alloys. This result indicates that the spins in the most concentrated alloy are not as "susceptible" as free spins in their response to external magnetic fields. Instead, their coupling to and interaction with each other limits their ability to respond to external fields and hence lowers their susceptibility χ. The type of interaction responsible for this behavior in AuMn alloys is an indirect interaction mediated by the conduction electrons.

W9.5 Spin Glasses and the RKKY Interaction

Clear evidence for the existence of the RKKY interaction has been found from studies of the magnetic properties of dilute alloys (e.g., Mn in Au, Ag, Cu, and Zn). When the spins of magnetic Mn^{2+} ions are coupled to each other via the conduction electrons, the average energy of the spin–spin interaction $\langle U_{\mathrm{RKKY}} \rangle$ is given by nV_0, where n is the concentration of Mn^{2+} ions per unit volume. This energy of interaction between spins competes with the energy of thermal disorder $k_B T$, with the result that the free-spin Curie law $\chi(T) = C/T$ is modified and becomes instead

$$\chi(T) = \frac{C}{T + \theta}. \tag{W9.1}$$

Here C is again the Curie constant as defined in Eq. (9.26) and $\theta \approx nV_0/k_B > 0$ is the Curie–Weiss temperature.[†] Equation (W9.1) is known as the *Curie–Weiss law* for the magnetic susceptibility and is valid for $T \gg \theta$ (i.e., for $k_B T \gg nV_0$).

Note that $\chi(T) = C/(T + \theta)$ with $\theta > 0$ is smaller than the free-spin susceptibility $\chi(T) = C/T$ for all T, indicating again that spin–spin interactions reduce the ability of the interacting spins to respond to external magnetic fields. This behavior has already been illustrated in Fig. W9.2, where, as stated previously, χ for the highest-concentration AuMn alloy at low T falls below the straight line that represents the Curie law behavior observed at higher T .

As $T \to \infty$ the Curie and Curie–Weiss laws become essentially identical since thermal fluctuations will always overcome magnetic interactions in this limit. The most significant difference is found for $T \ll \theta$, where $\chi(T) = C/(T + \theta)$ reaches a finite value while $\chi(T) = C/T$ for free spins diverges as $T \to 0$. The dependence of χ on T expressed by the Curie–Weiss law in Eq. (W9.1) is also observed in ferromagnetic and antiferromagnetic materials in their paramagnetic states above their respective critical temperatures T_c. For ferromagnets it is found that $\theta < 0$, whereas for antiferromagnets $\theta > 0$.

W9.6 Kondo Effect and s–d Interaction

One more interesting effect involving localized spins and the conduction electrons in metals can be mentioned. At sufficiently low temperatures the s–d or exchange interaction given in Eq. (9.32) can lead to a complicated many-body ground state of the system of the spin $\boldsymbol{S}$ and the conduction electrons of the metal. As already mentioned, the scattering of an electron from a magnetic ion can cause the spin of the scattered electron to flip (i.e., to change its direction), with a compensating change

[†] A. I. Larkin and D. E. Khmel'nitskii, *Sov. Phys. JETP*, **31**, 958 (1970).

TABLE W9.2 Competing Effects for Localized Spins in Metals: Thermal, RKKY, and Kondo Effects

$nV_0 \gg k_BT_K$: spin–spin interactions are dominant.	
$k_BT \gg nV_0$	Free spins
$k_BT \ll nV_0$	Frozen spins (spin glass behavior)
$k_BT_K \gg nV_0$: single-spin effects are dominant.	
$T \gg T_K$	Free spins
$T \ll T_K$	Compensated spins

occurring in the direction of the localized spin. The onset of this new ground state is typically signaled by the appearance of a minimum in the resistance of the metal as the temperature is lowered. It has been predicted that below a characteristic temperature T_K the spin $\boldsymbol{S}$ of the magnetic ion will be effectively canceled or compensated by the oppositely directed spins of the conduction electrons that interact with $\boldsymbol{S}$. This behavior is known as the *Kondo effect*, and the magnitude of the *Kondo temperature* T_K increases as the strength of the s–d interaction increases.

The s–d interaction, if sufficiently strong, can lead to complete mixing of the conduction electrons and the localized d electrons of the magnetic ion and therefore to the disappearance of the localized spin S. An example of this behavior is provided by Mn^{2+} ions, which do not retain well-defined magnetic moments in certain dilute alloys such as Mn in Al. In this case the characteristic temperature T_K for the s–d interaction is apparently very high, ≈ 1000 K, since for $T < T_K$, the spin will be compensated and hence effectively absent.

The three competing effects that ultimately determine the behavior and possibly even the existence of localized spins in metals are thermal effects, effects due to the spin–spin RKKY interaction, and the single-spin Kondo effect.[†] The characteristic energies that determine the strengths of these three effects are k_BT, nV_0, and k_BT_K, respectively. The possible regimes of behavior are defined in terms of the relative magnitudes of these three energies in Table W9.2. It can be seen that free-spin behavior should in principle always be observed in solids at sufficiently high T. The term *spin glass* used in the table is defined in the discussion of magnetism in disordered materials in Section W9.11.

W9.7 $\chi(T)$ for Ni

A test of the Curie–Weiss law $\chi(T) = C/(T - T_C)$ for the ferromagnet Ni is shown in Fig. W9.3, where χ_ρ^{-1} is plotted as a function of T. It can be seen that significant deviations from Curie–Weiss behavior occur just above $T_C = 627$ K. It is found experimentally for Fe that χ is proportional to $(T - T_C)^{-\gamma}$ as $T \to T_C$ from above. Here γ is measured to be 1.33 instead of the value 1 predicted by the Curie–Weiss law. The molecular field theory fails near T_C since it does not include the effects of fluctuations of the local magnetization.

[†] An alternative approach to the question of the existence of localized spins in metals has been developed by Anderson (P. W. Anderson, Phys. Rev., **124**, 41 (1961) and by Wolff (P. A. Wolff, Phys. Rev., **124**, 1030 (1961).) For a useful discussion of this approach, see White and Geballe (1979).

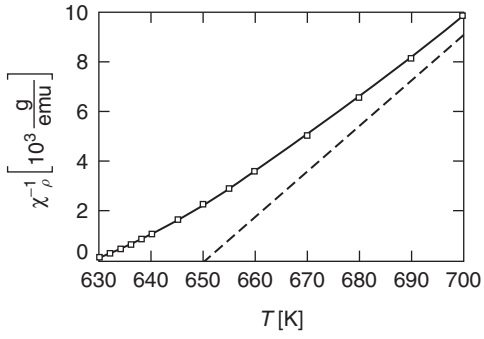

Figure W9.3. Test of the Curie–Weiss law $\chi(T) = C/(T - T_C)$ for the ferromagnet Ni in the form of a plot of χ_ρ^{-1} as a function of T . Deviations from Curie–Weiss behavior are observed just above $T_C = 627$ K. The straight line is the extrapolation of the results obtained for $T > 700$ K and is given by $\chi(T) = C/(T - \theta)$ where $\theta = 650$ K. [Data From J. S. Kouvel et al., *Phys. Rev.*, **136**, A1626 (1964).]

W9.8 Hubbard Model

An approach that attempts to include both itinerant and localized effects and also electron correlations within the same model is based on a proposal by Hubbard.[†] In the *Hubbard model* the oversimplified view is taken that the electrons in the partially filled shell of the free ion enter a single localized orbital in the solid. There are two important energies in the Hubbard model. The *Coulomb repulsion energy* $U > 0$ represents the effects of electron correlations between pairs of opposite-spin electrons occupying the same orbital on a given ion, and the *hopping* or *tunneling energy* is t. The parameter t is effectively the matrix element between states on neighboring ions which differ by one electron of a given spin direction and is therefore related to the energy required for an electron to hop from one site (i.e., one ion) to one of its NNs without changing its spin direction. In a one-state Hubbard model there is one orbital per atom and each orbital can be occupied by electrons in four different ways: (1) the orbital is empty: $(-,-)$, (2) and (3) the orbital is occupied by either a spin-up or a spin-down electron: $(\downarrow, -)$ or $(-,\uparrow)$, or (4) the orbital is doubly occupied: $(\downarrow,\uparrow)$.

In the limit $U \gg t$ and when there are just as many electrons as ions, there will be a strong preference for occupation of each orbital by a single electron (i.e., case 2 or 3 above). This limit corresponds to an antiferromagnetic insulator in which the effective exchange integral is $J = -4t^2/U$, with adjacent orbitals occupied by opposite spin electrons. In the opposite limit of $U \ll t$, the electrons are not localized but instead, form a band of itinerant electrons. Thus the Hubbard model is capable of describing a wide range of magnetic behavior in solids, depending on the relative values of the two parameters U and t. In addition, the Hubbard model has the advantage that it can be formulated so that the condition for local magnetic moment formation is not the same as that for the occurrence of long-range order in the spin system. The negative-U limit of the Hubbard model has been applied to charged defects in semiconducting and insulating solids. The defect is negatively charged when the orbital in question is

[†] J. Hubbard, *Proc. R. Soc.* A, **276**, 238 (1963); **277**, 237 (1964); **281**, 401 (1964).

doubly occupied, or positively charged when the orbital is unoccupied. The energy U can be effectively negative when lattice relaxations occur that favor negatively charged defects.

The Hubbard model goes beyond the one-electron tight-binding approximation presented in Chapter 7, in that it includes electron–electron interactions when two electrons reside on the same site. The application of the Hubbard model to high-T_c oxide-based superconductors is described briefly Chapter W16.

W9.9 Microscopic Origins of Magnetocrystalline Anisotropy

The microscopic origins of magnetocrystalline anisotropy can be viewed as arising from anisotropic interactions between pairs of spins when these interactions are significant and also from the interaction of a single spin with its local atomic environment (i.e., the crystal field). The *pair model* of Van Vleck, developed in 1937, attempts to explain the change of the energy of interaction of pairs of spins according to their directions relative to their separation $\mathbf{r}$. This type of interaction is called *anisotropic exchange*, in contrast to the isotropic Heisenberg exchange interaction of Eq. (9.30). The spin–orbit interaction is believed to be an important source of the magnetic anisotropy. In the pair model the first-order anisotropy coefficient K_1 is predicted to be proportional to a high power of the spontaneous magnetization M_s in the ferromagnet. This result can explain the observed rapid decrease of K_1 with increasing temperature, with M_s and K_1 both falling to zero at T_C.

The direction of the spin of a magnetic ion in a material can also depend on the nature of the crystal field acting on the ion. In this way the local atomic environment can influence the direction of the magnetization M, hence giving rise to anisotropy. In fact, the electronic energy levels of the ion are often modified by the interaction with the crystal field, as discussed in Section 9.3.

W9.10 $\chi_\parallel$ and $\chi_\perp$ for Antiferromagnetic Materials

The predicted differences between $\chi_\parallel$ and $\chi_\perp$ discussed in the textbook are clear evidence that the magnetic properties of antiferromagnetic materials can be expected to be anisotropic below T_N. For example, in MnO the preferred directions for the sublattice magnetizations $\boldsymbol{M}_{sA}$ and $\boldsymbol{M}_{sB}$, and hence the directions corresponding to $\chi_\parallel$, can be seen from Fig. 9.17 to be the $[\bar{1}01]$ and $[10\bar{1}]$ directions in the {111} planes. Also, if an antiferromagnet were perfectly isotropic below T_N, it would follow that $\chi_\parallel = \chi_\perp$. Since $\chi_\perp > \chi_\parallel$ for $T < T_N$, it can be energetically favorable for the spins to rotate so that the spin axis is perpendicular to the applied field. This "flopping" of the spin axis occurs at a critical applied magnetic field which is determined by the relative strengths of the magnetocrystalline anisotropy and the antiferromagnetic interactions.

W9.11 Magnetism in Disordered Materials

Spin glasses (i.e., dilute magnetic alloys) are the focus of this section, due to the fairly simple, yet important ideas involved in the explanation of their magnetic behavior. In general, nonuniform internal molecular fields $\mathbf{B}_{\text{eff}}$ whose magnitudes and directions vary from spin to spin are present in amorphous magnetic materials. The probability distribution $P(\mathbf{B}_{\text{eff}})$ of the magnitudes of these internal fields in spin glasses (e.g.

$Cu_{0.99}Fe_{0.01}$) will be nonzero even at $B_{\text{eff}} = 0$. Thus there will always be spins with $B_{\text{eff}} = 0$ which are effectively free to respond to thermal excitations and to external magnetic fields. This is clearly not the case in the magnetically ordered materials discussed in the textbook, in which every spin experiences a nonzero molecular field, at least below the critical temperature T_C or T_N for magnetic ordering.

In sufficiently dilute spin glasses and at relatively high temperatures each spin can in principle be thought of as being free or as interacting with at most one other spin in the material. The spins typically interact via the indirect RKKY interaction through the conduction electrons. In this case the contributions of the interacting spins to the magnetization M, the magnetic susceptibility χ, and the magnetic contribution C_M to the specific heat obey the following *scaling laws* involving temperature T and magnetic field H:

$$\begin{aligned} \frac{M(H,T)}{n} &= F_M\left(\frac{T}{n}, \frac{H}{n}\right), \\ \chi(T) &= F_\chi\left(\frac{T}{n}\right), \\ \frac{C_M(T)}{n} &= F_C\left(\frac{T}{n}\right). \end{aligned} \tag{W9.2}$$

Here n is the concentration of magnetic impurities, and F_M, F_χ, and F_C are functions only of H and T through the reduced variables H/n and T/n. These scaling laws follow from the $1/r^3$ dependence of the RKKY interaction on the separation r between spins, as presented in Eqs. (9.33) and (9.34).

Since the average separation $\langle r \rangle$ between randomly distributed spins can be approximated by $n^{-1/3}$, it follows that the average strength $\langle J_{\text{RKKY}}(r) \rangle$ of the interaction between spins is proportional to $\langle V_0/r^3 \rangle$ (i.e., to nV_0), where V_0 is a constant for a given combination of magnetic impurity and host material. The value for V_0 in dilute CuMn alloys[†] is $V_0 = 7.5 \times 10^{-50}$ J $\cdot$ m^3. Taking a Mn concentration of 0.1 at % = 1000 parts per million (ppm) in Cu yields $n = 8.45 \times 10^{25}$ Mn spins/m^3 and $nV_0 = 6.3 \times 10^{-24}$ J $\approx 4 \times 10^{-5}$ eV. This concentration corresponds to an average distance between Mn spins of about 2 nm. The value of J_{sd} for CuMn can be obtained from Eq. (9.35) using the result given above for V_0, a density of states for Cu of $\rho(E_F) = 2.34 \times 10^{47}$ J^{-1}m^{-3}. The value so obtained is $J_{sd} = 3.45 \times 10^{-19}$ J $= 2.16$ eV.

The scaling behavior of $\chi(T)$ predicted above has already been demonstrated in Fig. W9.2, where χ is shown plotted as a function of T/n for several AuMn alloys. The measured magnetization M for three of these AuMn alloys at a fixed value of T/n is shown in Fig. W9.4 plotted as M/n versus H/n. The scaling behavior predicted is again observed. The magnetization $M(H)$ shown here falls well below the corresponding Brillouin function $M = ng\mu_B J B_J(g\mu_B JB/k_B T)$, which would apply if the spins were free (i.e., completely noninteracting).

Experimental results for the magnetic contribution C_M to the specific heat of a series of dilute alloys of Mn in Zn are shown in Fig. W9.5, where C_M/n is plotted as a function of T/n. Scaling is observed for the more-concentrated alloys where RKKY

[†] F. W. Smith, *Phys. Rev. B*, **14**, 241 (1976).

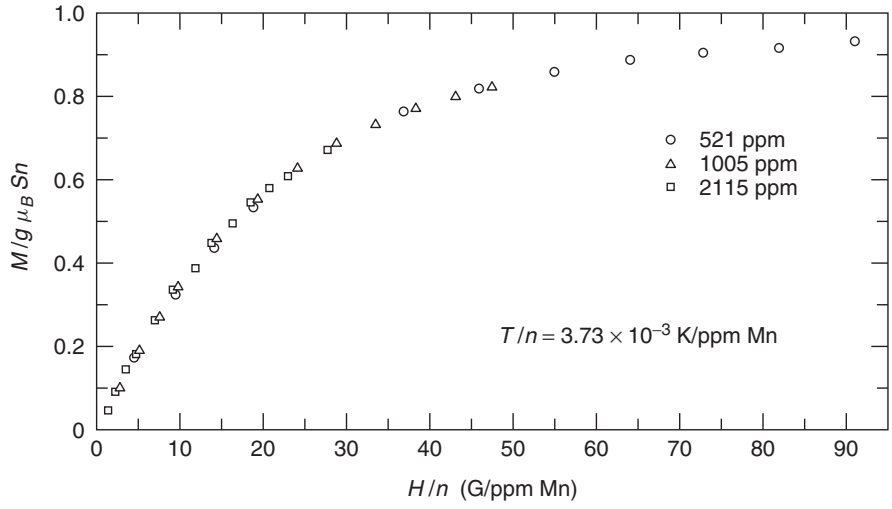

Figure W9.4. Contribution of the Mn spins to the magnetization M for three dilute alloys of Mn in Au at a fixed value of T/n plotted as $M/g\mu_B Sn$ versus H/n. The predicted scaling behavior $M(T)/n = F_M(H/n)$ is observed. [From J. C. Liu, B. W. Kasell, and F. W. Smith, *Phys. Rev. B*, **11**, 4396 (1975). Copyright © 1975 by the American Physical Society.

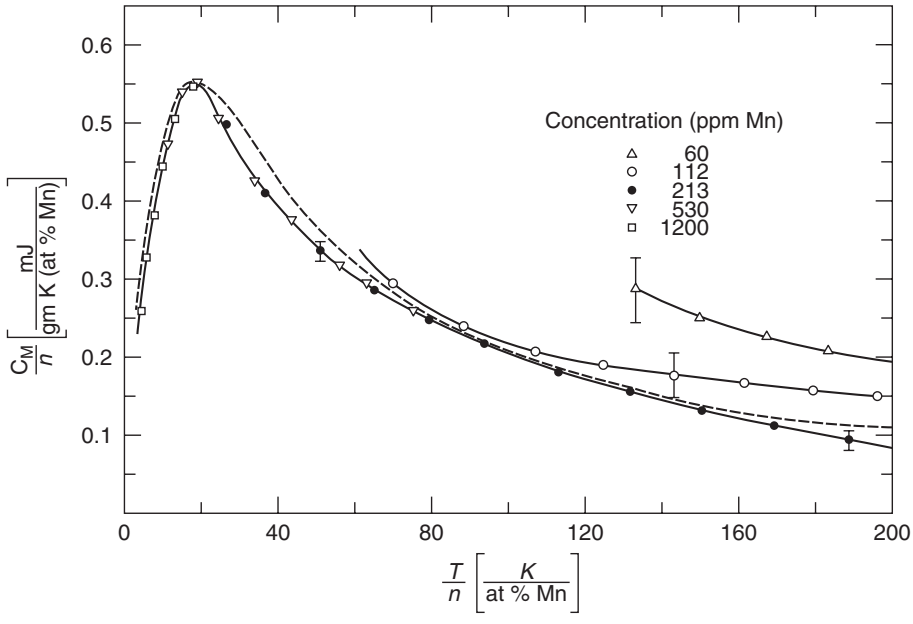

Figure W9.5. Experimental results for the magnetic contribution C_M to the specific heat of a series of dilute alloys of Mn in Zn, with C_M/n plotted as a function of T/n. Scaling is observed for the more concentrated alloys. [From F. W. Smith, *Phys. Rev. B*, **9**, 942 (1974). Copyright © 1974 by the American Physical Society.]

interactions dominate, whereas evidence for single-impurity effects, possibly due to the Kondo effect, is observed for the more dilute alloys at higher values of T/n. The peak observed in the measured specific heat at $T/n \approx 20$ K/(at % Mn) corresponds to a value of the ratio $k_B T/nV_0$ of thermal to RKKY interaction energies approximately equal to 2. At lower T (i.e., for $k_B T < nV_0$) interactions between the spins cause them to "freeze" in the local molecular field due to their neighboring spins. At $T = 0$ K the spin glass is magnetically "frozen" and the spins are oriented along the direction of their local molecular field. As T is lowered it is found experimentally that $C_M \propto n^2$, indicating that interactions first appear between pairs of spins. The typical size of an interacting cluster of spins increases as T decreases or n increases until the interactions extend throughout the entire spin system.

The magnetic behavior of dilute spin glasses can thus be understood as resulting from RKKY interactions between pairs of spins. Evidence for clusters of spins can be found in more concentrated spin glasses, such as Cu containing more than a few atomic percent Mn or in alloys such as Cu_xNi_{1-x} and Fe_xAl_{1-x}. Although the magnetic behavior is much more complicated in these concentrated alloys, the RKKY interaction still plays an important role. The term *mictomagnetism* is sometimes used to describe such materials in which the orientations of the spins are disordered and frozen at low temperatures.

