In addition, a-Si:H is much less expensive to produce than c-Si and so has found applications in the solar cells that provide power for electronic calculators and other electronic equipment. Other materials that are candidates for use in terrestrial solar cells include the chalcopyrite semiconductor $CuIn_{1-x}Ga_xSe_2$ with $E_g = 1.17$ eV from which cells with $\approx 17\%$ efficiency have been fabricated.

A possible solution to the problem associated with the choice of energy gap is to fabricate *two-color* or *multi color solar cells*, also known as *tandem solar cells*. In a two-color cell two *p-n* junctions fabricated from semiconductors with energy gaps E_{g1} and $E_{g2} > E_{g1}$ are placed in the same structure, with the semiconductor with the higher gap E_{g2} in front. In this way more of the energy of the incident photons with $\hbar\omega > E_{g2}$ would be collected by the front cell, while the back cell would collect energy from the photons with $E_{g2} > \hbar\omega > E_{g1}$ which had passed through the front cell. Although higher conversion efficiencies can be achieved in this way, the higher costs of fabricating such cells must also be taken into account. The cost per watt of output power of a photovoltaic solar cell will ultimately determine its economic feasibility.

W11.11 Thermoelectric Devices

The most common devices based on thermoelectric effects are *thermocouples*, which are used for measuring temperature differences. These are typically fabricated from metals rather than semiconductors. Thermoelectric effects in semiconductors have important applications in power generation and in refrigeration, due to the observed magnitude of the thermoelectric power S in semiconductors, ≈ 1 mV/K, which is 100 to 1000 times greater than the thermoelectric powers typically observed in metals. Thermoelectric energy conversion and cooling are achieved via the Peltier effect described in Section W11.4. An important advantage of these thermoelectric power sources and refrigerators fabricated from semiconductors is that they have no moving parts and so can have very long operating lifetimes.

Schematic diagrams of a thermoelectric power source or generator and a thermoelectric refrigerator are shown in Fig. W11.25. In the thermoelectric generator two semiconductors, one n-type and the other p-type, each carry a heat flux from a heat source at a temperature T_h to a heat sink at a temperature T_c; see Fig. W11.4 for a schematic presentation of the processes involved. In practice, many such pairs of semiconductors are used in parallel in each stage of the device. When a complete electrical circuit is formed, a net current density $J = I/A$ of majority carriers travels from the hot to the cold end of each semiconductor.

The net heat input into the semiconductors from the heat source is given by

$$\frac{dQ}{dt} = IT_h(S_p - S_n) + K\,\Delta T - \frac{I^2R}{2}, \tag{W11.51}$$

where the combined thermal conductance K and electrical resistance R of the pair of semiconductors are defined, respectively, by

$$K = \left[\left(\frac{\kappa A}{L}\right)_n + \left(\frac{\kappa A}{L}\right)_p\right],$$

$$R = \left[\left(\frac{\rho L}{A}\right)_n + \left(\frac{\rho L}{A}\right)_p\right]. \tag{W11.52}$$

Here κ is the thermal conductivity, ρ the electrical resistivity, and A and L the cross section and length of each semiconductor, respectively.[†] The semiconductors are thermally insulated and therefore lose no heat through their sides to the surroundings. The three terms on the right-hand side of Eq. (W11.51) represent the rates of heat flow either out of or into the heat source via the following mechanisms:

1. $IT_h(S_p - S_n) = I(\Pi_p - \Pi_n)$. This term represents the rate at which heat is removed from the heat source at temperature T_h via the Peltier effect at the junctions between each semiconductor and the metallic contact. The thermopower S_m of the metallic contacts cancels out of this term, and in any case, S_m is typically much smaller than either S_p or S_n. Note that both components of the Peltier heat are positive since "hot" electrons and "hot" holes enter the n- and p-type semiconductors, respectively, from the metallic contacts in order to replace the "hot" carriers that have diffused down the thermal gradients in the semiconductors.
2. $K\,\Delta T = K\,(T_h - T_c)$. This term represents the rate at which heat is conducted away from the heat source by charge carriers and phonons in the semiconductors.
3. $I^2R/2$. This rate corresponds to the Joule heat that is generated in the semiconductors, one half of which is assumed to flow into the heat source.