REFERENCES

Sugano, S., Y. Tanabe, and H. Kamimura, *Multiplets of Transition-Metal Ions in Crystals*, Academic Press, San Diego, Calif., 1970.

White, R. M., and T. H. Geballe, *Long Range Order in Solids*, Suppl. 15 of H. Ehrenreich, F. Seitz, and D. Turnbull, eds., *Solid State Physics*, Academic Press, San Diego, Calif., 1979.

PROBLEMS

W9.1 Using Hund's rules, find the values of S, L, and J for the atoms in the $4d$ transition element series (Y to Pd). Compare these values with the corresponding results given in Table 9.1 for the $3d$ series.

W9.2 From Fig. 9.5 it can be seen that, relative to the degenerate spherically symmetric level, the d_{xy}, d_{yz}, and d_{xz} orbitals are shifted lower in energy by $2\Delta_o/5$ for the octahedral case and higher in energy by $2\Delta_t/5$ for the tetrahedral case. The corresponding opposite shifts for the $d_{x^2-y^2}$ and d_{z^2} orbitals are by the amount $3\Delta_o/5$ or $3\Delta_t/5$ for the octahedral and tetrahedral cases, respectively. Show that these energy shifts are such that the total energy of the $3d^{10}$ configuration will be the same in both the spherically symmetric and crystal-field-split cases.

W9.3 Using the schematic energy-level diagrams shown in Fig. 9.5, calculate the crystal field stabilization energies (CFSEs) and spins S [assuming that orbital angular momentum L is quenched (i.e., $L = 0$)]:

(a) For the $3d^n$ ions in octahedral sites. Compare your results with the values presented in Table 9.2.

(b) For the $3d^n$ ions in tetrahedral sites.

(c) In a ferrite such as Fe_3O_4, will Fe^{2+} ions prefer to enter octahedral or tetrahedral sites on the basis of their crystal field stabilization energy CFSE? What about Fe^{3+} ions?

W9.4 Show that the induced saturation magnetization M_{sat} for a system of $n = 10^{26}/\text{m}^3$ free spins in a material makes a negligible contribution to the magnetic induction B .

W9.5 Derive the general expression for the Brillouin function $B_J(x)$ given in Eq. (9.24).

W9.6 Consider a dilute magnetic alloy that contains $n = 2 \times 10^{23}$ spins/m^3. At low T the spins can be saturated in a field $H \approx 4 \times 10^6$ A/m, with M_{sat} measured to be 5.56 A/m. At high T the spins obey a Curie–Weiss law $\chi(T) = C/(T + \theta)$ with Curie constant $C = 7.83 \times 10^{-6}$ K and Curie–Weiss temperature $\theta = 0.1$ K.

(a) From these data determine the spin J and g factor of the spins.

(b) Are the spins free? If not, what type of spin–spin interaction would you conclude is present in the alloy?

W9.7 Consider a spin S in a ferromagnet interacting only with its z NN spins ($z = 12$ for an FCC lattice).

(a) Using Eq. (9.41) show that the Curie–Weiss temperature θ is given by $\theta = zS(S + 1)J(\mathbf{R}_{NN})/3k_B$, where the exchange integral $J(r)$ is evaluated at the NN distance $\mathbf{R}_{NN}$.

(b) Using the approximate values $\theta \approx T_C = 1043$ K and $S \approx 1$ for BCC ferromagnetic α-Fe, calculate the value of $J(\mathbf{R}_{NN})$.

W9.8 Show that at the Néel temperature T_N, the predicted maximum value for the magnetic susceptibility χ according to the molecular field model is $\chi_{\text{max}} = -1/\lambda_{\text{AB}} > 0$. Explain why this prediction that χ_{max} is proportional to $1/\lambda_{\text{AB}}$ is physically reasonable.

W9.9 Calculate the Pauli paramagnetic susceptibility χ_P for Na metal according to the free-electron theory.

CHAPTER W10

Mechanical Properties of Materials

W10.1 Relationship of Hooke's Law to the Interatomic $U(r)$

Since the macroscopic deformation of a solid reflects the displacements of individual atoms from their equilibrium positions, it should not be surprising that the elastic response of a solid is determined by the nature of the interactions between neighboring atoms. In fact, Hooke's law can be derived from the form of the potential energy of interaction $U(r)$ for a pair of atoms, as shown for a pair of hydrogen atoms in Fig. 2.1 of the textbook.[†] The equilibrium separation of the two atoms corresponds to the minimum in the $U(r)$ curve at $r = r_0$. Since $U(r)$ is a continuous function, it can be expanded in a Taylor series about $r = r_0$, as follows:

$$U(r) = U(r_0) + (r - r_0)\left(\frac{dU}{dr}\right)_{r_0} + \frac{(r - r_0)^2}{2}\left(\frac{d^2U}{dr^2}\right)_{r_0} + \cdots. \tag{W10.1}$$

The first derivative, $(dU/dr)_{r_0}$, is equal to zero at the equilibrium separation $r = r_0$. In addition, cubic and other higher-order terms can be neglected since $(r - r_0) \ll r_0$ for the (typically) small displacements from equilibrium.

It follows that the force acting between a pair of atoms can be approximated by

$$F(r) = -\frac{dU(r)}{dr} = -(r - r_0)\left(\frac{d^2U}{dr^2}\right)_{r_0} = -k(r - r_0), \tag{W10.2}$$

where k is a constant. This result has the same form as Hooke's law since the displacement $(r - r_0)$ of atoms from their equilibrium positions is proportional to the restoring force F. This displacement is also inversely proportional to the curvature $(d^2U/dr^2)_{r_0}$ of the potential energy curve at $r = r_0$, which for a given material is a constant in a given direction.

It can be seen from Eqs. (10.21) and (W10.2) that *Young's modulus* E is proportional to the curvature $(d^2U/dr^2)_{r_0}$ of the potential energy. This is a reasonable result since the macroscopic deformations that correspond to the microscopic displacements of atoms from their equilibrium positions will be more difficult in materials where the potential energy well is deeper and hence $U(r)$ increases more rapidly as the atoms are displaced

[†] 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."

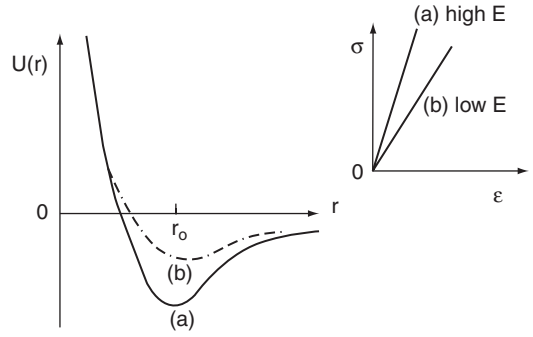

Figure W10.1. Schematic potential energies of interaction $U(r)$ for "deep" and "shallow" potential wells and corresponding stress–strain curves

from their equilibrium positions. This is illustrated schematically in Fig. W10.1 for the cases of "strong" and "weak" bonding between pairs of atoms, corresponding to "deep" and "shallow" potential wells, respectively. For the case of a material with strong bonding and a deep potential well, the curvature $(d^2U/dr^2)_{r_0}$ is high. Such a material will have a high stiffness E and a high slope for the initial linear portion of its stress–strain curve, as shown in the inset of this figure. The opposite will be true for a material having weak bonding, a shallow potential well, and a corresponding low curvature $(d^2U/dr^2)_{r_0}$. In this case the material will have a low stiffness E. It should be noted that the stress–strain curve will eventually become nonlinear as the stress increases, due to the nonparabolicity of the interatomic potential $U(r)$ for large displacements $(r - r_0)$.

Estimates for the magnitude of the elastic modulus E and its dependence on material properties can be obtained by noting that E, as a measure of the stiffness of a material, should be proportional to the stress needed to change the equilibrium separation between atoms in a solid.[†] For many materials with ionic, metallic, and covalent bonding, this stress is itself approximately proportional to the magnitude of the interatomic Coulomb force $F = q^2/4\pi\epsilon d^2$, where q is the ionic charge, d the interatomic separation, and ϵ the electric permittivity of the material. This stress should also be inversely proportional to the effective area, $\approx d^2$, over which the interatomic force acts. Thus the stress, and hence E, should be proportional to q^2/d^4.

A test of this relationship is presented in Fig. W10.2, where the bulk modulus B, defined in Section 10.6, is shown plotted as a function of the interatomic separation d in a logarithmic plot for three classes of materials with ionic, metallic, and covalent bonding, respectively. For each class of materials the measured values of B fall on a straight line with a slope close to -4, as predicted by the simple argument presented above. It is clear from this result that high elastic stiffness is favored in materials where the ions have large effective charges and are separated by small interatomic separations.

The magnitude of the elastic constants can also be estimated from the expression $E \approx q^2/4\pi\epsilon d^4$ by using $1/4\pi\epsilon \approx 9 \times 10^9$ N · m^2/C^2, $q = e = 1.6 \times 10^{-19}$ C, and $d \approx$

[†] See the discussion in Gilman (1969, pp. 29–42).

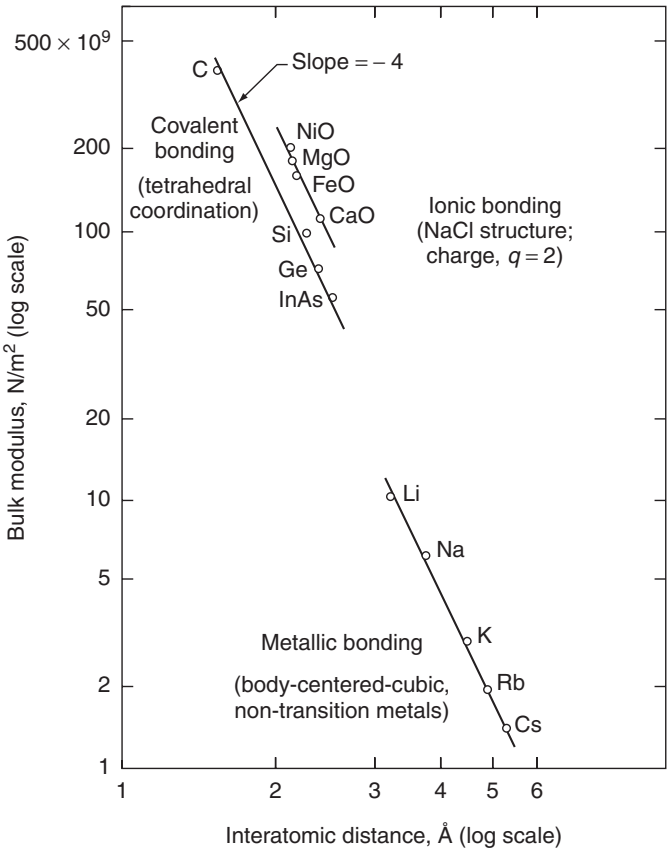

Figure W10.2. Logarithmic plot of the bulk modulus B versus the interatomic separation d for three classes of materials with ionic, metallic, and covalent bonding, respectively. (From A. G. Guy, *Introduction to Materials Science*, McGraw-Hill, New York, 1972. Reprinted by permission of the McGraw-Hill Companies.)

0.2 nm. The result obtained, $E \approx 100$ GPa, is consistent with the experimental values shown in Fig. W10.2 and listed in Table 10.2.

W10.2 Zener Model for Anelasticity

An interesting and useful model for describing anelastic processes has been proposed by Zener. This model deals with a *standard linear solid*, a solid in which the stress σ, the strain ε, and their first derivatives $\partial\sigma/\partial t$ and $\partial\varepsilon/\partial t$ are related to each other in a linear equation. Although Zener's model may not be sufficiently general to describe all types of anelastic effects, it is quite useful for the purpose of illustrating important general aspects of anelasticity.

In the Zener model the following equation is used to describe the anelastic effects illustrated in Fig. 10.9:

$$\sigma + \tau_\varepsilon \frac{\partial \sigma}{\partial t} = E_r \left(\varepsilon + \tau_\sigma \frac{\partial \varepsilon}{\partial t} \right). \tag{W10.3}$$

Here τ_ε is the time constant for the relaxation of stress under conditions of constant strain, and τ_σ is the time constant for relaxation of strain under conditions of constant stress.[†] The quantity E_r is the *relaxed elastic modulus*, that is, the stress/strain ratio σ/ε after all relaxation has occurred in the solid and when $\partial\sigma/\partial t$ and $\partial\varepsilon/\partial t$ are zero. If the changes in stress and strain in the material occur so rapidly (e.g., at sufficiently high frequencies) that relaxation cannot proceed to completion, it can be shown that the stress/strain ratio is given by the *unrelaxed elastic modulus* $E_u = E_r\tau_\sigma/\tau_\varepsilon$.

The solutions of Eq. (W10.3) for the conditions shown in Fig. 10.9*a* (i.e., after relaxation has occurred) are as follows:

$$\begin{aligned} &\sigma = \sigma_0 \text{ and } \partial\sigma/\partial t = 0: \qquad \varepsilon(t) = \varepsilon_\infty + (\varepsilon_0 - \varepsilon_\infty)e^{-t/\tau_\sigma}. \\ &\sigma = 0 \text{ and } \partial\sigma/\partial t = 0: \qquad \varepsilon(t) = \varepsilon_\infty e^{-t/\tau_\sigma}. \end{aligned} \tag{W10.4}$$

Here $\varepsilon_\infty = \sigma_0/E_r$. These expressions illustrate the kinetics to be expected for simple relaxation processes where the fraction of the relaxation completed in time t is $f(t) = 1 - e^{-t/\tau}$. Analogous equations can be derived for the time dependence of σ for the conditions shown in Fig. 10.9*b*.

The mechanical response of materials to dynamic conditions of stress and strain is of interest both for applications and for fundamental studies of anelasticity. Under dynamic conditions, stress and strain are often periodic functions of time, that is,

$$\sigma(t) = \sigma_0 e^{-i\omega t} \quad \text{and} \quad \varepsilon(t) = \varepsilon_0 e^{-i\omega t}, \tag{W10.5}$$

where the amplitudes σ_0 and ε_0 can be complex quantities. Upon substitution of $\sigma(t)$ and $\varepsilon(t)$, Eq. (W10.3) becomes

$$(1 - i\omega\tau_\varepsilon)\sigma_0 = E_r(1 - i\omega\tau_\sigma)\varepsilon_0. \tag{W10.6}$$

A *complex elastic modulus* E_c can then be defined as

$$E_c = \frac{E_r(1 - i\omega\tau_\sigma)}{1 - i\omega\tau_\varepsilon} = \frac{\sigma_0}{\varepsilon_0}. \tag{W10.7}$$

For a stress amplitude σ_0 that is real, this corresponds to a complex amplitude ε_0 for the strain.

Under dynamic conditions and due to either elastic aftereffects or strain relaxation, the strain ε will in general lag behind the stress σ by a phase angle ϕ (i.e., $\varepsilon(t) = \varepsilon_0 \exp[-i(\omega t - \phi)]$), whose tangent is given by

$$\tan\phi = \frac{\text{Im } E_c}{\text{Re } E_c} = \frac{\omega(\tau_\sigma - \tau_\varepsilon)}{1 + \omega^2\tau_\varepsilon\tau_\sigma}. \tag{W10.8}$$

The quantity $\tan\phi$, known as the *loss coefficient*, is often used as a measure of the magnitude of the *internal friction* or energy loss in a material. When $\tan\phi$ is small,

[†] While the use of a single relaxation time is appropriate for some materials, other materials, such as polymers, can have a large number of relaxation times, spanning many orders of magnitude.

it can be shown that $\tan\phi \approx \Delta U_{\rm el}/2\pi U_{\rm el} = 1/Q$, where $\Delta U_{\rm el}/U_{\rm el}$ is the fraction of elastic energy dissipated per oscillation. (Q is the *quality factor* of an electrical circuit, with $1/Q$ being a measure of energy dissipation.)

The predicted frequency dependence of the internal friction is illustrated in Fig. W10.3, where $\tan\phi$ is shown as a function of frequency, specifically $\omega(\tau_\sigma\tau_\varepsilon)^{1/2} = \omega\langle\tau\rangle$. It can be seen that $\tan\phi$ has a maximum value at $\omega\langle\tau\rangle = 1$ [i.e., at $\omega_{\max} = (\tau_\sigma\tau_\varepsilon)^{-1/2}$] and falls to zero for both $\omega \ll \omega_{\max}$ and $\omega \gg \omega_{\max}$. For low frequencies, $\omega \ll \omega_{\max}$, the solid is fully relaxed, the elastic modulus is E_r, and the internal friction is close to zero in the Zener model, since the strain has sufficient time to follow the applied stress (i.e., the phase angle $\phi \approx 0$). At high frequencies, $\omega \gg \omega_{\max}$, the solid is unrelaxed, the elastic modulus is E_u, and the internal friction is again close to zero.

Note that $E_u > E_r$ in Fig. W10.3, which follows from $\tau_\sigma > \tau_\varepsilon$. In this case the strain relaxes more slowly than the stress [see the definitions given earlier for τ_σ and τ_ε in Eq. (W10.3)]. It follows that the material will be stiffer at high frequencies than at low frequencies. The hysteresis loops for such material will actually be closed, straight lines with slopes given by E_r and E_u at very low and very high frequencies, respectively. Thus Hooke's law will be valid for $\omega \gg \omega_{\max}$ and $\omega \ll \omega_{\max}$. At $\omega = \omega_{\max}$ the hysteresis loop will have its maximum width and maximum area $\Delta U_{\rm el}$.

Zener has pointed out that although this model for a standard linear solid has several general features that are observed for real materials, it does not in fact correspond in detail to the behavior observed for any real solid. Nevertheless, measurements of internal friction as a function of frequency often show the behavior predicted by Zener's model, as shown in Fig. W10.4 for German silver, an alloy of Cu, Ni, and Zn.

W10.3 Typical Relaxation Times for Microscopic Processes

See Table W10.1, from which it can be seen that lattice vibrations, the motion of elastic waves, and the dissipation of heat are "fast" processes at $T \approx 300$ K, while the diffusion of interstitial atoms and the motion of grain boundaries can be considered to be "slow" processes.

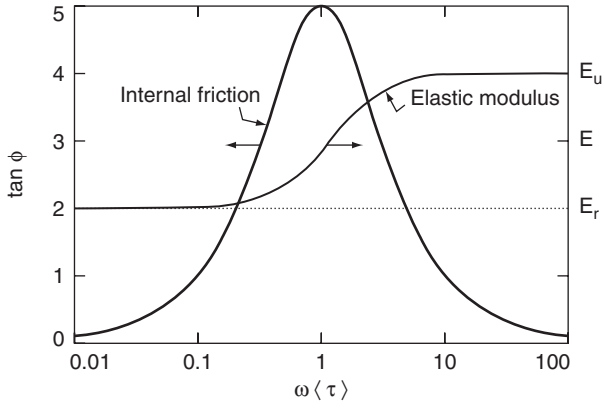

Figure W10.3. Magnitude of the internal friction $\tan\phi$ as a function of $\omega\langle\tau\rangle = \omega(\tau_\sigma\tau_\varepsilon)^{1/2}$. (Adapted from C. Zener, *Elasticity and Anelasticity of Metals*, University of Chicago Press, Chicago, 1948).

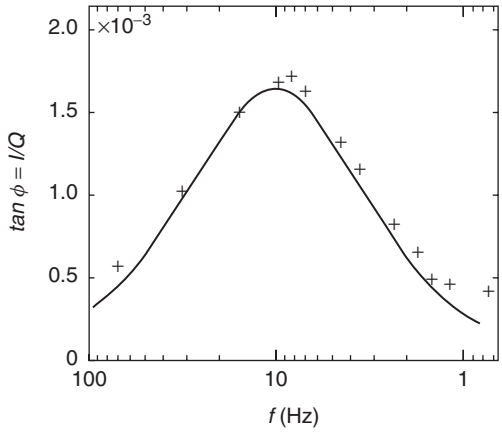

Figure W10.4. Magnitude of the internal friction $\tan\phi = 1/Q$ for German silver as a function of frequency. (From C. Zener, *Elasticity and Anelasticity of Metals*, University of Chicago Press, Chicago Copyright© 1948 by the University of Chicago. Reprinted by permission.)