The electrical power P made available to an external load resistance R_L can be shown to be given by the product of the current I and the terminal voltage V_t:

$$P = IV_t = I[(S_p - S_n)\,\Delta T - IR], \tag{W11.53}$$

where $(S_p - S_n)\,\Delta T$ is the total thermoelectric voltage due to the Seebeck effect. The efficiency of this thermoelectric generator in converting heat energy into electrical energy is given by $\eta = P/\dot{Q}$. It can be shown that η is maximized when the combined material parameter Z given by

$$Z = \frac{(S_p - S_n)^2}{(\sqrt{\rho_n \kappa_n} + \sqrt{\rho_p \kappa_p})^2} \tag{W11.54}$$

is maximized. When S_p and S_n have the same magnitude but are of opposite signs, and when the two semiconductors have the same thermal conductivities κ and electrical resistivities ρ, Z takes on the following simpler form:

$$Z = \frac{S^2}{\rho\kappa}. \tag{W11.55}$$

[†] It is assumed here for simplicity that the thermopowers S, thermal conductivities κ, and electrical resistivities ρ of the two semiconductors are independent of temperature. In this case the Thomson heat is zero and need not be included in the analysis.

High values of S are needed to increase the magnitudes of the Peltier effect and the thermoelectric voltage, low values of ρ are needed to minimize I^2R losses, and low values of κ are needed to allow higher temperature gradients and hence higher values of T_h. The dimensionless product ZT is known as the *thermoelectric figure of merit*. Despite extensive investigations of a wide range of semiconductors, alloys, and semimetals, the maximum currently attainable value of ZT is only about 1. When maximum power transfer is desired, independent of the efficiency of the transfer, the parameter to be maximized is then $Z' = S^2/\rho$.

Typical efficiencies for thermoelectric devices are in the range 10 to 12%. Thermoelectric power sources that obtain their heat input from the decay of radioactive isotopes are used on deep-space probes because of their reliability and convenience and because solar energy is too weak to be a useful source of electrical energy in deep space far from the sun.

Thermoelectric refrigeration employs the same configuration of semiconductors as used in thermoelectric generation, but with the load resistance R_L replaced by a voltage source V, as also shown in Fig. W11.25. In this case, as the current I flows around the circuit, heat is absorbed at the cooled end or heat "source" and is rejected at the other end, thereby providing refrigeration. As an example of thermoelectric refrigeration, when n- and p-type alloys of $Si_{0.78}Ge_{0.22}$ are used, the value $\Delta S = S_p - S_n = 0.646$ mV/K is obtained. With $T_h = 270$ K and $I = 10$ A, each n-p semiconductor pair can provide a cooling power of $P = IT_h\,\Delta S = 1.74$ W. While the use of thermoelectric refrigeration is not widespread due to its low efficiency compared to compressor-based refrigerators, it is a convenient source of cooling for electronics applications such as computers and infrared detectors.

Since different semiconductors possess superior thermoelectric performance in specific temperature ranges, it is common to employ cascaded thermoelements in thermoelectric generators and refrigerators, as shown in the multistage cooling device

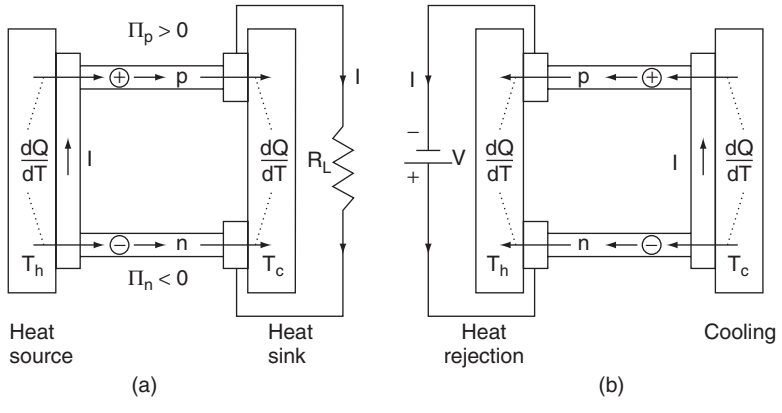

Figure W11.25. Schematic diagrams of (*a*) a thermoelectric power generator and (*b*) a thermoelectric refrigerator. In the thermoelectric generator or thermopile two semiconductors, one n-type and the other p-type, each carry a heat flux from a heat source to a heat sink. In the thermoelectric refrigerator the same configuration of semiconductors is employed, but with the load resistance R_L replaced by a voltage source V. In this case, as the current I flows around the circuit, heat is absorbed at the cooled end or heat "source" and is rejected at the other end, thereby providing refrigeration.

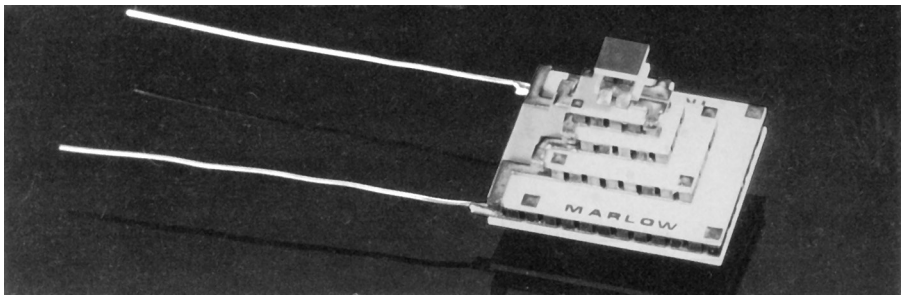

Figure W11.26. Cascaded thermoelements are employed in thermoelectric generators and refrigerators, as shown in the cooling module pictured here. (From G. Mahan et al., *Phys. Today*, Mar. 1997, p. 42. Copyright © 1997 by the American Institute of Physics.)

pictured in Fig. W11.26. In this way each stage can operate in its most efficient temperature range, thereby improving the overall efficiency and performance of the device. Temperatures as low as $T = 160$ K can be reached with multistage thermoelectric refrigerators.

The semiconductor material properties involved in the dimensionless figure of merit ZT for both power generation and for refrigeration are usually not independent of each other. For example, when the energy gap E_g or the doping level N_d or N_a of a semiconductor are changed, the electronic contributions to all three parameters, S, ρ, and κ, will change. It is reasonable, however, to assume that the lattice or phonon contribution κ_l to $\kappa = \kappa_e + \kappa_l$ is essentially independent of the changes in the electronic properties. To illustrate these effects, the values of S, ρ, and κ and their changes with carrier concentration are shown at room temperature in Fig. W11.27 for an idealized semiconductor. It can be seen that the quantity $Z = S^2/\rho\kappa$ has a maximum value in this idealized case near the middle of the range at the relatively high carrier concentration of $\approx 10^{25}$ m^{-3}. As a result, the dominant thermoelectric materials in use today are highly doped semiconductors.