TABLE W10.1 Typical Relaxation Times τ for Microscopic Processes in Solids at $T = 300$ K

Time Scale for τ (s)	Microscopic Process
10^{-14}	Electron collisions in metals
10^{-12}	Vibrations of atoms (lattice vibrations)
10^{-10}	
10^{-8}	Radiative recombination of electrons and holes
10^{-6}	
	Elastic wave traverses solid (as in brittle fracture)
10^{-4}	
	Dissipation of heat (thermal relaxation)
10^{-2}	
$10^{0} = 1$	
	(Time of typical tensile test $= t_{\text{test}}$)
10^{+2}	
10^{+4}	Diffusion of interstitial atoms
(1 week $\approx 6 \times 10^{5}$ s)	
10^{+6}	
(1 year $\approx 3 \times 10^{7}$ s)	Motion of grain boundaries
10^{+8}	Creep
	Flow of inorganic glasses

W10.4 Further Discussion of Work Hardening

The phenomenon of *work hardening* is difficult to treat theoretically, the most difficult aspect being to predict how the density and distribution of dislocations vary with the strain in the material. There is in fact no unique correlation between the level of strain and the resulting distribution of dislocations. The experimental situation is complicated by the fact that there can exist three distinct regions of work hardening when the plastic deformation is presented in the form of a shear stress–shear strain

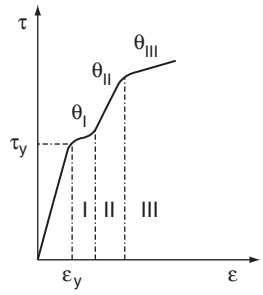

Figure W10.5. Shear stress–shear strain $\tau-\varepsilon$ curve for a typical single-crystal FCC metal. Three inelastic regions are shown, with the rate of work hardening in each region characterized by the slope $d\tau/d\varepsilon$, denoted by θ_{I}, θ_{II}, and θ_{III}, respectively

curve (i.e., τ versus ε). Such a curve is shown schematically in Fig. W10.5 for a typical FCC metal in the form of a single crystal. Beyond the elastic region which extends up to the *shear yield stress* τ_y, there can exist in some materials three inelastic regions, I, II, and III. The rate of work hardening in each region can be characterized by the slope $d\tau/d\varepsilon$, which is denoted by θ_{I}, θ_{II}, and θ_{III}, respectively. The higher the slope, the greater the rate at which work hardening occurs for a given increment in applied shear stress τ.

Although all may not be present in a given material, these regions have the following characteristics:

Region I. Plastic deformation in region I begins with the onset of "*easy glide*" or *slip* occurring on the primary slip system, as described in Section 10.14. A relatively low rate of work hardening occurs in region I. This region corresponds to the existence of long, straight slip lines in a single crystal. Region I is absent in polycrystals.

Region II. This is the linear work-hardening region, with $\theta_{\mathrm{II}} \approx 10\theta_{\mathrm{I}}$ and $\theta_{\mathrm{II}} \approx G/300$, where G is the shear modulus (i.e., the slope $d\tau/d\varepsilon$ in the elastic region). Plastic deformation in this region results in the interaction of dislocations and occurs via the mechanism of slip. The resulting distribution of dislocations is very inhomogeneous. The shear stress in region II is often observed to be proportional to the square root of the dislocation density ρ, that is,

$$\tau_y(\rho) = \tau_{y0} + \alpha G b\sqrt{\rho}. \tag{W10.9}$$

Here τ_{y0} is the shear yield stress (i.e., the shear stress needed to move a dislocation when no other dislocations are present), b is the Burgers vector, and α (≈ 0.3 to 0.6) is a constant. Note that ρ is given by the total length of all the dislocations divided by the volume of the material and has units of m^{-2}. It is clear from this expression that ρ is an increasing function of shear stress [i.e., $\tau_y(\rho) - \tau_{y0}$]. Typical values for single-crystal or polycrystalline Cu are $\rho \approx 10^{16}$ m^{-2} for $\tau_y \approx$ 100 MPa.

Region III. In this region the slope $d\tau/d\varepsilon$ decreases continuously with increasing stress, with the dependence of τ on ε usually observed to be close to parabolic, that is,

$$\tau(\varepsilon) = \theta_{\text{III}}\sqrt{\varepsilon - \varepsilon'}, \tag{W10.10}$$

where ε' is a constant.

Various theories can reproduce the form of Eq. (W10.9) observed in the linear region II or the parabolic dependence of τ on ε observed in region III. None of the theories of work hardening is completely satisfactory, however, which should not be surprising given the complexity of the problem. One of the first approaches, presented by Taylor, considered the source of work hardening to be the interactions between edge dislocations and the pinning that results. If l is the average distance that dislocations move before being pinned, the resulting shear strain ε corresponding to a dislocation density ρ is

$$\varepsilon = K\rho bl, \tag{W10.11}$$

where K is a constant that depends on orientation.

For a material containing a uniform distribution of edge dislocations, the average separation between the dislocations is $L \approx \rho^{-1/2}$. The applied shear stress required to move two dislocations past each other must overcome the effective internal stress acting on one dislocation due to the other. This can be written as

$$\tau = \frac{kGb}{L}, \tag{W10.12}$$

where k is a constant. Since $L \approx \rho^{-1/2}$, it follows that

$$\tau \approx kGb\sqrt{\rho}, \tag{W10.13}$$

which has the form of Eq. (W10.9). When Eqs. (W10.11) and (W10.13) are combined, the following dependence of τ on ε is obtained:

$$\tau(\varepsilon) \approx kG\sqrt{\frac{b\varepsilon}{Kl}} \approx k'G\sqrt{\frac{\varepsilon}{l}}, \tag{W10.14}$$

where k' is another constant. This prediction corresponds to the parabolic dependence of τ on ε observed in region III. The predictions of Taylor's theory therefore agree with the observed dependencies of τ on ρ and on ε despite the simplifying assumptions made, including the assumption of a uniform distribution of edge dislocations. Taylor's theory does not, however, explain the linear work hardening observed in region II.

W10.5 Strengthening Mechanisms

Dispersion Strengthening. *Dispersion strengthening* is a process in which small particles of a hard phase such as alumina (Al_2O_3) or silica (SiO_2) are distributed uniformly in the matrix of a weaker material (e.g., a copper alloy), either by precipitation in situ or by sintering the materials together. This process strengthens the

weaker host material and increases its resistance to plastic deformation. Dispersion-strengthened materials can have high hardness at high temperatures when the dispersed particles are of a refractory nature and very hard. This is an advantage of this strengthening method over precipitation hardening. The Orowan expression relating the yield stress σ_y to the interparticle spacing Λ is described in Chapter W21 with regard to the dispersion strengthening of steels

Precipitation Hardening. *Precipitation hardening* is a process in which a second phase is precipitated from a supersaturated solid solution in a matrix via heat treatment. Important examples include the precipitation of particles of Fe_3C or Fe_4N in iron and of particles of the intermetallic compound $CuAl_2$ in Al, as described in detail in Chapter W21. Both dispersion strengthening and precipitation hardening arise from short-range interactions between dislocations and the dispersed particles or the precipitate. As a result, the dislocations are pinned and cannot move freely through the material. The Orowan expression mentioned earlier is also applicable to these short-range interactions between dislocations and precipitate particles.

Long-range interactions between precipitate particles and dislocations are also possible due to the internal stresses created by the difference in average atomic volumes of the precipitate and the host matrix. Mott and Nabarro obtained the following estimate for the average shear strain ε_{av} in a single crystal due to a volume fraction f of spherical precipitate particles:

$$\varepsilon_{\text{av}} = 2\varepsilon f. \tag{W10.15}$$

Here $\varepsilon = \Delta r/r_0 = (r - r_0)/r_0$ is the fractional radial misfit resulting from the insertion of a particle of radius r in a cavity of radius $r_0 < r$ within the host matrix. The resulting strain leads to an increase in the critical shear yield stress by the amount

$$\Delta\tau_y = G\varepsilon_{\text{av}} = 2G\varepsilon f, \tag{W10.16}$$

where G is the shear modulus. According to this prediction, the critical shear yield stress should be independent of the particle sizes and interparticle separations. In fact, the precipitate particles will have little effect on the motion of the dislocations when the particles are small and closely spaced and also when they are large and far apart. Only at intermediate sizes and separations will they have a strong effect.

Solid-Solution Strengthening. An example of *solid-solution strengthening* is doubling of the yield strength of Fe–C solid-solution alloys at a C/Fe atom ratio of only $1/10^4$. As mentioned in Section 10.12, interstitial C atoms in octahedral sites cause tetragonal distortions of the BCC crystal structure of α-Fe. These lattice distortions in turn impede the motion of dislocations, thereby strengthening the Fe. This strengthening mechanism is described further for the case of steels in Chapter W21.

W10.6 Creep Testing

Typical creep tests at $0.5T_m < T < T_m$ and constant applied stress are shown in Fig. W10.6, where three distinct stages are shown for the dependence of the nominal strain on time. Results are shown at two applied stresses σ. It can be seen that the creep rate $\partial\varepsilon/\partial t$ is an increasing function of σ, as expected, and also of temperature T.

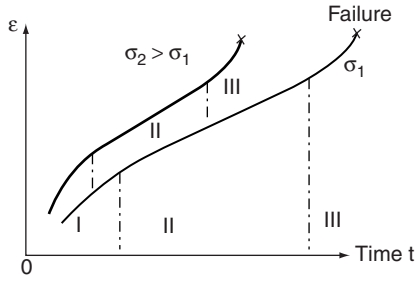

Figure W10.6. Typical creep test for $0.5T_m < T < T_m$ and constant applied stress. Three distinct stages are evident for the dependence of the nominal strain ε on time.

In stage I of *primary* creep the creep strain rate $\partial\varepsilon/\partial t$ actually slows down, probably as a result of work hardening, and reaches a value that typically remains constant in the most important stage II of *secondary* or *quasiviscous* creep. In stage III of *tertiary* creep the creep rate increases, nonuniform deformation begins, and failure eventually occurs. The *creep strength* of a material can be defined as the stress that will produce a given strain in a given time at a given temperature T. For example, a typical low-carbon nickel alloy has a creep strength of 60 MPa for $10^{-3}\%$ elongation per hour at $T = 534°\text{C}$. The stress for fracture σ_f due to creep is lower the longer the time of loading. Extrapolation of the results of creep tests to longer times is required for predicting the performance of materials in service (e.g., predicting when failure will occur under a given load or stress condition). This is due to the fact that creep tests generally do not extend to the point of failure, particularly when carried out at low stress levels and low temperatures.

Various models have been proposed to describe the dependencies of creep or the creep rate $\dot{\varepsilon} = \partial\varepsilon/\partial t$ on time, temperature, and stress. There is no universal model, but expressions such as

$$\varepsilon(t) = \varepsilon_0 + \varepsilon_p(1 - e^{-mt}) + \dot{\varepsilon}_s t, \tag{W10.17}$$

$$\frac{\partial\varepsilon}{\partial t} = A\sigma^n \exp\left(-\frac{Q_c}{k_B T}\right) \tag{W10.18}$$

have been proposed. In Eq. (W10.17), ε_0 is the initial strain in the material, the second term describes creep in stage I, and the term $\dot{\varepsilon}_s t$ (which is linear in time) represents stage II. Equation (W10.18) is proposed to be valid for the secondary creep rate in stage II, with A and n being constants and Q_c the thermal activation energy for creep. For a number of pure metals it has been found that $n = 5$ and that $Q_c \approx E_a(\text{diff})$, the measured thermal activation energy for self-diffusion in the metal.

A useful way of graphically illustrating the stress and temperature regions in which various deformation mechanisms are dominant (i.e., rate controlling) is the *Weertman–Ashby map*, shown in Fig. W10.7 for pure nickel. This map presents a plot of normalized tensile stress σ/G (where G is the shear modulus) versus T/T_m and corresponds to a critical strain rate $\dot{\varepsilon}_c$ of 10^{-8} s^{-1}. *Coble creep* and *Nabarro creep* correspond to diffusion of vacancies within the boundaries of the grains and within the bulk of the grains, respectively, and can be seen in Fig. W10.7 to be dominant in different regimes of temperature and stress.

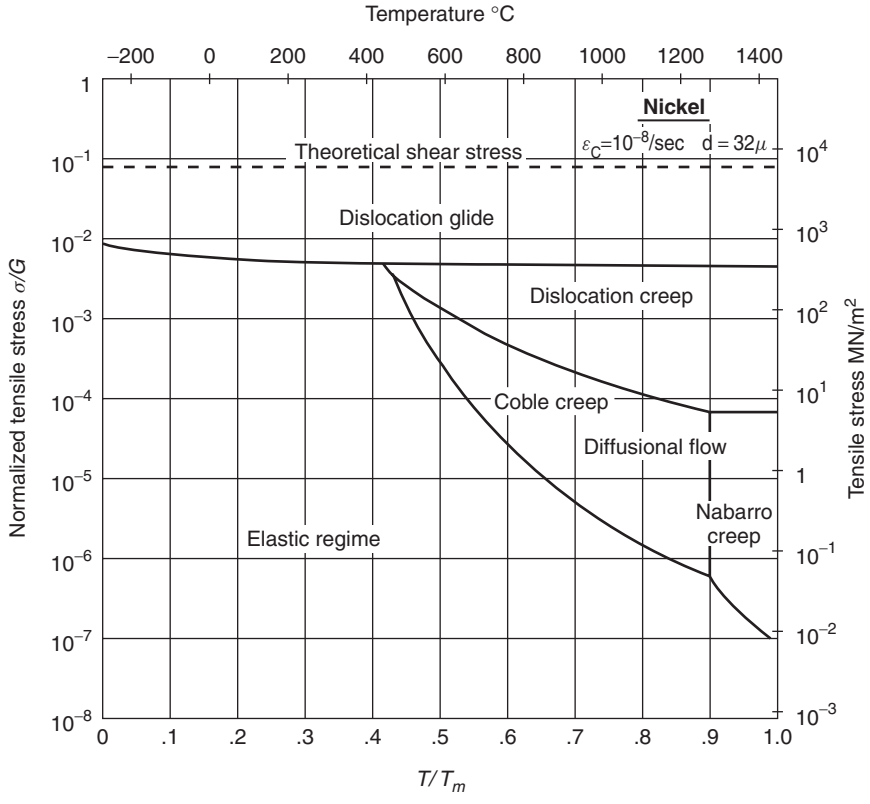

Figure W10.7. The Weertman–Ashby map presented here for pure nickel is a semilogarithmic plot of normalized tensile stress σ/G versus T/T_m for a critical strain rate $\dot{\varepsilon}_c$ of 10^{-8} s^{-1}. (Reprinted from *Acta Metallurgica*, Vol. 20, M. F. Ashby, p. 887. Copyright © 1972, by permission from Elsevier Science.)

W10.7 Further Discussion of Fatigue

When fatigue occurs under conditions of low true-stress amplitude σ_a, the response of the material is primarily elastic and the number of cycles to failure N_f is large. In this case the range $\Delta\varepsilon_e$ over which the elastic component of the strain varies can be described by

$$\Delta\varepsilon_e = \frac{2\sigma_a}{E} = \frac{2\sigma'_f}{E}(2N_f)^b, \tag{W10.19}$$

where b is the fatigue strength exponent and σ'_f is the fatigue strength coefficient, equal to the stress intercept for $2N_f = 1$. The quantity σ'_f is approximately equal to σ_f, the fracture stress under monotonic loading. The exponent b can be expressed in terms of the cyclic hardening coefficient n' by

$$b = -\frac{n'}{1+5n'}. \tag{W10.20}$$

Fatigue life thus increases with decreasing $|b|$, i.e. decreasing n'.

When fatigue occurs under conditions of higher stress amplitude σ_a and the response of the material has an inelastic or plastic component, the number of cycles to failure N_f will be smaller. The range of variation $\Delta\varepsilon_p$ of the plastic strain component can be described by the *Manson–Coffin relation*,

$$\Delta\varepsilon_p = 2\varepsilon'_f(2N_f)^c, \tag{W10.21}$$

where ε'_f, the ductility coefficient in fatigue, is equal to the strain intercept for $2N_f = 1$, and c is the ductility exponent in fatigue. Smaller values of c correspond to longer fatigue life. In the limit of high strain and low number of cycles c is given by

$$c = -\frac{1}{1+5n'}. \tag{W10.22}$$

As a result, fatigue life in this limit increases with increasing n'.

When a material is subjected under cyclic loading to both elastic and plastic strain, the fatigue strength will be determined by the total strain:

$$\Delta\varepsilon_t = \Delta\varepsilon_e + \Delta\varepsilon_p = \frac{2\sigma'_f}{E}(2N_f)^b + 2\varepsilon'_f(2N_f)^c. \tag{W10.23}$$

The separation of a $\Delta\varepsilon_t - N_f$ curve into its elastic and plastic components is illustrated schematically in Fig. W10.8. It can be seen that $\Delta\varepsilon_t$ approaches the plastic curve at high strain levels and the elastic curve at low strain levels.

W10.8 Hardness Testing

Hardness is often measured by the indentation of a harder material, typically a diamond indenter, into a softer material or by a scratch test. Indentation methods can be quantitative, while scratch testing gives essentially qualitative results. The most common methods of indentation hardness testing include the *Brinnell* and *Rockwell* tests and microindentation or microhardness tests such as the *Knoop* and *Vickers* tests. Hardness values are expressed using hardness scales with the same names. A common scale for

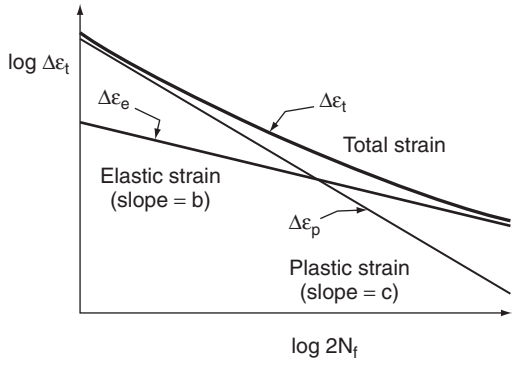

Figure W10.8. Separation of a $\Delta\varepsilon_t$–N_f fatigue curve into its elastic and plastic components.

minerals is *Mohs* hardness, determined by a scratch test, which extends from 1 for talc to 10 for diamond.

The Knoop hardness test is a microindentation test that uses an indenter in the form of an elongated pyramid while the Vickers test uses a square pyramid of diamond. The Knoop and Vickers hardnesses are defined as the ratio of the applied force or load to the surface area of the indentation. The Vickers hardness VHN is given by

$$\text{VHN} = \frac{1.854F}{d^2}, \tag{W10.24}$$

where F is the load in kilograms force (kgf) and d is the length of the diagonal of the square indentation in millimeters. Some Vickers hardness values for metals and other hard materials are given in Table 10.6. These hardness values, as with many other mechanical properties, are sensitive to processing treatments that the material may have received, especially those affecting the surface region.

The indentation of the Knoop indenter in the material under test is shallower than that of the Vickers indenter, thus making the Knoop method more appropriate for brittle materials and for thin layers. Because of the shallowness of the indentation, the surfaces of materials to be tested for Knoop hardness must be very smooth.

W10.9 Further Discussion of Hall–Petch Relation

The *Hall–Petch relation* was originally justified on the basis of the assumption that the effect of grain boundaries is to pin dislocations, but more recent interpretations emphasize the emission of dislocations by grain boundaries. An approach by Li[†] takes the onset of plastic deformation in polycrystalline materials as due to the activation of dislocation sources, which are assumed to be grain-boundary ledges. The shear yield stress for the motion of a dislocation relative to a distribution of other dislocations has been given in Eq. (W10.9) by

$$\tau_y(\rho) = \tau_y + \alpha G b \sqrt{\rho}, \tag{W10.25}$$

where ρ is the dislocation density and the other symbols are as defined earlier. If it is assumed that there is a uniform distribution of dislocation sources on the surfaces of all grain boundaries, regardless of their size, the dislocation density ρ will be proportional to S_v, the grain boundary area per unit volume. If the grains are all taken to be cubes of volume d^3, S_v will be given by

$$S_v = \frac{1}{2}\frac{6d^2}{d^3} = \frac{3}{d}, \tag{W10.26}$$

where the initial factor of $\frac{1}{2}$ accounts for the fact that each cube face (i.e., each grain boundary) is shared by two grains. The Hall–Petch relation of Eq. (10.43) is obtained when the result that $\rho \propto S_v \propto 1/d$ is used in Eq. (W10.25).

[†] J. C. M. Li, *Trans. TMS-AIME*, **227**, 239 (1963).

The yield stress can also be increased by solid-solution strengthening, as discussed in Section W10.5. The typical example is dilute alloys of C in BCC α-Fe, where $\sigma_y = \sigma_0 + k_y N_C^{1/2}$. Here N_C is the atomic fraction of C present in Fe.

W10.10 Analysis of Crack Propagation

When fracture occurs in a ductile material in which significant amounts of plastic deformation can occur, the critical stress will be increased above the prediction of Eq. (10.48) since the strain energy required for the generation of plastic deformation near the crack must be included. Plastic deformation of the material surrounding the crack tip can take the form of a dense array of dislocations and microcracks whose presence can slow down and even stop the propagation of the crack. The effective surface energy γ_p associated with the plastic deformation is equal to the work per unit area required to carry out the plastic deformation. When γ_p is added to γ_s in Eq. (10.48), *Griffith's criterion* in its general form becomes

$$\sigma_c = \sqrt{\frac{(2\gamma_s + \gamma_p)E}{\pi a}}. \tag{W10.27}$$

For many ductile materials $\gamma_p \gg \gamma_s$, so that

$$\sigma_c = \sqrt{\frac{\gamma_p E}{\pi a}} \tag{W10.28}$$

for the case of ductile fracture. The effect of the plastic deformation is to blunt the crack tip, thus relaxing the stress concentration there by increasing the local radius of curvature. As a result, ductile fracture requires higher stress levels than brittle fracture.