The parameter Z has relatively low values in both insulators and metals. At the lower carrier concentrations found in insulators, Z is low due to the resulting increase in the electrical resistivity ρ and also at the higher carrier concentrations found in metals due both to the resulting increase in the electronic contribution to the thermal conductivity κ and to the decrease of S. The decrease in S with increasing carrier concentration occurs because a smaller thermovoltage is then needed to provide the reverse current required to balance the current induced by the temperature gradient. These decreases in S with increasing n or p can also be understood on the basis of Eqs. (W11.17) and (W11.18), which indicate that $S_n \propto (E_c - \mu)$ while $S_p \propto (\mu - E_v)$. Either $(E_c - \mu)$ or $(\mu - E_v)$ decrease as the chemical potential μ approaches a band edge as a result of doping. It is important that thermal excitation of electrons and holes not lead to large increases in carrier concentrations at the highest temperature of operation, $T_{\max}$, since this would lead to a decrease in S. It is necessary, therefore, that the energy gap E_g of the semiconductor be at least 10 times $k_B T_{\max}$.

A useful method for increasing the efficiency η of thermoelectric devices is to increase the temperature T_h of the hot reservoir, thereby increasing both the Peltier heat $\Pi = TS$ and the figure of merit ZT. In this way the *Carnot efficiency limit* $(T_h - T_c)/T_h$ will also be increased. The temperature T_h can be increased by reducing

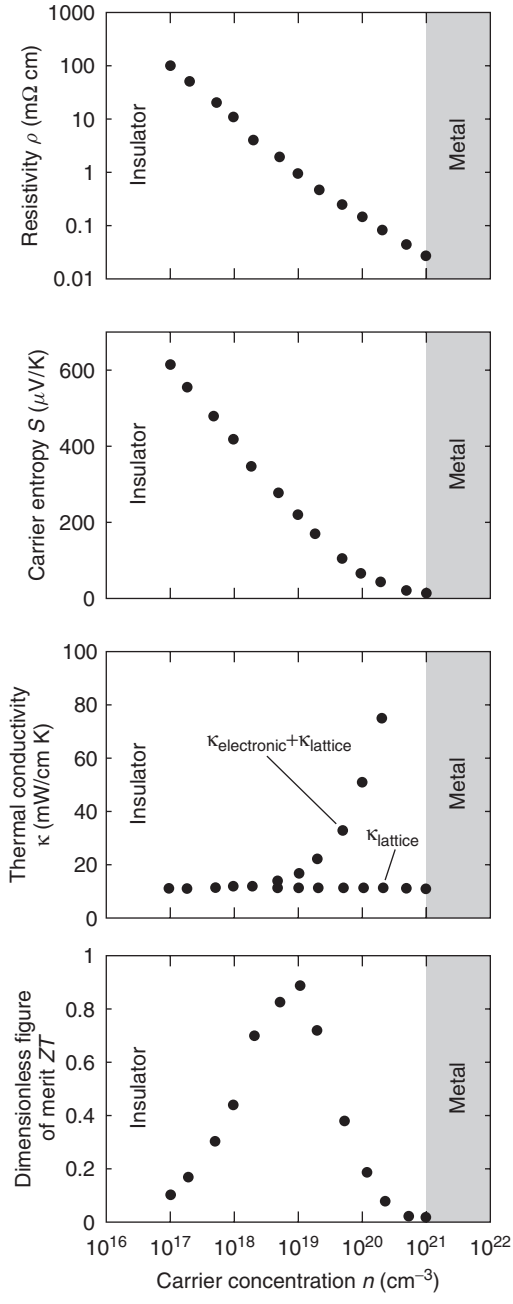

Figure W11.27. Effects of changing the carrier concentration on the thermoelectric parameter $Z = S^2/\rho\kappa$ and the values of S (the thermopower or carrier entropy), ρ, and κ for an idealized semiconductor. The energy gap E_g increases to the left in this figure. (From G. Mahan et al., *Phys. Today*, Mar. 1997, p. 42. Copyright © 1997 by the American Institute of Physics.)

the phonon mean free path, thereby decreasing κ_l through a disturbance of the periodic lattice potential. This is typically accomplished by alloying or by introducing lattice defects such as impurities. Another method of decreasing κ_l is to choose a semiconductor with a high atomic mass M since the speed of the lattice waves is proportional to $M^{-1/2}$.