Correlations of *fracture toughness* K_{1c} with density ρ, Young's modulus E, and with strength σ_f for several classes of engineering materials (alloys, plastics, elastomers, composites, ceramics, glasses, etc.) have been presented by Ashby in the form of materials property charts.[†] These charts and the accompanying discussions are helpful in that they present and condense a large body of information and reveal correlations between the properties of materials. A striking feature of the charts is the clustering of members of a given class of materials. This clustering and the relative positions of the various clusters on the charts can be understood in terms of the type of bonding, the density of atoms, and so on, in the materials. Within each cluster the position of a given material can be influenced by the synthesis and processing that it receives. The following charts are also presented by Ashby: E versus ρ, σ_f versus ρ, E versus σ_f, and E/ρ versus σ_f/ρ.

The rate of elastic strain energy release by a crack is $G(\text{el})$, defined by

$$G(\text{el}) = -\frac{1}{2d}\frac{\partial \Delta U_{\text{el}}}{\partial a} = \frac{\pi \sigma^2 a}{E}. \tag{W10.29}$$

[†] M. F. Ashby, Materials Property Charts, in *ASM Handbook*, Vol. 20, ASM International, Materials Park, Ohio, 1997.

At the point of fracture $G(\text{el}) = G_c(\text{el})$ and the critical fracture stress can therefore be expressed in terms of $G_c(\text{el})$ by

$$\sigma_c = \sqrt{\frac{EG_c(\text{el})}{\pi a}}. \tag{W10.30}$$

By comparing this result with Eqs. (W10.27) and (10.49), it can be seen that

$$K_c = \sqrt{EG_c(\text{el})}. \tag{W10.31}$$

The quantity $G_c(\text{el})$ is also known as the *critical crack extension force*, with units of N/m.

REFERENCE

Gilman, J. J., *Micromechanics of Flow in Solids*, McGraw-Hill, New York, 1969.

PROBLEMS

W10.1 A bar of a solid material undergoes two consecutive deformations along the x axis corresponding to nominal normal strains ε_1 and ε_2, as defined by $\varepsilon_1 = (x_1 - x_0)/x_0$ and $\varepsilon_2 = (x_2 - x_1)/x_1$.

(a) Show that these two nominal strains are not additive [i.e., that $\varepsilon_{\text{total}} = (x_2 - x_0)/x_0 \neq \varepsilon_1 + \varepsilon_2$].

(b) Show, however, that the corresponding true strains $\varepsilon_{\text{true}}(1)$ and $\varepsilon_{\text{true}}(2)$, as defined in Eq. (10.8), are additive.

(c) Find the difference between ε and $\varepsilon_{\text{true}}$ for $\Delta l = 0.1 l_0$.

W10.2 From the expressions given for the shear modulus G and the bulk modulus B in Table 10.4, show that Poisson's ratio ν for an isotropic solid must satisfy $-1 < \nu < \frac{1}{2}$.

W10.3 Derive the expression for the elastic energy density $u_{\text{el}}(\varepsilon)$ for a cubic crystal given in Eq. (10.32).

W10.4 Using the general definitions for strains as $\varepsilon_1 = \partial u_x/\partial x$, $\varepsilon_5 = \partial u_x/\partial z + \partial u_z/\partial x$, and so on, show that the equation of motion, Eq. (10.35), can be written as the wave equation given in Eq. (10.36).

W10.5 Consider the values of E, G, B, and ν given in Table 10.2 for several polycrystalline cubic metals.

(a) Show that the values of E, G, and ν are consistent with the expressions for isotropic materials given in Table 10.4.

(b) Show that the same cannot be said for the values of B.

W10.6 If the changes in stress and strain in a material occur so rapidly (e.g., at sufficiently high frequencies) that no relaxation occurs, show that the stress/strain ratio is given by the unrelaxed elastic modulus, $E_u = E_r \tau_\sigma / \tau_\varepsilon$.

W10.7 **(a)** For the conditions shown in Fig. 10.9*a* after relaxation has occurred, derive the solutions of Eq. (W10.3) presented in Eq. (W10.4).

(b) Also derive the analogous equations for the time dependence of σ for the conditions shown in Fig. 10.9*b*.

W10.8 Let σ_0 be real and set $\varepsilon_0 = \varepsilon_{00}e^{-i\phi}$ in Eq. (W10.5) so that the strain $\varepsilon(t)$ lags behind the stress $\sigma(t)$ by a phase angle ϕ. Using these expressions (i.e., $\sigma(t) = \sigma_0 \exp(-i\omega t)$ and $\varepsilon(t) = \varepsilon_{00} \exp[-i(\omega t + \phi)]$), in Eq. (W10.6), show that $\tan\phi$ is given by Eq. (W10.8).

W10.9 The relaxation time τ for a piece of cross-linked natural rubber is 30 days at $T = 300$ K.

(a) If the stress applied to the rubber at $T = 300$ K is initially 1 MPa, how long will it take for the stress to relax to 0.5 MPa?

(b) If the relaxation time for the rubber at $T = 310$ K is 20 days, what is the activation energy E_a for the relaxation process? See Eq. (10.41) for the definition of E_a.

W10.10 Repeat Problem 10.9 for the (0001), (1100), and $(10\bar{1}0)$ planes of HCP Cd and for the three $\langle 11\bar{2}0\rangle$ directions in the (0001) plane.

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:

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

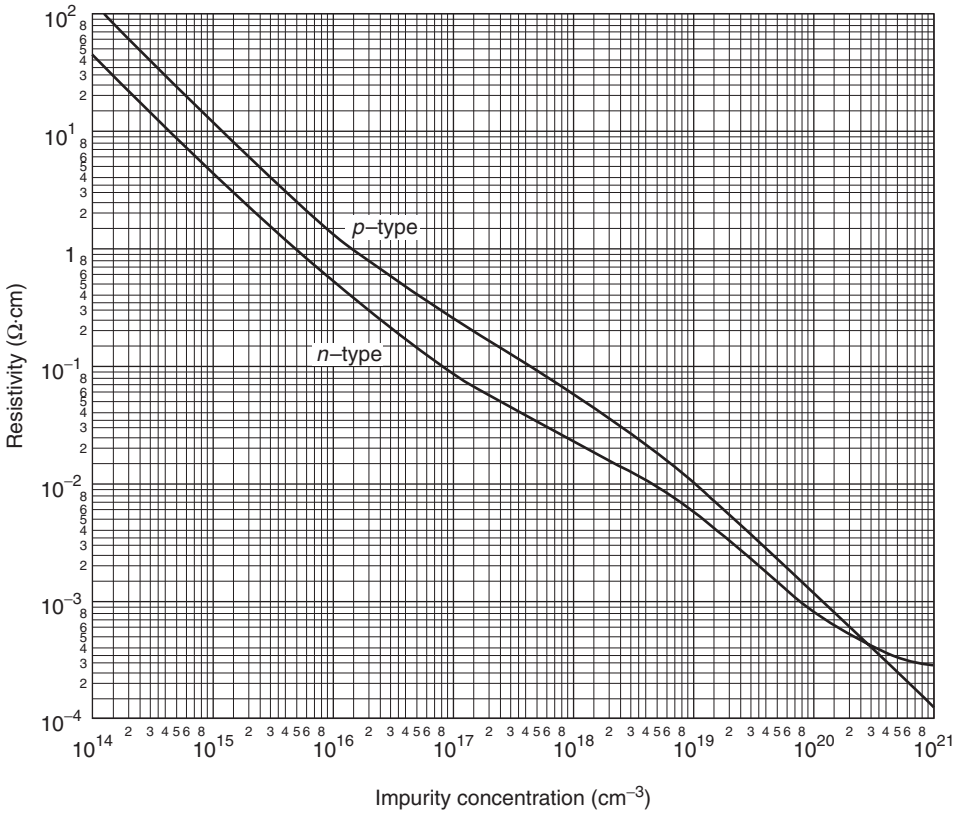

Figure W11.1. Effects of n- and p-type doping on the electrical resistivity of Si at $T = 300$ K, with ρ plotted versus the dopant concentration on a logarithmic plot. (From J. C. Irvin, *The Bell System Technical Journal*, **41**, 387 (1962). Copyright © 1962 AT&T. All rights reserved. Reprinted with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

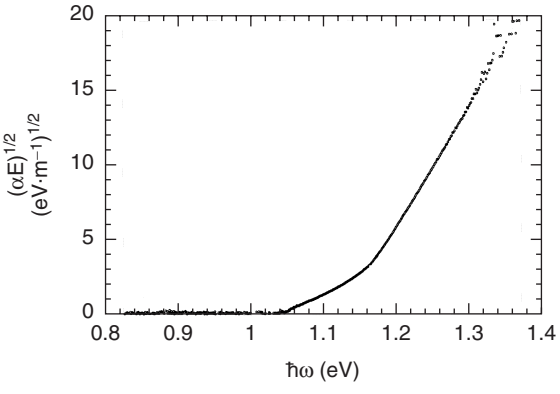

Figure W11.2. Optical absorption edge for Si at $T = 300$ K with the absorption coefficient α plotted as $(\alpha\hbar\omega)^{1/2}$ versus the photon energy $E = \hbar\omega$. The energy gap $E_g = 1.11$ eV and the energy of the phonon $\hbar\omega_{\text{phonon}} \approx 0.06$ eV participating in this 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

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

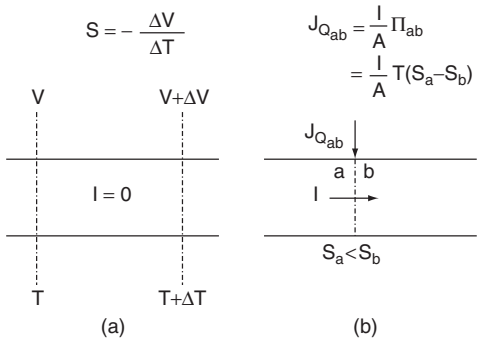

Figure W11.3. Seebeck and Peltier effects. (*a*) In the Seebeck effect a voltage difference ΔV exists in a material due to the temperature difference ΔT. The Seebeck coefficient or thermopower of the material is given by $S = -\Delta V/\Delta T$. (*b*) In the Peltier effect a flow of heat into (or out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

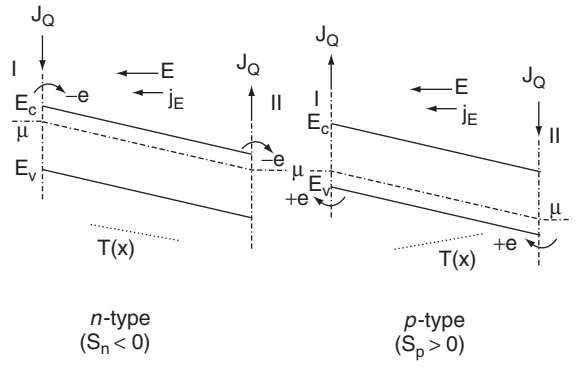

Figure W11.4. Peltier effect in a semiconductor. An electric field **E** is applied across a semiconductor, and as a result, the energy bands and the chemical potential μ slope downward from left to right. In the n-type semiconductor, electrons flow from left to right and in the p-type semiconductor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

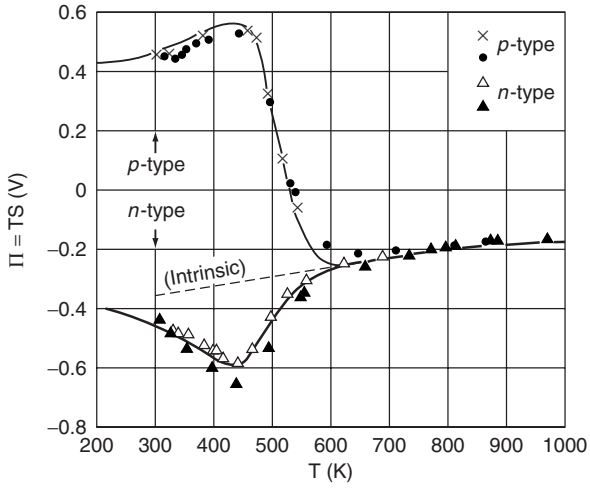

Figure W11.5. Experimental results (points) and theoretical predictions (solid lines) for the Peltier coefficient Π for n- and p-type Si are shown as functions of temperature. The Si samples show intrinsic behavior above $T \approx 600$ K. (From T. H. Geballe et al., *Phys. Rev.*, **98**, 940 (1955).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

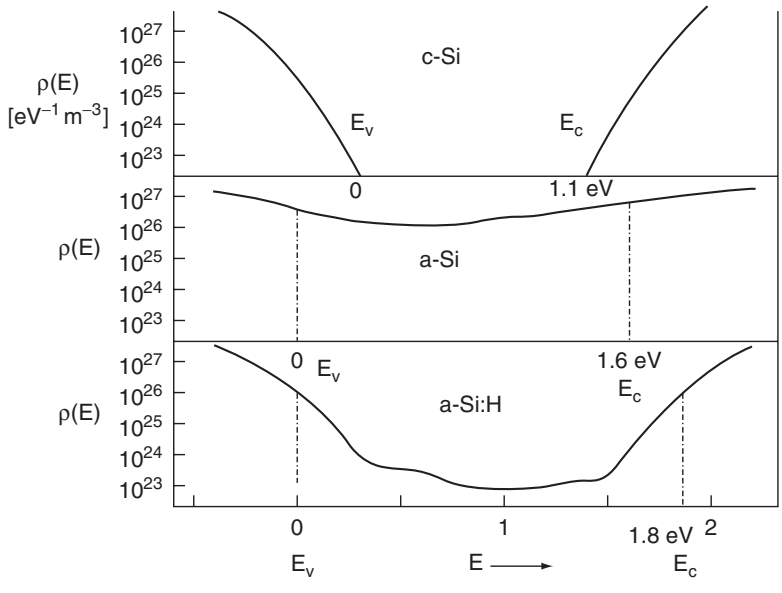

Figure W11.6. Electron densities of states in crystalline Si, a-Si, and a-Si:H in the region of the energy gap.

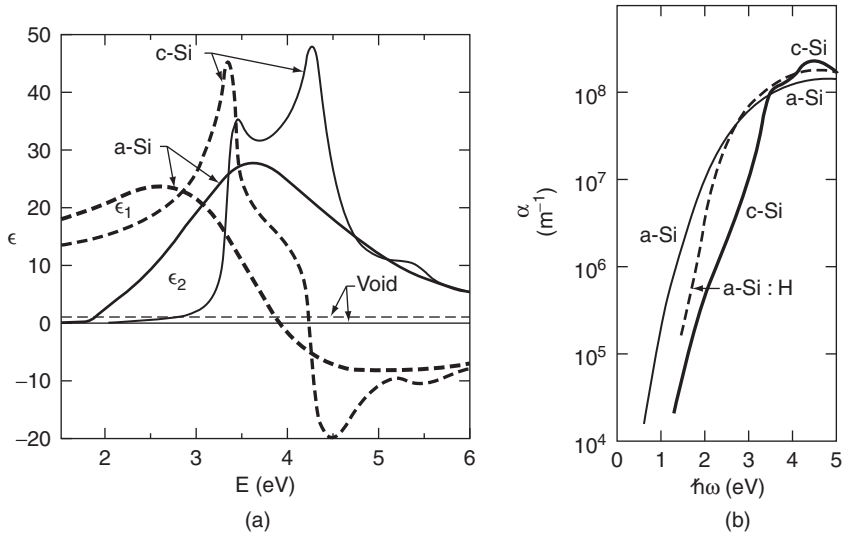

Figure W11.7. Comparison of the optical properties of crystalline and amorphous Si. (*a*) The quantities ϵ_1 (dashed lines) and ϵ_2 (solid lines) of c-Si and a-Si are plotted versus photon energy $E = \hbar\omega$. (From B. G. Bagley et al., in B. R. Appleton and G. K. Celler, eds., *Laser and Electron-Beam 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

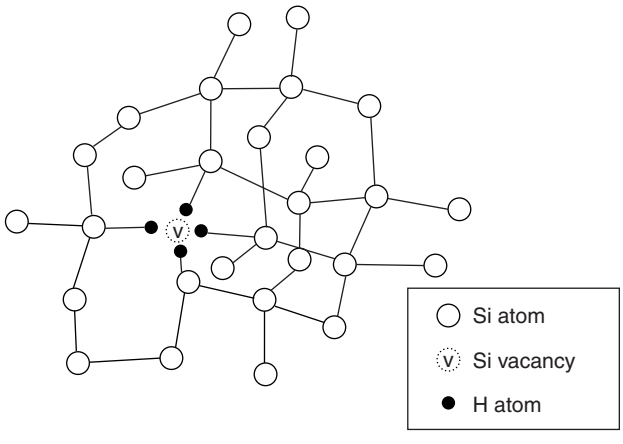

Figure W11.8. Model of a segment of the continuous random network corresponding to the bonding in a-Si:H. Four H atoms are shown completing the Si bonds at a Si monovacancy.

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.

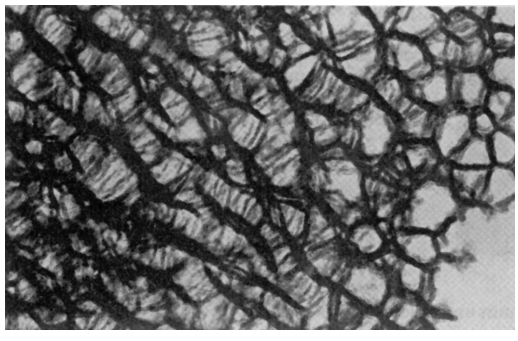

Figure W11.9. Transmission electron micrograph of porous Si in which the Si columns are about 10 nm in diameter and the pore spaces are about 50 nm wide. (Reprinted with permission of A. G. Cullis. From R. T. Collins et al., *Phys. Today*, Jan. 1997, p. 26.)

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

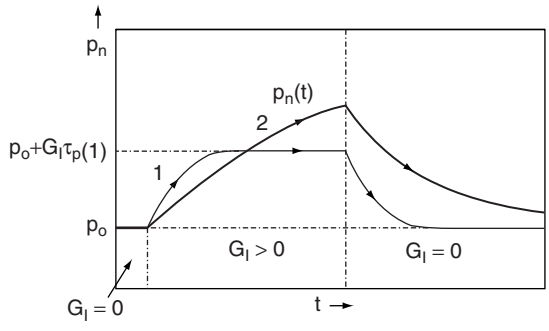

Figure W11.10. Buildup and decay of the minority-carrier hole concentration $p_n(t)$ in an n-type semiconductor under low-level carrier injection for two different minority-carrier lifetimes, with $\tau_p(1) < \tau_p(2)$.

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).

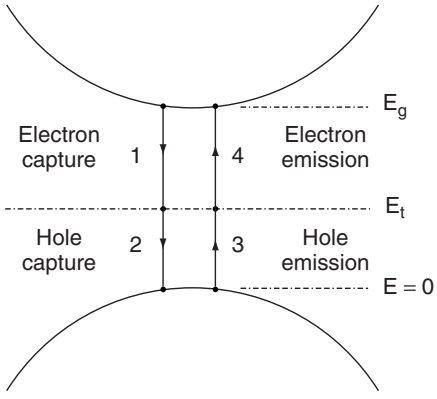

Figure W11.11. Possible transitions involving electrons and holes and resulting from a defect level at the energy E_t in the gap. 1, Capture of an electron; 2, capture of a hole; 3, emission of a hole; 4, emission of an electron.

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

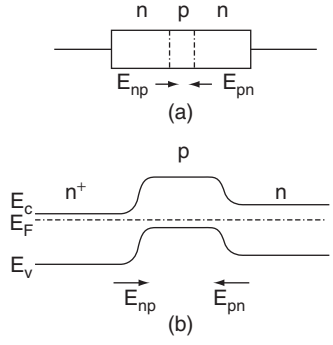

Figure W11.12. An *npn* bipolar junction transistor: (*a*) directions of the built-in electric fields at the two junctions; (*b*) electron energy bands across the transistor at zero bias.

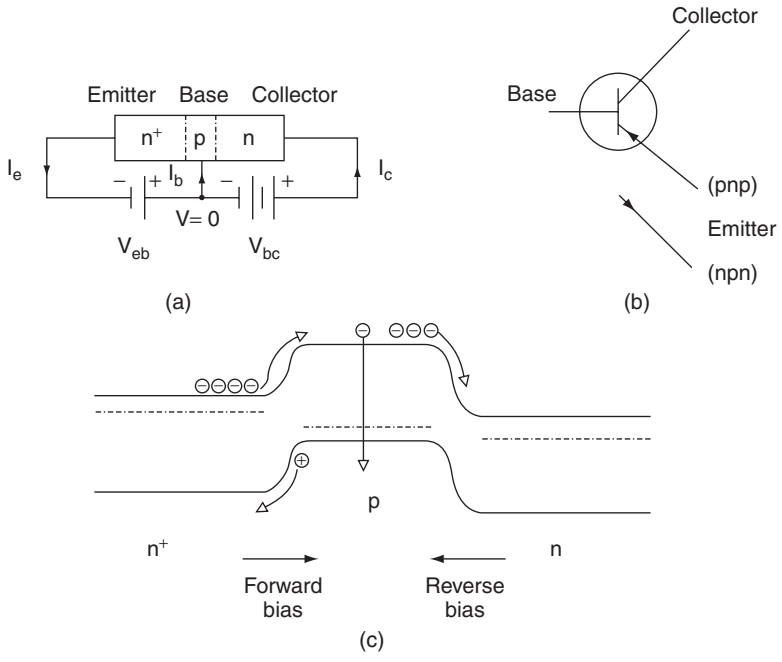

Figure W11.13. Operation of an *npn* transistor. (*a*) The emitter-base *n*-*p* junction is forward biased, while the base–collector *p*-*n* junction is given a stronger reverse bias. The directions of the three resulting currents I_e, I_b, and I_c for the emitter, base, and collector are shown. (*b*) 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.

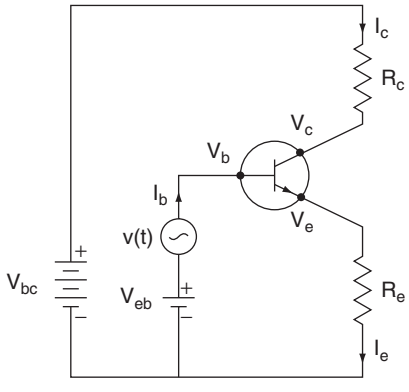

Figure W11.14. Simplified circuit illustrating the use of an *npn* transistor as an amplifier of a small ac voltage $v(t)$. The dc voltage sources V_{bc} and V_{eb} provide the biasing of the two junctions and the source of the input signal $v(t)$ appears in the base circuit.

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

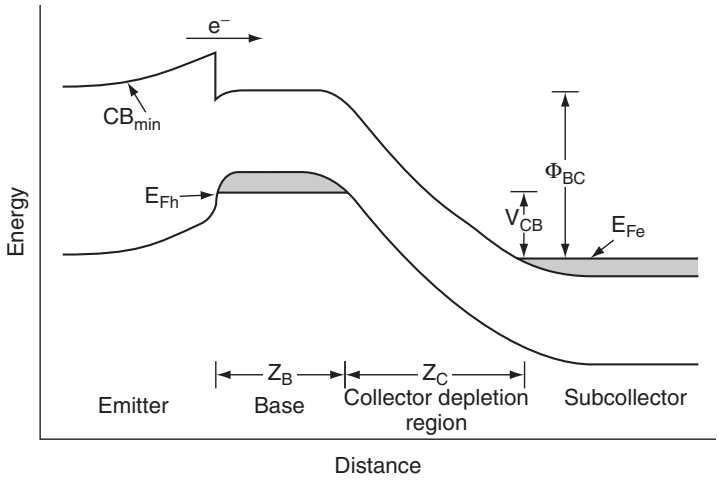

Figure W11.15. Electron energy-band diagram for a heterojunction bipolar transistor (HBT). In the *npn* HBT shown here the emitter has a wider bandgap than the base and collector semiconductors. A potential barrier exists in the valence band that hinders the back injection of holes from the *p*-type 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

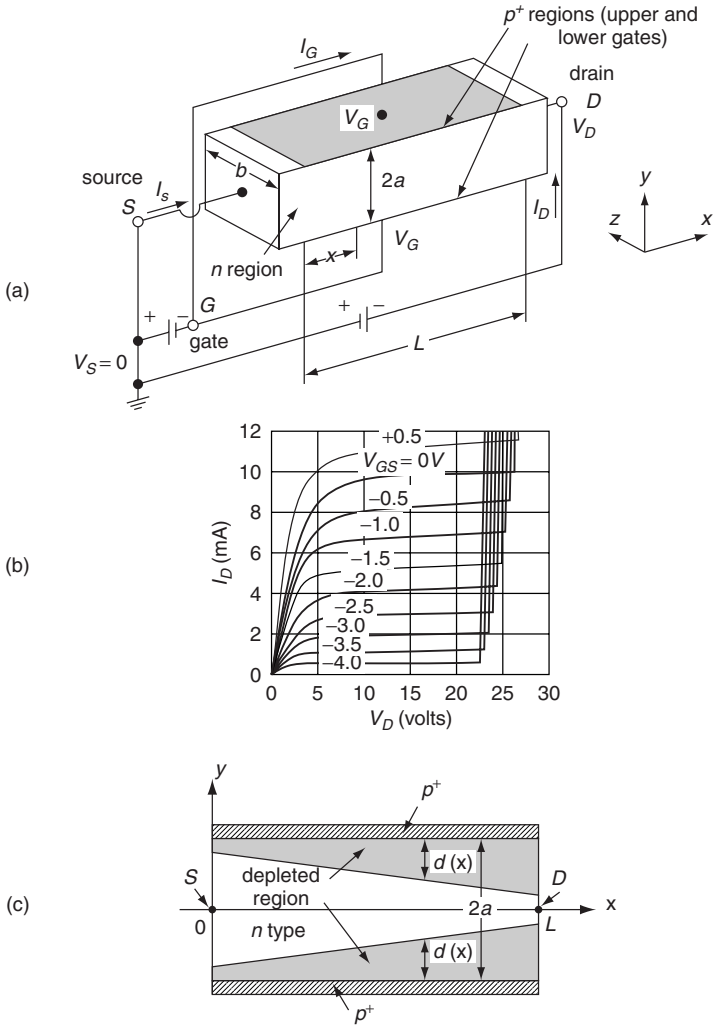

Figure W11.16. Properties of a junction FET. (*a*) Configuration of a junction FET in a rectangular bar of n-type Si. The two metallic electrodes at the ends of the bar are the source and drain, and the conducting channel between them is controlled by the p-type gates at the center of the bar. (*b*) 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.

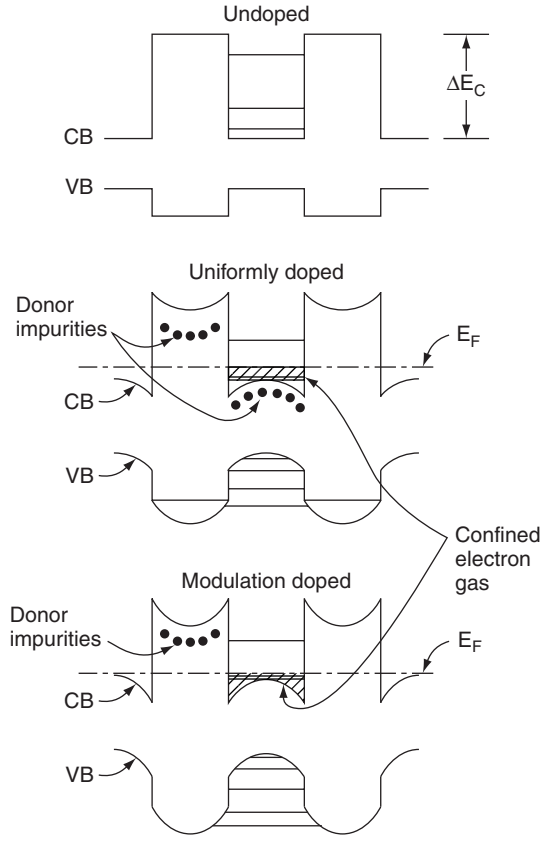

Figure W11.17. Modulation doping in GaAs-$Al_xGa_{1-x}As$ superlattices. The carriers are spatially separated from the charged scattering centers associated with the dopant impurity ions. (From R. Dingle et al., *Appl. Phys. Lett.*, **33**, 665 (1978). Copyright © 1978 by the American Institute of Physics.)

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

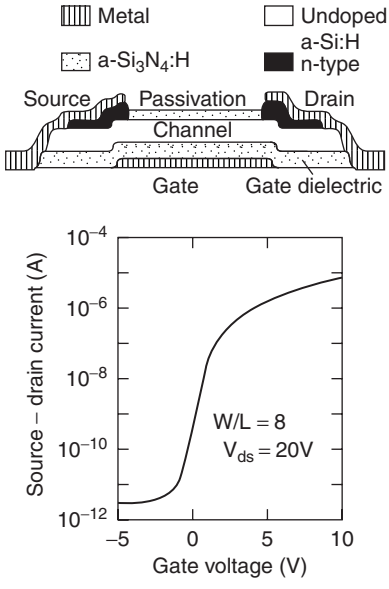

Figure W11.18. Common configuration of an a-Si:H field-effect TFT, along with its source-to-drain current I_d versus gate voltage V_g transfer characteristic. (From R. A. Street, *Mater. Res. Soc. Bull.*, **17**(11), 71 (1992).)

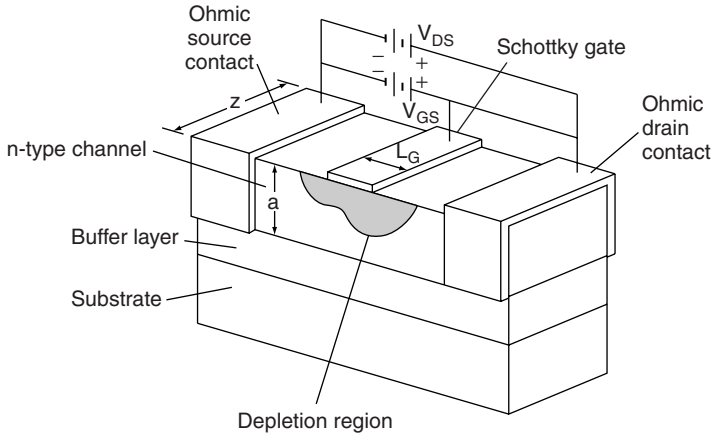

Figure W11.19. SiC metal–semiconductor field-effect transistor (MESFET). The gate configuration in the MESFET consists of a rectifying metal–semiconductor Schottky barrier at the surface of a doped, epitaxial layer of SiC. (From K. Moore et al., *Mater. Res. Soc. Bull.*, **23**(3), 51 (1997).)

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).

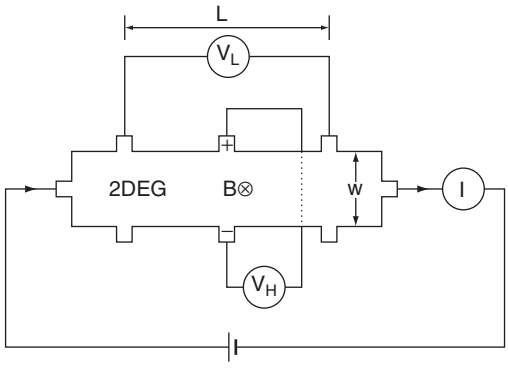

Figure W11.20. Geometry of the measurement of the quantum Hall effect for the two-dimensional electron gas.

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

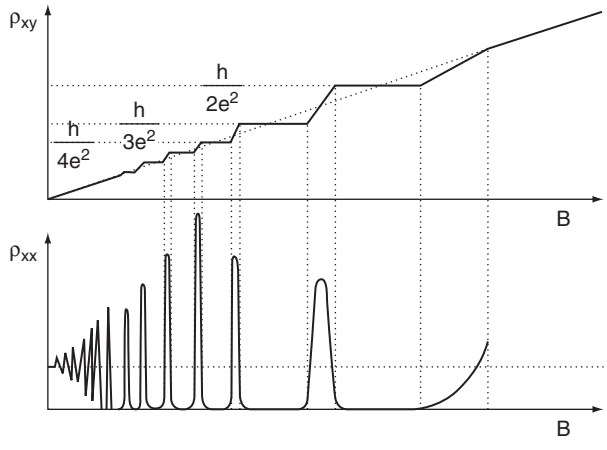

Figure W11.21. Experimental results for the Hall resistivity ρ_{xy} and magnetoresistivity ρ_{xx} for the two-dimensional electron gas. (Reprinted with permission of H. Iken. Adapted from B. I. Halperin, The quantized Hall effect, *Sci. Am.*, Apr., 1986, p 52.)

† 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

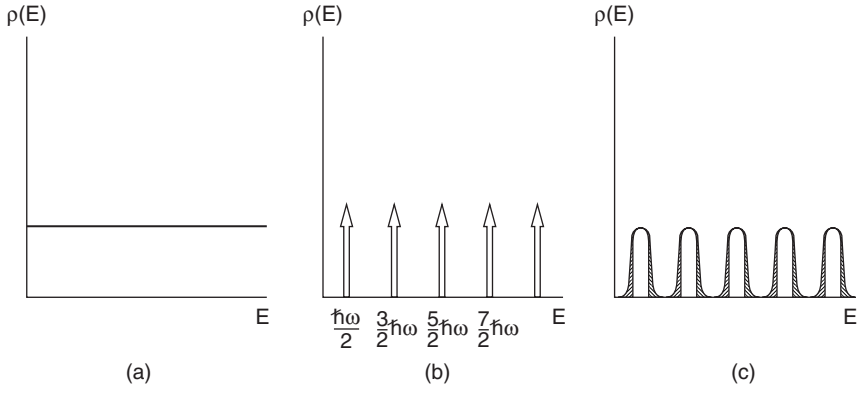

Figure W11.22. Density of states for a two-dimensional electron gas: (*a*) in the absence of a magnetic field; (*b*) in the presence of a magnetic field, but with no disorder; (*c*) in the presence of a magnetic field and with disorder. The smaller Zeeman spin splitting of the Landau levels is not shown.

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).

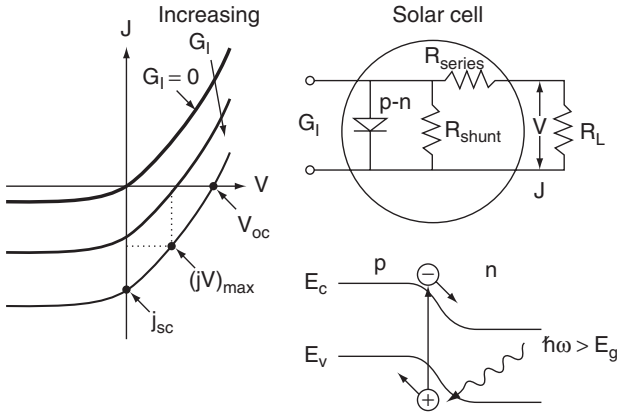

Figure W11.23. Predicted current–voltage characteristics for a photovoltaic solar cell in the form of a *p-n* junction, both in the dark ($G_I = 0$) and illuminated ($G_I > 0$), shown schematically when the solar cell is connected to an external circuit. The generation rate of photo-excited electron–hole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

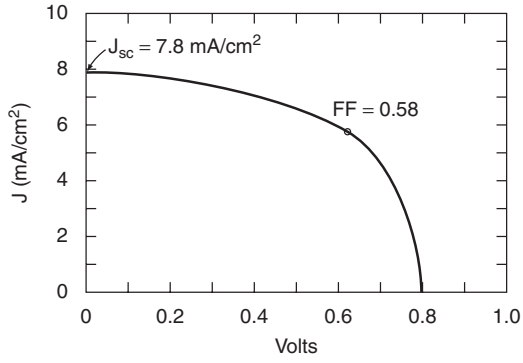

Figure W11.24. Typical J–V curve for an a-Si:H Schottky-barrier solar cell under illumination of 650 W/m^2. (From M. H. Brodsky, ed., *Amorphous Semiconductors*, 2nd ed., Springer-Verlag, New York, 1985.)

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

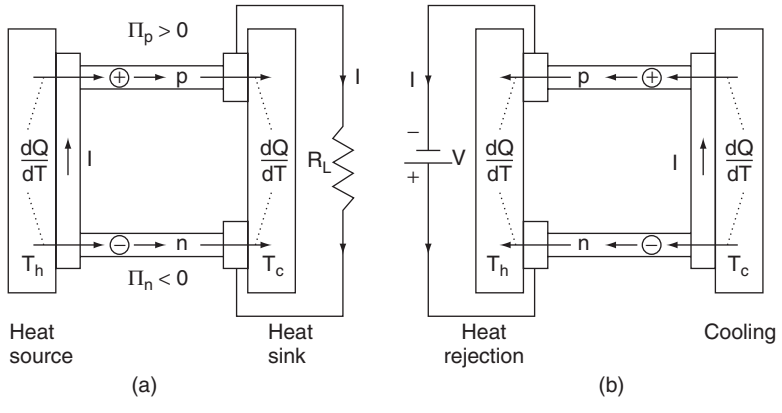

Figure W11.25. Schematic diagrams of (*a*) a thermoelectric power generator and (*b*) a thermoelectric refrigerator. In the thermoelectric generator or thermopile two semiconductors, one n-type and the other p-type, each carry a heat flux from a heat source to a heat sink. In the thermoelectric refrigerator 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.

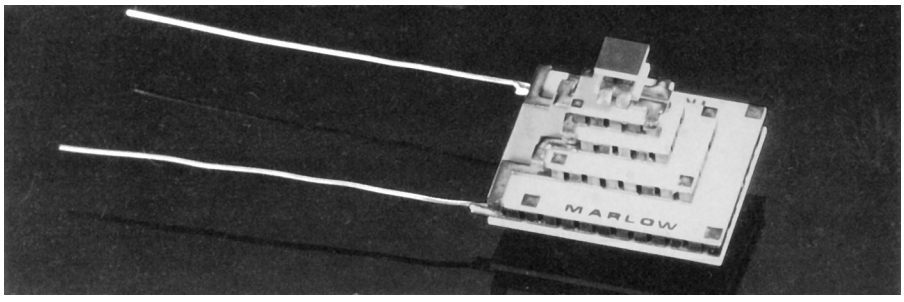

Figure W11.26. Cascaded thermoelements are employed in thermoelectric generators and refrigerators, as shown in the cooling module pictured here. (From G. Mahan et al., *Phys. Today*, Mar. 1997, p. 42. Copyright © 1997 by the American Institute of Physics.)

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

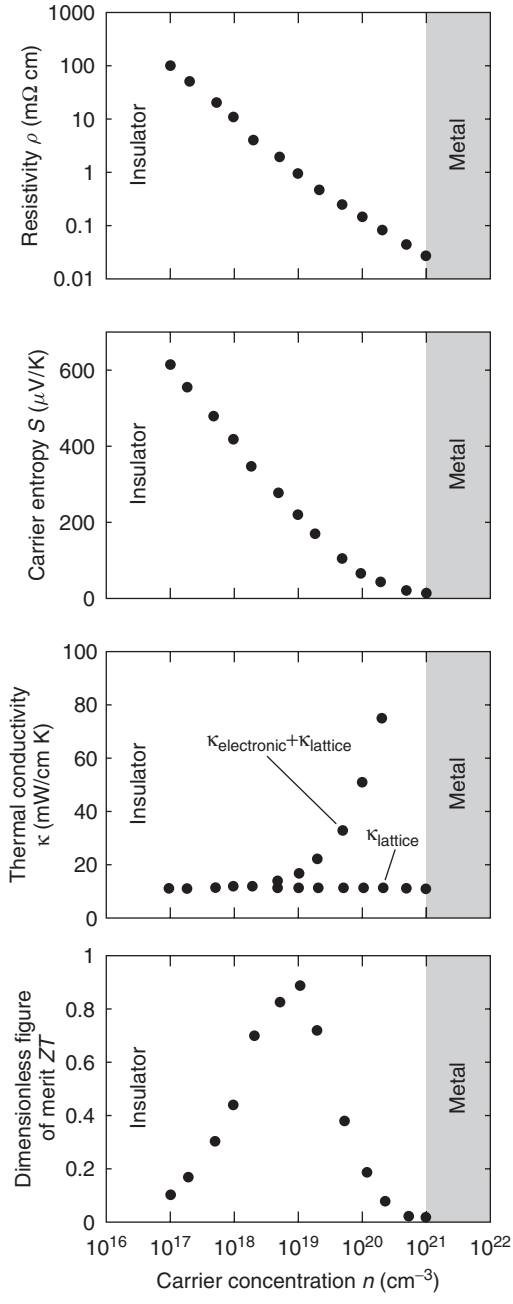

Figure W11.27. Effects of changing the carrier concentration on the thermoelectric parameter $Z = S^2/\rho\kappa$ and the values of S (the thermopower or carrier entropy), ρ, and κ for an idealized semiconductor. The energy gap E_g increases to the left in this figure. (From G. Mahan et al., *Phys. Today*,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

CHAPTER W12

Metals and Alloys

A variety of theoretical tools is available for the study of metallic solids. Electronic band-structure methods include the augmented plane wave (APW) method, the orthogonalized plane wave (OPW) method, the Green function [Korringer, Kohn, and Rostoker (KKR)] method, the pseudopotential method, and the cellular (Wigner–Seitz) method. These approaches are discussed in solid-state physics textbooks (e.g., Fletcher or Ashcroft and Mermin). These methods all rely on the perfect periodicity of the solid and utilize Bloch's theorem to limit the focus of attention to a unit cell. They are not directly applicable to disordered alloys or solids with impurities or defects.