Current research into the development of new or improved thermoelectric materials involves studies of a wide range of materials, including the semiconductors PbTe, Si:Ge alloys, Bi_2Te_3, and Bi:Sb:Te alloys, which are in current use. It can be shown in these "conventional" semiconductors that maximizing ZT is equivalent to maximizing $N(m^*)^{3/2}\mu/\kappa_l$, where N is the number of equivalent parabolic energy bands for the carriers, and m^* and μ are the electron or hole effective mass and mobility, respectively. Other novel materials under investigation include crystals with complicated crystal structures, such as the "filled" *skudderite* antimonides with 34 atoms per unit cell and with the general formula RM_4Sb_{14}. Here M is Fe, Ru, or Os, and R is a rare earth such as La or Ce. These crystals can have very good thermoelectric properties, with $ZT \approx 1$. This is apparently related to the lowering of κ_l due to the motions of the rare earth atoms inside the cages which they occupy within the skudderite structure.

Appendix W11A: Landau Levels

In this appendix an electron in the presence of a uniform magnetic field is considered. The Hamiltonian is

$$H = \frac{1}{2m_e^*}(\mathbf{p} + e\mathbf{A})^2, \tag{W11A.1}$$

where $\mathbf{A}$ is the vector potential. The magnetic induction is given by $\mathbf{B} = \nabla \times \mathbf{A}$, which automatically satisfies the condition $\nabla \cdot \mathbf{B} = 0$. A uniform magnetic field in the z direction may be described by the vector potential $\mathbf{A} = -\mathbf{B}y\hat{i}$. The Schrödinger equation $H\psi = E\psi$ for motion in the xy plane becomes

$$\frac{1}{2m_e^*}(p_x - eBy)^2\psi + \frac{p_y^2}{2m_e^*}\psi = E\psi. \tag{W11A.2}$$

This may be separated by choosing $\psi(x, y) = u(y)\exp(ik_x x)$, so

$$\left[\frac{p_y^2}{2m_e^*} + \frac{m_e^*\omega_c^2}{2}\left(y - \frac{\hbar k_x}{eB}\right)^2 - E\right]u(y) = 0, \tag{W11A.3}$$

where $\omega_c = eB/m_e^*$ is the cyclotron frequency. This may be brought into the form of the Schrödinger equation for the simple harmonic oscillator in one dimension by making the coordinate transformation $y' = y - \hbar k_x/eB$. The energy eigenvalues are $E = (n + 1/2)\hbar\omega_c$, where $n = 0, 1, 2, \ldots$. The effect of electron spin may be included by adding the Zeeman interaction with the spin magnetic moment. Thus

$$E = \left(n + \frac{1}{2}\right)\hbar\omega_c + g\mu_B B m_s, \tag{W11A.4}$$

where μ_B is the Bohr magneton, $g \approx 2$, and $m_s = \pm\frac{1}{2}$. The energy is independent of the quantum number k_x.

From Eq. (W11A.1) it is seen that the solution to the Schrödinger equation in a region of space where the vector potential is varying as a function of position is

$$\psi(\mathbf{r}) = \exp\left(i\mathbf{k}\cdot\mathbf{r} - i\frac{e}{\hbar}\int^{\mathbf{r}} \mathbf{A}(\mathbf{r}')\cdot d\mathbf{r}'\right). \tag{W11A.5}$$

REFERENCES

Grove, A. S., *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.

Hovel, H. J., *Solar Cells*, Vol. 11 in R.K. Willardson and A. C. Beer, eds., *Semiconductors and Semimetals*, Academic Press, San Diego, Calif., 1975.

Mott, N. F., and E. A. Davis, *Electronic Processes in Non-crystalline Materials*, 2nd ed., Clarendon Press, Oxford, 1979.

Zallen, R., *The Physics of Amorphous Solids*, Wiley, New York, 1983.