Quantum-chemistry calculations can be done for clusters of finite size, but the computational time grows rapidly as the size of the cluster is increased, making such calculations impractical for the study of large collections of atoms with present-day computers.

The next three sections introduce methods that have found some utility in describing realistic solids: the density-functional method, the embedded-atom method, and the tight-binding approximation. Although lacking the accuracy of the band-structure or quantum-chemistry computations, they are nevertheless useful in studying large-scale systems, are relatively simple to implement on the computer, and are, for many purposes, adequate.

W12.1 Density-Functional Theory

Density-functional theory is a method currently being used to obtain a theoretical understanding of metals, metallic alloys, surfaces of metals, and imperfections in metals. The method is a natural outgrowth of the Thomas–Fermi method introduced in Chapter 7 of the textbook.[†] It is based on the observation by Hohenberg and Kohn that all the ground-state properties of a many-body quantum-mechanical system of electrons may be obtained from a knowledge of the electron density, $n(\mathbf{r})$. They proved that $n(\mathbf{r})$ determines the potential $V(\mathbf{r})$ that the electrons move in, up to an insignificant additive constant. Furthermore, an energy functional $E[n]$ may be constructed and it may be shown to attain its minimum value when the correct $n(\mathbf{r})$ is employed.

The uniqueness proof is based on the minimum principle from quantum mechanics. Begin by noting that if the potential energy function $V(\mathbf{r})$ were known, one could solve

[†] 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."

the Schrödinger equation and obtain the electron density $n(\mathbf{r})$. If there were two different potentials $V(\mathbf{r})$ and $V'(\mathbf{r})$ leading to the same $n(\mathbf{r})$, the Schrödinger equation could be solved for each potential and the respective ground-state wavefunctions ψ and ψ' would be determined. By the minimum principle, the ground-state energy obeys the inequality

$$E = \langle\psi|(T+V)|\psi\rangle < \langle\psi'|(T+V)|\psi'\rangle = \langle\psi'|(T+V')|\psi'\rangle + \langle\psi'|(V-V')|\psi'\rangle$$
$$= E' + \langle\psi'|(V-V')|\psi'\rangle = E' + \int n(\mathbf{r})[V(\mathbf{r}) - V'(\mathbf{r})]\,d\mathbf{r}. \tag{W12.1}$$

Repeating the argument with the primed and unprimed variables interchanged leads to $E' < E + \int n(\mathbf{r})[V'(\mathbf{r}) - V(\mathbf{r})]\,d\mathbf{r}$. Adding the two inequalities leads to the contradiction $E + E' < E' + E$. Q.E.D.

The energy of the system is written in the form

$$E[n] = \int n(\mathbf{r}) \left[\frac{3}{5}E_F(\mathbf{r})\right] d\mathbf{r} + \int n(\mathbf{r})V(\mathbf{r})\,d\mathbf{r} + E_{ii}$$
$$+ \frac{1}{2}\frac{e^2}{4\pi\epsilon_0}\int d\mathbf{r}\int d\mathbf{r}'\frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + E_{\mathrm{xc}}[n]. \tag{W12.2}$$

Here $E_F = \hbar^2 k_F^2/2m$, where $k_F(\mathbf{r}) = [3\pi^2 n(\mathbf{r})]^{1/3}$ is a local Fermi wave vector, and $V(\mathbf{r})$ is the potential due to the ions. The first four terms are the kinetic energy, the energy of interaction of the electrons with the ions, the ion–ion interaction, and the Coulomb repulsion energy of the electrons. The quantity E_{xc} is the energy arising from exchange and correlation effects. The variational problem may be reduced to the solution of a set of partial-differential equations called the Kohn–Sham equations. These are of the form

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r}) - E_j\right]\psi_j(\mathbf{r}) = 0, \tag{W12.3}$$

where $E_{xc}[n] = \int n\epsilon_{xc}\,d\mathbf{r}$ and

$$v_{\mathrm{eff}}(\mathbf{r}) = V(\mathbf{r}) + \frac{e^2}{4\pi\epsilon_0}\int\frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + v_{\mathrm{xc}}(\mathbf{r}), \tag{W12.4}$$

$$v_{\mathrm{xc}}(\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}[n(\mathbf{r})]}{\delta n(\mathbf{r})}. \tag{W12.5}$$

The electron density is constructed from the Kohn–Sham wavefunctions as

$$n(\mathbf{r}) = \sum_{j=1}^{N}|\psi_j(\mathbf{r})|^2. \tag{W12.6}$$

In the local-density approximation (LDA) it is assumed that E_{xc} depends only on n and not on its derivatives, and one writes

$$v_{\mathrm{xc}} \approx \frac{d}{dn}(n\epsilon_{\mathrm{xc}}). \tag{W12.7}$$

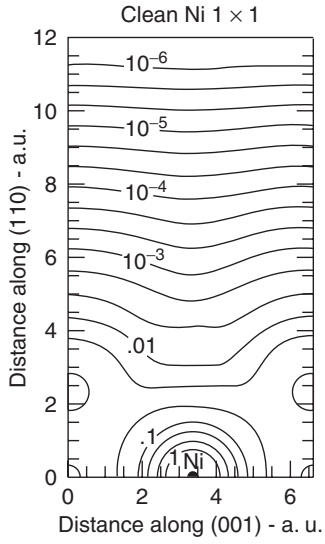

Figure W12.1. Surface-charge density for Ni. Distance is measured in atomic units (a.u.). [Adapted from D. R. Hamann, *Phys. Rev. Lett.*, **46**, 1227 (1981). Copyright 1981 by the American Physical Society.]

Various research groups have presented useful functional forms for $\epsilon_{\rm xc}(n)$. The results of the calculations of $n(\mathbf{r})$ generally compare favorably with experiment or with quantum-chemistry calculations for finite systems. Density-functional theory has also been extended to include corrections involving ∇n terms. An example of calculational results for the surface-charge density of Ni is given in Fig. W12.1.

W12.2 Embedded-Atom Method

The embedded-atom method attempts to calculate the energy of realistic metals by making simplifying assumptions about how atoms interact with each other and with the common sea of electrons. The energy is written as a sum of two terms

$$E = E_{\rm rep} + E_{\rm embed}. \tag{W12.8}$$

The first term is the interatomic-repulsive energy associated with the nuclei and their core electrons. The repulsive energy is given by the sum of pairwise potentials:

$$E_{\rm rep} = \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}} U_{ij}(\mathbf{R}_{ij}). \tag{W12.9}$$

The second term is the interaction of the atoms with the electron gas in which the atoms find themselves embedded. The embedding energy is approximated as the sum of the energies of interaction of each atom with a *uniform* electron gas. The electron density at site i is computed by superimposing the local electronic densities from all

other atoms. Thus

$$E_{\text{embed}} = \sum_i F_i \left[\sum_j {}' n_j(\mathbf{R}_i - \mathbf{R}_j) \right]. \tag{W12.10}$$

The embedding energy, $F_i(n_0)$, is computed using density-functional theory. A point charge ze is placed at the origin. The jellium model is used for the electron gas. The charge density is given by $\rho(\mathbf{r}) = e[n_0 + z\delta(\mathbf{r}) - n(\mathbf{r})]$. Detailed calculations were carried out for a number of elements.† Typical results are presented in Fig. W12.2. Values for the densities at which the minimum occurs and the corresponding well depths are presented in Table W12.1.

Often $F_i(n_0)$ is approximated by a function of the form

$$F_i(n_0) = A_i n_0 - B_i \sqrt{n_0}. \tag{W12.11}$$

The first term corresponds to the effect of the filled shells of the ion. For example, in the inert gases, where all the shells are filled, the embedding energy is observed to grow approximately linearly with the electron density, with a slope given by A_i. The second term arises from the bonding of the valence electrons of the atom with the ambient electrons. If the volume of the embedded atom is Ω, the number of electrons that the atom overlaps with is $N = n_0\Omega$. In a tight-binding description, in which each ambient electron is assigned to a neighboring site, one would could construct a wavefunction as a superposition of the form $|\psi\rangle = (|1\rangle + \cdots + |N\rangle)/\sqrt{N}$, where each term represents a state localized on a given site. The tunneling-matrix element linking the atom to the ith neighbor would be of the form $t = \langle\psi_0|V|i\rangle/\sqrt{N}$. A band whose width is given by $2Nt$ would form. If the state at the bottom of that band is occupied, this would result in a reduction of energy $\Delta E_i = -\langle\psi_0|V|i\rangle\sqrt{N} \equiv -B_i\sqrt{n_0}$. It is interesting to note that the metallic bond is unsaturated (i.e., only part of the band is occupied). If the full band were occupied, the band energy would not be reduced and B_i would be zero.

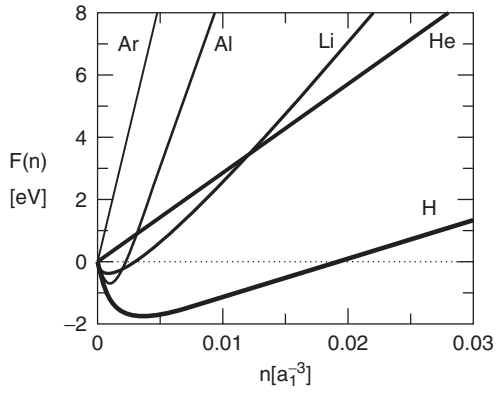

Figure W12.2. Embedding energy as a function of electron density for several elements. Here a_1 is the Bohr radius. [Adapted from M. J. Puska, R. M. Nieminen, and M. Manninen, *Phys. Rev. B*, **24**, 3037 (1981). Copyright 1981 by the American Physical Society.]

† M. J. Puska, R. M. Nieminen, and M. Manninen, *Phys. Rev. B*, **24**, 3037 (1981).

TABLE W12.1 Position and Depth of the Minimum of the Embedding Energy

Atom	n_0 $(a_1^{-3})^a$	$F(n_0)$ (eV)
H	0.0026	−1.8
He	0	—
C	0.0035	−1.8
N	0.0045	−1.4
O	0.0037	−4.1
F	0.0010	−5.1
Ne	0	—
Na	<0.0005	<−0.6
Al	0.0005	−0.2
Cl	0.0005	−4.0

Source: Data from M. J. Puska, R. M. Nieminen, and M. Manninen, *Phys. Rev. B*, **24**, 3037 (1981).

[a] a_1 = Bohr radius = 0.0529 nm.

The embedded-atom method allows rapid computation of the ground-state energy of a configuration of many atoms. By varying the atomic positions it is possible to search for the minimum energy. Such quantities as the lattice constants, cohesive energy, elastic constants, and surface energies could be obtained, as well as information concerning the effects of impurities and defects.

W12.3 Peierls Instability

As an example of the utility of the tight-binding method, this section is devoted to a special phenomenon that occurs when a one-dimensional metal is constructed. With the trend toward miniaturization proceeding at the pace that it is, such a situation is not out of the realm of the possible. When the Fermi surface of an electron gas approaches certain special points in the Brillouin zone, structural instabilities may result. The special points could lie at boundaries of the Brillouin zones or could lie within the zone. Peierls showed that in a one-dimensional solid, a half-filled band results in an instability that converts the metal into an insulator. The instability produces a dimerization of adjacent atoms and doubles the size of the unit cell.

The model is depicted in Fig. W12.3, where the lattice is shown before and after dimerization. The lattice will be idealized by a tight-binding model in which the atoms are connected by springs of spring constant k_s. Prior to dimerization the electronic

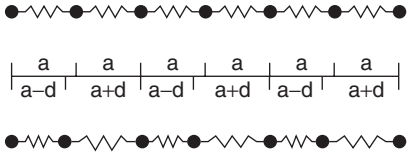

Figure W12.3. One-dimensional solid, before and after dimerization due to the Peierls instability.

energies are given by [see Eq. (7.81)]

$$E(k) = E_0 - 2t\cos ka, \tag{W12.12}$$

where E_0 is the site energy and t is the tunneling-matrix element. After dimerization two bands appear, with the respective energies

$$E_\pm = E_0 \pm \sqrt{2(t^2+\Delta^2) + 2(t^2-\Delta^2)\cos 2ka} \tag{W12.13}$$

where the tunneling-matrix elements for the springs of length $a \pm d$ have been written as $t \mp \Delta$. It is assumed that for small d the shift in Δ is proportional to d (i.e., $\Delta = \alpha d$). The lower band is occupied and the upper band is empty, so the solid becomes an insulator.

The total energy per unit length consists of the sum of the electronic energy and the elastic energy. Its change is given by

$$\frac{\delta U}{L} = \sum_s \int_{-\pi/2a}^{\pi/2a} \frac{dk}{2\pi}\left[2t\cos ka - \sqrt{2(t^2+\Delta^2) + 2(t^2-\Delta^2)\cos 2ka}\right] + \frac{k_s d^2}{2a}. \tag{W12.14}$$

The integral is expressible in terms of *E[m]*, the complete elliptic integral of the second kind,

$$\frac{\delta U}{L} \approx -\frac{2\Delta^2}{\pi at}\left(\ln\frac{4t}{\Delta} - \frac{1}{2}\right) + \frac{k_s\Delta^2}{2a\alpha^2}. \tag{W12.15}$$

For small Δ the result may be written as

$$\frac{\delta U}{L} = \frac{4t}{\pi a}\left[1 - E\left(1 - \frac{\Delta^2}{t^2}\right)\right] + \frac{k_s\Delta^2}{2a\alpha^2}. \tag{W12.16}$$

For small-enough Δ this will be negative, predicting that the instability will always occur. Minimizing δU with respect to Δ leads to

$$\Delta = 4t\exp\left[-\left(1 + \frac{\pi k_s\alpha^2 t}{4}\right)\right], \tag{W12.17}$$

with

$$\frac{\delta U}{L} = -\frac{16t}{\pi a}\exp\left[-2\left(\frac{\pi k_s\alpha^2 t}{4} + 1\right)\right]. \tag{W12.18}$$

Peierls instabilities are believed to play a role in solids constructed from linear organic molecules such as polyacetylene.

W12.4 Corrosion and Oxidation

Corrosion occurs because metals in contact with ionic solutions often function as electrodes of batteries. To see how this comes about, consider the energy needed to

extract an atom, A, from a metal in contact with a solution, and to ionize it, resulting in the ion, A^{z+}, of charge state z, and z electrons

$$A \longrightarrow A^{z+} + ze^{-}. \tag{W12.19}$$

First the cohesive energy of the atom, E_{coh}, must be provided to remove the atom from the solid into the vacuum. Then the free-space ionization energy, IE, must be added to create the ion A^{z+} in vacuum. Upon placing the charges back into solution, the solvation energy of the ion, $U_i(A^{z+})$, is regained, as well as the solvation energy of the z electrons, zU_e. Dividing this by the electronic charge, $-e$, gives a possible expression for the standard potential for the electrode half-reaction:

$$V(A \longrightarrow A^{z+} + ze^{-}) = -\frac{E_{coh} + IE - U_i(A^{z+}) - zU_e}{e}. \tag{W12.20}$$

In practice only a relative scale for the standard potential is defined. The standard potential is determined experimentally relative to a standard reaction, usually taken as that for $H_2 \rightarrow 2H^+ + 2e^-$. The standard potential V is arbitrarily defined to be zero for this reaction.

As an example of a battery, consider the Daniell cell (Fig. W12.4). Two metals, Zn and Cu, are in contact with electrolytic solutions of $ZnSO_4$ and $CuSO_4$, respectively. These metals are connected to each other electrically through some external conduction path. The electrolytes are separated from each other by a saturated salt bridge, which selectively permits passage of the SO_4^{2-} ions but blocks the passage of Cu^{2+} and Zn^{2+} ions. At the anode, Zn undergoes the oxidation reaction $Zn \rightarrow Zn^{2+} + 2e^-$, with Zn^{2+} ions going into solution and electrons going into the external circuit. The reduction reaction $Cu^{2+} + 2e^- \rightarrow Cu$ occurs at the cathode, where Cu^{2+} ions are deposited on the electrode as they recombine with circuit electrons. The net result is that the Zn corrodes and the Cu gets plated. The potential difference of this cell is computed from the difference of the standard potentials, determined by the half-reactions taking place at the electrodes:

$$Zn \longrightarrow Zn^{2+} + 2e^{-}(+0.76\ V), \qquad Cu^{2+} + 2e^{-} \longrightarrow Cu\ (-0.34\ V) \tag{W12.21}$$

and is 1.1 V. The larger this voltage, the larger the ionic current will be (according to Ohm's law), and the faster the corrosion of the Zn will be. For materials with smaller standard potential differences, the corrosion would be slower. If the sign difference

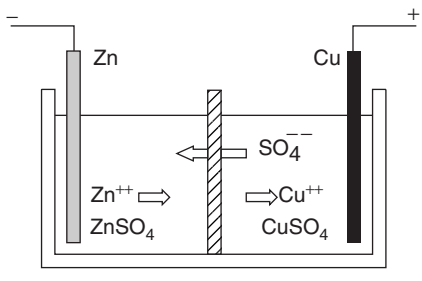

Figure W12.4. Daniell cell.

were negative instead of positive, no battery action, and consequently no corrosion, would occur. For example, if Zn were replaced by Ag, the oxidation half-reaction would be

$$2\mathrm{Ag} \longrightarrow 2\mathrm{Ag}^{+} + 2e^{-}(-1.6\ \mathrm{V}) \tag{W12.22}$$

and the standard difference would be -1.26 V, so no battery action would occur.

It is important to relate the electrode processes to the thermodynamic energies involved. The reaction $\mathrm{Cu} \rightarrow \mathrm{Cu}^{2+} + 2e^{-}$ (aqueous) involves a change of Gibbs free energy $\Delta G = -15.66$ kcal/mol $= -0.680$ eV, and the reaction $\mathrm{Zn}^{2+} + 2e^{-} \rightarrow \mathrm{Zn}$ (aqueous) has $\Delta G^0 = -35.14$ kcal/mol $= -1.525$ eV (at $T = 25^{\circ}\mathrm{C}$). The net Gibbs free energy change for the reaction is the sum of these and is -2.205 eV. Since two electrons are transferred per reaction, $z = 2$, so the open-circuit electromotive force (EMF) is $\mathscr{E}^0 = \Delta G/(-ze) = 1.10$ V. In a battery the electrical energy is supplied from the change in Gibbs free energy of the constituents.

The overall reaction for the Daniell cell may be written as $\mathrm{Zn} + \mathrm{Cu}^{2+} \rightleftharpoons \mathrm{Zn}^{2+} + \mathrm{Cu}$. For standard conditions ($T = 25^{\circ}\mathrm{C}$, $P = 1$ atm) the EMF is determined by ΔG^0. However, conditions are usually not standard and the appropriate Gibbs free energy change is

$$\Delta G = \Delta G^0 + Nk_BT \ln \frac{a_{\mathrm{Zn}^{2+}} a_{\mathrm{Cu}}}{a_{\mathrm{Cu}^{2+}} a_{\mathrm{Zn}}}, \tag{W12.23}$$

where N is the number of atoms transferred and a_i refers to the activity of species i. The EMF becomes

$$\mathscr{E} = \mathscr{E}^0 - \frac{k_BT}{ze} \ln \frac{a_{\mathrm{Zn}^{2+}}}{a_{\mathrm{Cu}^{2+}}} = \mathscr{E}^0 - \frac{k_BT}{ze} \ln \frac{a_{\mathrm{ZnSO_4}}}{a_{\mathrm{CuSO_4}}}, \tag{W12.24}$$

since $a_{\mathrm{Cu}} = a_{\mathrm{Zn}} = 1$ (by definition). Since the activities are approximately proportional to the concentrations, as the concentration of Cu^{2+} drops, so does the EMF of the cell.

It should be noted that there are similarities between electrolytic solutions and semiconductors. In the electrolyte charge is carried by the ions, whereas in the semiconductor the carriers are electrons and holes. The standard potentials of electrolytes replace the bandgap potentials of semiconductors.

Next consider a piece of iron with a drop of water on it. The outer surface of the drop is assumed to be in contact with air. Oxygen is absorbed into the water, and a concentration gradient is established with the part of the water in contact with the iron relatively deficient in oxygen. Some of the iron is oxidized and goes into solution according to the reaction

$$\mathrm{Fe} \longrightarrow \mathrm{Fe}^{2+} + 2e^{-}(+0.44\ \mathrm{V}) \tag{W12.25}$$

with the electrons entering the metal across the electrolyte–metal interface. Near the outer boundary of the water–iron interface, the oxygen is reduced by accepting the two electrons from the metal and combining with solvated protons (hydronium ions, often denoted by $\mathrm{H_3O^+}$) in solution, according to either of the two reactions

$$\begin{aligned} &\tfrac{1}{2}\mathrm{O_2} + 2\mathrm{H_3O^+} + 2e^{-} \longrightarrow 3\mathrm{H_2O}\ (+0.615\ \mathrm{V}), \\ &2\mathrm{H_3O^+} + 2e^{-} \longrightarrow 2\mathrm{H_2O} + \mathrm{H_2}\ (+1.23\ \mathrm{V}). \end{aligned} \tag{W12.26}$$

In the first case the standard potential difference is 0.175 V and in the second case it is 0.79 V. In both cases the difference is positive, so the reaction can proceed. The net result is that iron is corroded from the metal. In solution the iron ions combine with oxygen to precipitate as rust. The rust (hydrated Fe_2O_3) is deposited on the metal surface as a porous material, so additional water can come in contact with the iron.

The pH of an aqueous solution is a measure of the concentration of hydronium ions and is defined by $pH = -\log_{10} n_{H_3O^+}$, with n given in units of moles per liter (mol/L). Nernst noted that the half-potentials are dependent on the pH of the water, and shift downward with increasing pH. Thus the acidity or basicity of the electrolyte can have a strong effect on the corrosion process.

Two strategies for eliminating corrosion present themselves. One is to coat the metal with a protective overlayer and thus block ionic flow. The second is to try to alloy the metal to make its oxidation potential more negative. It is noteworthy that gold, with its standard potential for the reaction $Au \rightarrow Au^{3+} + 3e^-$ at -1.50 V, is the most negative of the elements and is therefore the most "noble" of them all. This may be understood in terms of Eq. (W12.20), because the ionization energy of Au is high (9.22 eV) and the ionic radius is large (0.137 nm), which implies that the solvation energy U_i will be small.

The extent of damage caused by corrosion is more dependent on the morphology of the oxide than on the metals themselves. It is worth contrasting the oxidation of Fe discussed above with the oxidation of Al. In the latter case the Al_2O_3 layer that is produced forms a crystal on the surface of the Al and remains in registry with the substrate. For additional oxygen atoms to come in contact with the Al, they must first diffuse through the oxide layer. Although this is possible, especially at elevated temperatures, it becomes more and more difficult as the oxide layer builds up. Thus the oxidation process becomes self-arresting. For this reason, Al_2O_3 is called a *passivation layer* in electronics application. The process of depositing such a layer, called *anodization*, is discussed further in Section 19.11. In the iron case the porous nature of the rust permits the corrosion to continue until all the iron is consumed. Chromium is added to steel to form stainless steel. A passivation layer of Cr_2O_3 is formed. It should be noted that the standard potential for the electrode reaction $Cr^{3+} + Fe = Fe^{3+} + Cr$ is -0.93 V, which is quite negative and implies that Cr_2O_3 is more likely to be produced than Fe_2O_3.

Differences in potential may exist even for a grain of single crystal between different faces, or between the surface and the interior, and these may act as the driving force for battery action and corrosion. Stress differentials across a material may also produce potential differences. This makes metals with microcracks vulnerable to corrosion.

W12.5 Coatings

The surface of a metal or alloy is often modified by applying a coating or by building the coating directly into the surface. There are numerous reasons why this is done, including enhancement of corrosion resistance (CR), wear resistance (WR), fatigue resistance (FR), oxidation resistance (OR), and thermal resistance (TR), reducing the coefficient of friction, or enabling an electric contact to be made. For example, integrated circuits based on Si have TiN and Ti deposited on them as diffusion-barrier metal films. One may also want to increase adhesion, use the surface as a catalyst, or endow the surface with special optical properties.

Traditional methods for applying coatings included such techniques as electroplating and chemical reactions. Modern materials for these coatings include SiC, TiC, TiN, TiB_2, WC, W_2C, AlN, CrN, and Si_3N_4. Coating techniques include sputtering, chemical vapor-deposition (CVD) at high temperatures (800 to 1000°C), physical vapor deposition (PVD) at lower temperatures (250 to 500°C), energetic ion implantation, and thermal reactions.

Thin coatings ($\approx$ 10 μm) of SiC, TiC, TiN, Cr_7C_3, CrN, ZrC, or ZrN are applied to tools to improve their WR and ability to cut, and where high levels of microhardness are needed. Even diamond films, the hardest substance available, and the best thermal conductor at room temperature, can be CVD-coated onto tools. The hardest coatings are made of Si_3N_4, SiC, and TiB_2.

Coatings are used in ultrahigh-vacuum systems because of their low sticking coefficients for adsorbing gases, their low yield of secondary electrons (which are ejected from a metal following the impact of a primary electron or ion), and the absence of long-lived electronic excitations, which could result in photodesorption processes. In addition, they prevent ultraclean metal parts from fusing together via the formation of diffusion bonds, in which atoms from one metal migrate over to intermediate positions between the two metals to form bridging bonds.

The coefficient of friction is often reduced substantially by applying a coating. The metals Ag, Au, or Pb may be applied to steel as a lubricant. When there is frictional heating, the coating melts and acts as a lubricant. A layer of Ti applied to steel lowers the coefficient of sliding friction. Lowering friction proves to be of considerable importance in the fabrication of semiconductors, where there are moving parts that insert, position, and remove the wafers from the vacuum system. As these parts move, there is friction. Associated with the friction is wear, and as particles are broken off, the semiconductor can become contaminated. Since liquid lubricants are of no use in a vacuum system, coatings are used instead.

There can also be improved resistance to corrosion. Typically, 50-μm layers are used. Protection is afforded by such coatings as alumina, NiCr, SiC, and CoCr. Chromium, Ni, Ta, and Ti are applied to steel and Pd or Pt are applied to Ti for this purpose. A combination of Co, Cr, Al, and Y is applied to Ni alloys. The CR is due, in part, to the dense granular structure, which tends to be equiaxed (hexagonally tiled). This presents to the surrounding electrolytic medium a material of uniform electronegativity. It also serves as an obstacle for diffusion of oxygen into grain boundary channels in the underlying metal. Yttrium coated on steel or Cr on Cu inhibits oxidation, and ZrO_2 improves the OR of Ni alloys.

Ion implantation produces a high density of interstitials, dislocations, and other defects near the surface which can act as traps for other dislocations and therefore harden the material and improve the WR. The compounds BN, CrN, SiC, Si_3N_4, TiC, TiN, ZrC, and ZrN are used to harden steels.

Electrical contacts may be deposited on Si using Ag, Al, Pt, or Au coatings. For GaAs, Al coatings may be employed, and for alumina, Cu coatings are used. The formation of silicides of Pt, Pd, and Ti on Si creates Schottky barriers, which serve as rectifiers with small forward-biased impedance.

An alloy of Co, Ni, Cr, Al, and Y acts to provide a high degree of OR for use in such applications as jet turbines. Thermal-insulation layers are often used in conjunction with these, in which case they are called thermal-barrier coatings. The goal is

to achieve low thermal diffusivity ($\kappa/\rho c_p$). Materials for TR include MgO, Y_2O_3, and ZrO_2, which have low thermal conductivities and moderate heat capacities and densities.

W12.6 Shape-Memory Alloys

It is possible to start with a hot metallic object of a particular shape, cool it, distort it, and remove the external stress, to produce what will appear to be a plastically deformed object. At a later time, however, the object may be reheated and it will return to its original shape. The ability to revert to the original shape provides the name for this class of metals — shape-memory alloys (SMA). Underlying this "talent" lies some interesting physics. Typical SMA materials include the alloys FePt, FeNiC, NiFeAlB, AuCd, NiAl, NiTi, and CuZnAl. There are also SMA materials composed of ceramic materials (e.g., PbLaZrTiO).

The SM alloys are ordered and exist in two crystalline phases. The low-temperature phase is called *martensite* (M) and the high-temperature phase is called *austenite* (A). These names stem from the nomenclature used in steel metallurgy. More generally, the high-T phase may be called the *parent phase* and the low-T phase the *daughter phase*, although here the symbols A and M are used. Phase A has a higher degree of symmetry than phase M. There is a phase transition governing the A $\leftrightarrow$ M transformation (from A to M, and vice versa). This is illustrated in Fig. W12.5, where the volume is plotted against temperature. Plots of other physical quantities, such as electrical resistance, are similar in structure and show hysterisis. Suppose that one starts in the M phase and heats the sample. At a temperature T_{A_s}, one begins to form some austenite. The amount of A formed depends on $T - T_{A_s}$. At temperature T_{A_f}, one will have reached 100% A. Above that temperature the A material is simply heated. If one then cools the sample, at a temperature T_{M_s}, one begins creating the M phase. At temperature T_{M_f}, this conversion is complete, and below T_{M_f} there is 100% M. Note the presence of a small hysteresis loop. Typical values of these temperatures for some SMA materials are given in Table W12.2.

Figure W12.6 shows the A and M phase unit cells for the NiAl intermetallic compound. The A phase has the higher-symmetry CsCl structure, while the M phase has the lower-symmetry tetragonal structure (four atoms per unit cell). The phase

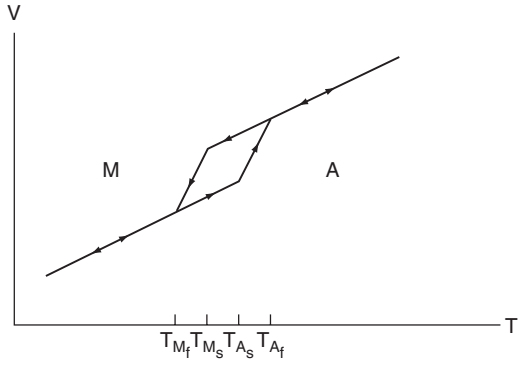

Figure W12.5. Variation of volume with temperature for a shape-memory alloy. Various critical temperatures described in the text are indicated.

TABLE W12.2 Start and Finish Temperatures for the Austenite (A) and Martensite (M) Phases of Some Shape-Memory Alloys

Shape-Memory Alloy	Temperature (°C)			
	T_{A_s}	T_{A_f}	T_{M_s}	T_{M_f}
$Au_{49.5}Cd_{50.5}$	40	42	37	35
$Zn_{25.75}Al_{4.01}Cu_{70.24}$	20	45	30	−5
$Zn_{25.60}Al_{3.90}Cu_{70.50}$	78	90	83	62
$Al_{23.9}Ni_{4.2}Cu_{71.9}$	35	80	71	26
$Ni_{58.9}Fe_{13.98}Al_{26.95}B_{0.17}$	93	172	127	56
$Ti_{50}Pd_{22}Ni_{28}$	201	252	200	107

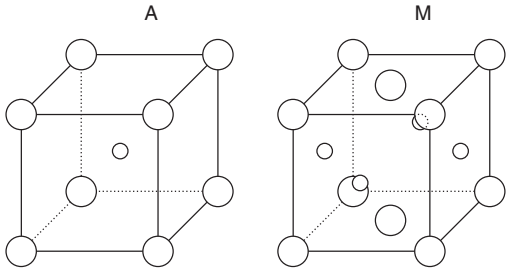

Figure W12.6. Example of the austenite and martensite unit cells in NiAl alloys.

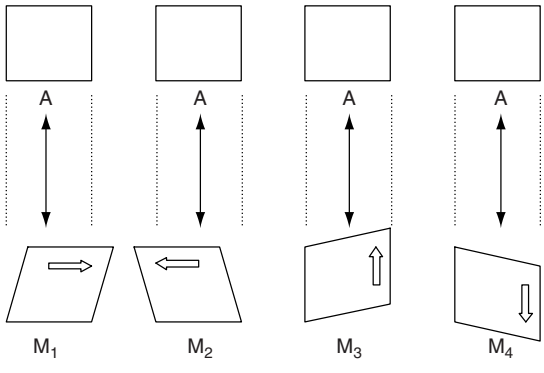

Figure W12.7. Four possible distortions of a square (phase A) to a rhombus (phase M).

transformation is reversible and is first order. No atomic-scale diffusion is taking place and no slippage of atomic planes is occurring. Everything about the transition is predictable, with randomness playing little role other than accelerating thermally assisted transitions. The material is said to be *thermoelastic*. In reality, the unit cell for the SMA materials is much larger, as may be seen by looking at the stoichiometry of the materials (see Table W12.2). It is useful to think of the unit cell as being composed of subunit cells with vacancies that may appear on different faces.

When the martensitic transition occurs, upon cooling there are a number of different states the subunits can assume in the low-symmetry phase. This is illustrated in Fig. W12.7, where the A phase is represented by a square and the M phase is

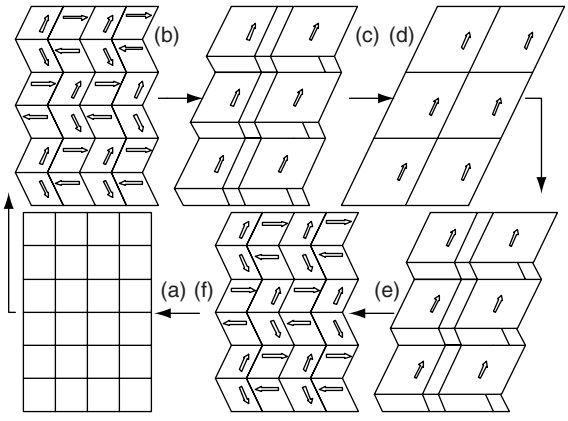

Figure W12.8. Stages in the shape-memory process.

represented by a rhombus (which has lower symmetry). The four orientations are labeled by a set of arrows. These structures self-accommodate (i.e., when the A-to-M transition occurs, there is no change in the macroscopic size of the object). The material consists of the various types of rhombi intermeshed with each other. This is illustrated in Fig. W12.8, where several such rhombi are drawn. In Fig. W12.8*a* one starts with an austenite crystal at a temperature above T_{A_f}, represented by a rectangle. The crystal is then cooled to the martensite phase. Figure W12.8*b* shows that the large-scale shape is still rectangular but now has rhombus "domains" that accommodate each other. A stress is then applied to the crystal to change its shape to a parallelogram. Figure W12.8*c* shows that one type of domain grows at the expense of the others, and eventually, in Fig. W12.8*d* the desired shape is obtained. If the stress is removed, the parallelogram shape is retained.

When a rhombus is forced to have a different orientation than its state of minimum free energy would allow, stress is built into it. The system adjusts in such a manner as to relieve this stress. This determines which rhombus will be the next to alter its shape. Modification of the structure takes place in a sequential manner. In this way the system has a memory, which consists of the sequence of stress-relaxing deformations that take place. In some ways the process is similar to magnetizing a ferromagnet, with a self-consistent strain replacing the role played by the self-consistent magnetic field. Unlike the magnetic case, however, there is only one return path that the alloy can follow when it is heated, and that is determined by the original orientations of the rhombi.

Now the sample is heated. The domains retrace their evolution (see Fig. W12.8*e* and *f*) until, when T_{A_f} is passed, the crystal has reverted to its original shape. If the temperature is lowered again, the parallelogram shape is not regained unless it is reshaped by external forces.

SMA materials exhibit a high degree of strain recovery, meaning that they revert to their original size and shape when the stress causing the strain is relaxed. For example, a NiAl alloy can have a strain recovery of 7%. The stress–strain curve exhibits superelasticity. What appears to be plastic deformation in the M phase disappears when the sample is heated to the A phase. In addition, it is possible to induce the martensitic transformation by applying an external stress field. A more complete description of the

material involves a three-dimensional phase diagram with stress plotted as a function of both strain and temperature.

Applications of SMA materials benefit from their ability to store a large amount of mechanical strain or elastic energy in a small volume. They may be used for such diverse applications as circuit breakers, switches, automatic window openers, steam-release valves, hydraulic controls for aircraft, rock cracking, sealing rings, and actuators. They can even be used to unfurl antennas on satellites, where a bulky motor assembly may be replaced by a simple SMA. A limitation on their use, however, is their slow response time, being limited by thermal conduction.

W12.7 Metallic Glasses

If a liquid metal alloy were to be rapidly quenched (i.e., its temperature lowered sufficiently rapidly) it is possible to solidify it without forming a crystalline state. Such a material is called a *metallic glass*. Since the thermal conductivity of metals is high and since the crystalline state is generally the state of lower free-energy, metals have a strong tendency to crystallize quickly. However, if a small droplet of liquid alloy is projected onto a cold surface, the resulting "splat" can cool very rapidly (with rates on the order of -10^6 K/s) and become a metallic glass. Alternatively, one could inject a fine stream of the molten alloy into a high-conductivity cold liquid to form the glass, or vapor-deposit onto a cryogenic substrate. In many ways the formation of a metallic glass is similar to that of window glass, but the thermal relaxation times are orders of magnitude faster. The metallic glasses are essentially solids, with diffusion rates often less than 10^{-22} m^2/s, orders of magnitude smaller than in crystals. The random close-packing model for metallic glasses is discussed in Chapter 4. Rapid quenching is described further in Chapter W21.

These materials are amorphous and hence do not have dislocations, but rather, a high degree of disorder on the atomic scale. They are strong, stiff, and ductile. In addition, they are corrosion resistant. Furthermore, being largely homogeneous, they allow sound to propagate without appreciable attenuation due to scattering. This is because, for most acoustic applications, the wavelength of sound is long compared with the scale size of the inhomogeneities, and the sound propagates through an effectively isotropic medium. Things are different, however, when short-wavelength phonons are involved, such as in the thermal-conduction process. Due to the lack of a crystal lattice the metallic glasses are generally poor thermal and electrical conductors, with very short phonon and electron collisional mean free paths.

Examples of metallic glasses include AuSi near the eutectic composition of 19 at % Si, $Pd_{80}Si_{20}$, $Pd_{78}Si_{16}Cu_6$, and $Ni_{36}Fe_{32}Cr_{14}P_{12}B_6$. They include transition metals (Co, Fe, La, Mn, Ni, Pd, Pt, Zr) alloyed with (B, C, N, P, Si) near the eutectic composition. Some are ferromagnetic (e.g., $Pd_{68}Co_{12}Si_{20}$ or $Fe_{83}P_{10}C_7$) and some are antiferromagnetic (e.g., $Mn_{75}P_{15}C_{10}$). The ferromagnets are readily magnetized or demagnetized, since there are no large-scale defects that pin the domain walls. The magnets are soft in the amorphous state because the domain wall thickness is much larger than the domain size. This is likely to be due to the absence of well-defined magnetic anisotropy in the magnetic metallic glass as a result of the lack of crystalline order. As discussed in Section 17.2 strong magnetic anisotropy favors magnetic domains with narrow domain walls. The metallic glass $Fe_{80}B_{11}Si_9$ is commonly used in power magnetic applications such as power distribution due to its high Curie temperature, $T_C = 665$ K, and hence its good thermal stability.

It is found that the more elements present in the alloy, the more complex the unit cell of a crystal is, and hence the longer it would take to crystallize. An example is the alloy $Zr_{41.2}Ti_{13.8}Cu_{12.5}Ni_{10.0}Be_{22.5}$ which forms a metallic glass at cooling rates of only 10 K/s. The high resistance to crystallization is believed to be due to the low melting point of the corresponding crystalline alloy and the fact that the alloy is composed of atoms of quite different sizes. Since one wants the glass to form rather than the crystal, it is preferable to work with materials with long crystallization times. This accounts for the high integers in the stoichiometry.

A further aid in the formation of the metallic glass is to have a composition corresponding to the eutectic point, as in the case of AuSi, whose binary phase diagram is sketched in Fig. W12.9. Since the eutectic temperature is low, diffusion will be sluggish when the solid is formed, and the formation of crystals will be slow. If the temperature drop is sufficiently fast, the eutectic metal will become a glass.

The metallic glass is only slightly less dense than the corresponding crystal. It tends to form a random close-packed structure (see Chapter 4) of a binary system with two sphere sizes (Fig. W12.10). The bonding is primarily metallic. There is some evidence of short-range order [i.e., there are different polyhedral arrangements (e.g., tetrahedra, octahedra, trigonal prisms and cubic biprisms)], which appear in definite proportions but are not spatially ordered. The bulk modulus of a metallic glass is found to be comparable to its crystalline counterpart. The shear modulus, however, is typically reduced by 25%. They have fairly low values of yield stress and can undergo large plastic deformations of up to about 50%. If a crack were to form and stress were concentrated in the neighborhood of its tip, the tip region would yield, the sharpness of the tip would be reduced, and the stress would be relieved. This healing mechanism curtails crack propagation and makes the material tough (i.e., able to withstand large stresses without fracturing). Repetitive cycling of the stress on and off does not work-harden the material, since no dislocations are present.

As the temperature is raised from room temperature to about half the melting temperature, activated hopping of atoms becomes important. The atoms can search for the lowest free-energy state and the solid can begin to crystallize. This prevents the metallic glasses from being employed in high-temperature applications.