Zemansky, M. W., and R. H. Dittman, *Heat and Thermodynamics*, 6th ed., McGraw-Hill, New York, 1981.

PROBLEMS

W11.1 Prove that holes behave as positively charged particles (i.e., that $q_h = -q_e = +e$) by equating the current $\mathbf{J}_e = (-e)(-\mathbf{v}_e) = +e\mathbf{v}_e$ carried by the "extra" electron II in the valence band in Fig. 11.6 with the current $\mathbf{J}_h$ carried by the hole.

W11.2 Derive the expressions for the intrinsic carrier concentration $n_i(T)$ and $p_i(T)$, given in Eq. (11.29), and for the temperature dependence of the chemical potential $\mu(T)$, given in Eq. (11.30), from Eq. (11.27) by setting $n_i(T) = p_i(T)$.

W11.3 Consider the high-temperature limit in an n-type semiconductor with a concentration N_d of donors and with no acceptors. Show that the approximate concentrations of electrons and holes are given, respectively, by $n(T) \approx n_i(T) + N_d/2$ and $p(T) \approx p_i(T) - N_d/2$). [*Hint*: Use Eq. (11.35).]

W11.4 Calculate the average scattering time $\langle\tau\rangle$ for defect or phonon scattering at which the broadening of the two lowest energy levels for electrons confined in a two-dimensional quantum well of width $L_x = 10$ nm causes them to overlap in energy. Take $m_c^* = m$.

W11.5 Derive the expression $R_H = (p\mu_h^2 - n\mu_e^2)/e(n\mu_e + p\mu_h)^2$ for the Hall coefficient for a partially compensated semiconductor from the general expression for R_H for two types of charge carriers given in Eq. (11.48).

W11.6 If ΔV is the voltage drop that exists as a result of a temperature difference ΔT in a semiconductor in which no current is flowing, show that ΔV and ΔT have the same sign for electrons and opposite signs for holes and that the correct expression for calculating the thermoelectric power is $S = -\Delta V/\Delta T$.

W11.7 **(a)** Using Vegard's law given in Eq. (11.62) and the data presented in Table 11.9, find the composition parameter x for which $Al_{1-x}B_xAs$ alloys (assuming they exist) would have the same lattice parameter as Si.

(b) What value of E_g would Vegard's law predict for an alloy of this composition? [*Hint*: See Eq. (11.64).]

W11.8 Using the data presented in Table 2.12 for $r_{\text{cov}}(\text{Ga})$ and $r_{\text{cov}}(\text{As})$ and assuming that $d(\text{Ga} - \text{As}) = r_{\text{cov}}(\text{Ga}) + r_{\text{cov}}(\text{As})$, calculate the parameters E_h, C, E_g, and f_i for GaAs based on the dielectric model of Phillips and Van Vechten. *Note*: Estimate k_{TF} using the definition given in Section 7.17.

W11.9 Plot on a logarithmic graph the carrier concentrations n and p and their product $n\,p$ at $T = 300$ K as a function of the concentration of injected carriers $\Delta n = \Delta p$ from 10^{20} up to 10^{26} m^{-3} for the n-type Si sample with a donor concentration $N_d = 2 \times 10^{24}$ m^{-3} described in the textbook in Section 11.12. Identify on the graph the regions corresponding to low- and high-level carrier injection.

W11.10 By integrating Eq. (11.71), show that the buildup of the hole concentration $p(t)$ from its initial value p_0 is given by Eq. (11.74). Also, by integrating Eq. (11.76), show that the decay of the hole concentration $p(t)$ to its equilibrium value p_0 is given by Eq. (11.77).

W11.11 Using the fact that the additional output voltage ΔV_c in the collector circuit of the *npn* transistor amplifier described in Section W11.8 is equal to $[I_c(v) - I_c(v = 0)]R_c$, show that the voltage gain G is given by R_c/R_e.