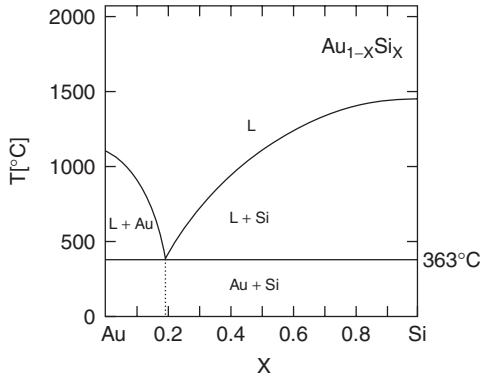

Figure W12.9. AuSi tends to form a metallic glass near the eutectic composition, indicated by the dashed line on the binary phase diagram. [Adapted from J. J. Gilman, Metallic glasses, *Phys. Today*, May 1975, p. 46. See also H. Okamoto et al., *Bull. Alloy Phase Diagrams*, **4**, 190 (1983).]

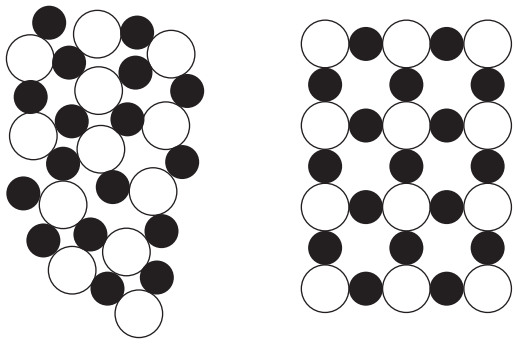

Figure W12.10. Arrangement of a binary-alloy metallic glass (*a*) compared with the crystalline state (*b*).

Possible applications for metallic glasses include transformers, tape-recording heads, filaments to reinforce rubber tires, transmission belts, and tubing. Their hardness makes them suitable for cutting instruments. Their low acoustic-attenuation feature makes them appropriate for use where sound vibrations are likely to be prevalent, such as in loudspeakers.

In crystalline metals, different crystallographic faces have different work functions and hence there is a contact potential difference between them. In an ionic solution it is possible for corrosion to take place as ionic currents between the faces are established. Due to the amorphous nature of the metallic glass, there is overall isotropy, and these contact potential differences do not exist. This tends to make the metallic glasses corrosion resistant.

W12.8 Metal Hydrides

The ability of hydrogen to adsorb on metals, dissociate, diffuse into the bulk, and then form chemical compounds provides a way to store hydrogen in metals. The density of hydrogen in metals can even exceed that of liquid hydrogen. This is attractive since the process can often be reversed and the hydrogen may be released simply by warming the metal. Hydrogen is a fuel with a high energy content and produces only water vapor when it is burned. This makes it an attractive chemical-energy source for a future technology.

Some metals can store only a fraction of a hydrogen atom per metal atom (e.g., $TaH_{0.5}$), whereas others can store more (e.g., Th_4H_{15} or CeH_3). The metal Ta has a BCC crystal structure, whereas Th and Ce have FCC crystal structures. The hydrogen atom, being small, generally occupies interstitial sites, as is illustrated in Fig. W12.11. In the left diagram there is an FCC metal with a hydrogen at one of the eight tetrahedral interstitial sites per unit cell. In the right diagram the hydrogen is at one of the four octahedral interstitial sites. In some cases all the FCC interstitial sites are occupied, such as in Th_4H_{15} and CeH_3. For an FCC cell there are eight tetrahedral interstitial sites, four octahedral interstitial sites, and four atoms per unit cell. For CeH_3 it may happen that all the interstitial sites are occupied. In Th_4H_{15} there could be more than one hydrogen per site. The hydrogen atoms generally have a high diffusivity through the

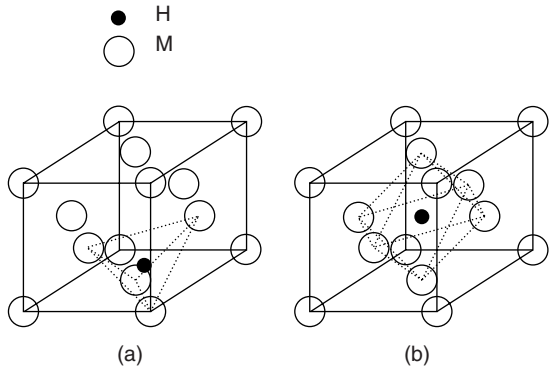

Figure W12.11. (*a*) Hydrogen at a tetrahedral interstitial site in an FCC unit cell; (*b*) hydrogen at the octahedral interstitial site in the same cell.

metal and readily hop from site to site. Some of this hopping ability is due to thermal activation, but there is also an appreciable part due to quantum-mechanical tunneling. This is similar to what occurs in the free NH_3 molecule, where the tetrahedron formed by the atoms periodically inverts as the N atom tunnels through the barrier presented by the three H atoms. (In the actual motion there is a concerted motion in which all atoms participate.) The hopping rates may be as large as a terahertz. At high-enough concentrations the absorbed hydrogen can induce structural phase transitions in the metal. This provides the means for monitoring the hydrogen content. It is also responsible for *hydrogen embrittlement*, in which a metal may be weakened by the presence of H. Imperfections, such as vacancies in the metal, can act as centers for concentrating H, and as a result, recrystallization may take place. This causes a large stress concentration and the imperfection may propagate because of it.

The presence of H may also cause drastic changes in the electrical and magnetic properties of the metal. Hydrogen generally tends to suppress magnetism. This might be expected because the origin of magnetism stems from the spin-dependent exchange interaction between neighboring metal atoms, and this, in turn, depends on the wave-function overlap. As new bonds are formed to create the hydride, less of the wave-function is left to participate in magnetism.

In some instances the H causes the metal to become a semiconductor. Electrons are extracted from the conduction band of the metal and are tied up in chemical bonds to form the hydride. It is also found that the metals may become superconductors with transition temperatures considerably higher than the bare metals, perhaps due to the enhanced electron–phonon coupling (see Chapter 16). Examples include Th_4H_{15} and PdH. Some of the anomalies observed for the hydrides are similar to those observed in the high-temperature cuprates (e.g., an absence of an isotope effect for the superconducting transition temperature).

W12.9 Solder Joints and Their Failure

Solder joints play a crucial role in the operation of electronic-circuit boards since they provide both the mechanical and, more important, the electrical connections for the various components and chips. Two modes of failure of these joints may be identified. The first is aging. In the normal course of operation the joints are subject to

thermal cycling. Due to the mismatch of coefficients of thermal expansion, heating leads to stresses. These stresses cause the motion of dislocations, which may pile up to form microscopic cracks or void spaces. The resulting embrittlement makes the joint susceptible to fracture. A second source of failure results from intermetallic compound (IMC) formation. Compound particles nucleate and grow within the joints and produce mechanical stresses due to lattice-constant mismatch, and these can also cause embrittlement. Since a typical circuit board may contain many hundreds of joints, even a small probability for failure in a joint may compound to a severe lifetime limitation for the board. The processes responsible for failure are identified by examining the joints under high-power optical microscopes.

Examples of IMC formation that results from use of the common eutectic Pb–Sn solder (see Fig. 6.8) on copper are $\mathrm{Sn} + 3\mathrm{Cu} \rightarrow \mathrm{Cu_3Sn}$ or $6\mathrm{Cu} + 5\mathrm{Sn} \rightarrow \mathrm{Cu_6Sn_5}$. Similarly, Ni can form a highly brittle compound when reacting with solder. The growth of the layer thickness of an IMC, z, is governed by an empirical equation of the form

$$\frac{dz}{dt} = A_0 \frac{e^{-E_a/k_B T}}{z^n}, \tag{W12.27}$$

where A_0 is a constant, E_a an activation energy, and n an empirical exponent ranging from $\frac{1}{2}$ to 1. It is found that the thicker the IMC layer, the more susceptible it is to brittle fracture.

Ideally, solder joints should be designed to eliminate, or at least minimize, these problems. One might try using spring-shaped elastic-component leads to relieve the thermal stresses that develop. This conflicts with the desire for a higher concentration of components on the board. It is better to match the coefficients of thermal expansion to eliminate the thermal stresses altogether. However, this often leads to a degradation of the electrical properties of the leads. It was found that decreasing the solder-joint thickness results in a reduced tendency for fractures to occur. This may be because of the ability of the joint to anneal its defects to the surface. One may also try to make the material more homogeneous so that dislocations are less likely to be present. Alternatively, one may try to alloy the material and insert dopants that would trap the dislocations and prevent them from propagating to form cracks.

To date there is no preferred method. Each has its benefits and its drawbacks. The design of joints is still in the "arts" stage.

W12.10 Porous Metals

Porous metals define a class of materials that find application in such diverse areas as filters, heat exchangers, mufflers and other noise-abatement devices, fuel cells, electrolytic cells, hydrogen-storage media, and thermal insulators. They may be fabricated using several techniques, including sintering and slip-casting. The sintering method involves mixing powders of the metal, M, with powders of another material, A, with a higher melting point. When the metal M melts, it flows around the particles of A and forms a solid metallic cage as it is cooled. If the pores are interconnected, material A can then be removed by chemical means, so the porous metal M remains. In the slip-casting method a solid foam is created from a nonmetallic material, and a dispersion of fine metal powder is absorbed by this sponge. When heated, the metal particles fuse together and the nonmetallic powder is burned away. Again the metallic foam

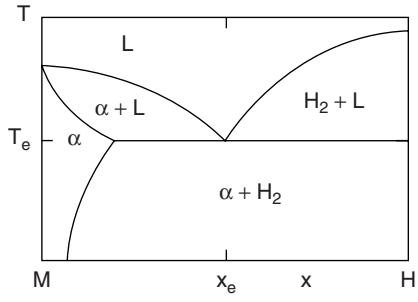

Figure W12.12. Binary phase diagram for a metal–hydrogen alloy. (Adapted from V. Shapovalov, Porous metals, *Mater. Res. Soc. Bull.*, Apr. 1994, p. 24.)

is produced. Chemical vapor-deposition techniques may be employed to build up a thickness of metal on a porous substrate and then to remove the substrate by chemical or thermal means, leaving behind a metal film.

The materials are characterized by a filling factor, which tells what fractional volume of space is occupied by the metal, a distribution of pore sizes and shapes, and a topology describing the interconnection between the pores. They are found to be poor electrical conductors, both because of the low filling factors and the high degree of boundary scattering along the thin conducting paths.

The term *gasar* has been coined to describe a foam produced by a gas–metal eutectic transition. Due to the small size of the hydrogen atom (especially when it is ionized to a proton), it has little difficulty being adsorbed in many metals, as discussed in Section W12.8. The resulting hydrogen–metal alloy phase diagram often has a eutectic transition. Such a diagram is illustrated in Fig. W12.12. The compound is of the form $M_{1-x}H_x$. Hydrogen is bubbled into the liquid metal to increase x to the eutectic composition x_e. The material is then cooled below the eutectic temperature T_e. This produces a eutectic composition consisting of a mixture of the α phase of the metallic hydride and H_2 gas. The gas is able to desorb from the hydride, leaving behind a porous structure. Gasars have proven to be the strongest of the porous–metal structures. This is probably due to a homogeneous pore size distribution, which permits loading stresses to be distributed uniformly. If residual hydrogen is trapped in the metal, the gasar is found to be a good thermal conductor, since hydrogen is light and mobile and therefore is able to convect the heat through the pore structure. The material is also able to damp acoustic waves efficiently, since the trapped gas makes inelastic collisions with the surrounding cage as the cage vibrates back and forth.

REFERENCES

Frear, D. R., and F. G. Yost, Reliability of solder joints, *Mater. Res. Soc. Bull.*, Dec. 1993, p. 49.

Gilman, J. J., Metallic glasses, *Phys. Today*, May 1975, p. 46.

Kohn, W., Overview of density functional theory, in E. K. U. Gross and R. M. Dreizler, eds., *Density Functional Theory*, Plenum Press, New York, 1995.

Shapovalov, V., Porous metals, *Mater. Res. Soc. Bull.*, Apr. 1994, p. 24.

Smith, W. F., *Structure and Properties of Engineering Alloys*, 2nd ed., McGraw-Hill, New York, 1993.

Tien, J. K., and T. Caulfield, *Superalloys, Supercomposites and Superceramics*, Academic Press, San Diego, Calif., 1994.

Wayman, C. M., Shape memory alloys, *Mater. Res. Soc. Bull.*, Apr. 1993, p. 42.

Westbrook, J. H., et al., Applications of intermetallic compounds, *Mater. Res. Soc. Bull.*, May 1996, p. 26.

Westlake, D. G., et al., Hydrogen in metals, *Phys. Today*, Nov. 1978, p. 32.

CHAPTER W13

Ceramics

W13.1 Ternary Phase Diagrams

As the number of components of a system increases, the number of possible subsystems increases rapidly and the complexity grows exponentially. For example, a two-component system has only two possible unary subsystems and one binary subsystem for a total of three different types of subsystems. A three-component system has three unary subsystems, three binary subsystems, and a ternary subsystem, for a total of seven different types of subsystems. In the general case a C-component system will have $C!/[C'!(C - C')!]$ subsystems with C' components, and will have a total of $2^C - 1$ possible subsystems. Often, it is desirable to optimize a particular physical property of the system, so the composition and temperature must be chosen carefully to achieve this optimization. Obviously, the process becomes more challenging as the number of components is increased. Phase diagrams provide a type of road map upon which it is possible to chart the composition of the material and indicate the various phase boundaries.

Often, materials with interesting physical properties are constructed out of just three components, which will be labeled by A, B, and C. These may be elements or compounds. For example, the electro ceramic $Pb_xZr_yTi_zO_3$ (PZT) is constructed from the compounds $A = PbO$, $B = TiO_2$, and $C = ZrO_2$, and the composition is $(PbO)_x \cdot (ZrO_2)_y \cdot (TiO_2)_z$. Here x, y, and z are constrained by the valence balance condition $2x + 4y + 4z = 6$, so that only two of the variables may be varied independently. The high-temperature superconductor $YBa_2Cu_3O_{7-x}$ is but one of many phases constructed from Y_2O_3, BaO, and Cu_2O. Glasses are often made from ternary mixtures, such as soda-lime, made from SiO_2, CaO, and Na_2O.

According to the Gibbs phase rule (see Section W6.4), Eq. (W6.9), the number of degrees of freedom, F, is related to the number of components, C, and the number of phases, P, by $F = C - P + 2$. For constant temperature and pressure, two of the degrees of freedom are removed, leaving $F' = C - P$ degrees of freedom. For a three-component system, such as PZT, $C = 3$. Since there must be at least one phase present, $p \geq 1$ and $F' \leq 2$. The two degrees of freedom are conveniently displayed using the Gibbs triangle, as illustrated in Fig. W13.1.

Imagine that there is a totality of one unit of components, so the chemical formula is $A_aB_bC_c$, with $a + b + c = 1$ and (a, b, c), each lying in the range 0 to 1. The composition may be represented graphically as a point inside an equilateral triangle. The height of this triangle is taken to be 1. In Fig. W13.1 point O represents $A_aB_bC_c$. The perpendicular distances to the sides of the triangle are a, b, and c, and the fractions of components A, B, and C present are also a, b, and c. The corners of the triangle represent pure-component (unary) compounds. If the point O were at A, then

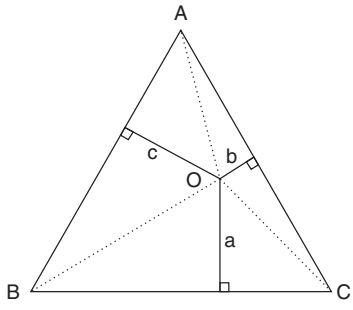

Figure W13.1. Point O represents the composition $A_aB_bC_c$, where $a + b + c = 1$.

$b = c = 0$ and $a = 1$. The composition would be 100% A. The edges of the triangle represent binary compounds. For example, a point on the base of the triangle will have composition B_bC_c, with $b + c = 1$. If the point O is at the center of the triangle, then $a = b = c = \frac{1}{3}$ and 33.3% of each component is present.

It is a simple matter to prove that $a + b + c = 1$. Note that the area of equilateral triangle ABC (with side $L = 2/\sqrt{3}$) is half the base times the height: $(\frac{1}{2})(L)(1) = 1/\sqrt{3}$. On the other hand, the area of ABC may be written as the sums of the areas of the three triangles AOB, BOC, and COA, which gives $1/\sqrt{3} = (\frac{1}{2})L(a + b + c)$, so $a + b + c = 1$. Thus any point within the triangle ABC will always correspond to a total of one unit of components.

An alternative method for determining the composition is to make the construction shown in Fig. W13.2. Lines are passed through point O parallel to the three sides. The intersections of these lines with the sides are labeled by the points D, E, F, G, H, and I. It can be shown that the relative amounts of A, B, and C present are proportional to the lengths of segments of the sides, that is,

$$\frac{c}{\text{AI}} = \frac{b}{\text{IH}} = \frac{a}{\text{HC}}, \quad \frac{a}{\text{FG}} = \frac{b}{\text{GC}} = \frac{c}{\text{BF}}, \quad \frac{c}{\text{DE}} = \frac{a}{\text{EB}} = \frac{b}{\text{AD}}. \tag{W13.1}$$

This construction may be generalized to the case of a scalene triangle. In Fig. W13.3, point O represents 1 mol of material with composition $A_aB_bC_c$, where $a + b + c = 1$. Through point O, construct-lines FOI, HOE, and DOG are drawn parallel to sides CB, AC, and BA, respectively. Each side is divided into three segments by these lines. It may be shown that the following identity holds for the lengths of the segments:

$$\text{DE:EC:BD} = \text{CF:FG:GA} = \text{IB:AH:HI} = a\text{: } b\text{: } c. \tag{W13.2}$$

The ternary diagram is used to depict the various phases of the material at thermal equilibrium. At times one is interested only in the phase boundaries at a given temperature and pressure. The diagram is then called an *isothermal-ternary diagram*. Alternatively, the temperature field could be represented by drawing isothermal contours on the diagram. Since this proves to be more useful, this representation will be used here.

Refer to Fig. W13.4, where a three-dimensional temperature–composition diagram is drawn. Viewed from the top, one has a ternary phase diagram. This diagram will be used to follow a process in which a liquid solidifies. At sufficiently high temperatures

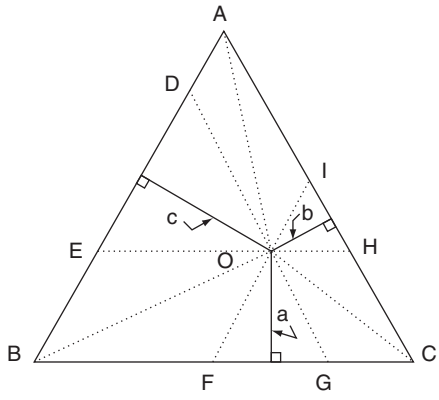

Figure W13.2. Material $A_aB_bC_c$ is represented by point O. The segments AI:IH:HC are in the same proportion as $c:b:a$.

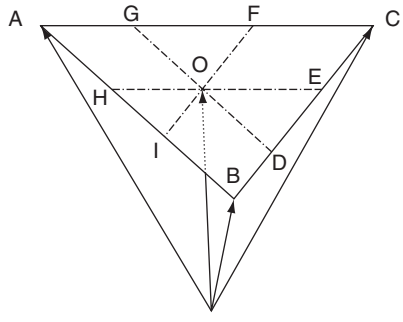

Figure W13.3. Composition triangle ABC together with various construction lines.

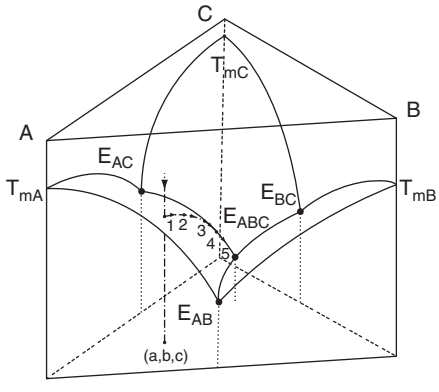

Figure W13.4. Three sheets of the liquidus surface on a plot of temperature versus composition.

the material is assumed to be liquid. As the temperature is slowly lowered, the material begins to crystallize. The degree of crystallization, and the fractions and compositions of solid and liquid formed, are determined by the liquidus surfaces. Of course, the mean composition taken over all the phases always remains the same. In Fig. W13.4 the liquidus surface is presented for the simple case in which solid solutions are not formed. The liquidus surface consists of three separate sheets, corresponding to the three primary compositions A, B, and C. Various eutectic points are depicted by the letter E with subscripts. Thus E_{AB} denotes the eutectic point for the composition A_aB_b for the special case where $a + b = 1$ and $c = 0$. E_{ABC} is the ternary eutectic point and is the lowest point for which some liquid remain. There is a horizontal eutectic plane (not shown) in the phase diagram passing through the point E_{ABC} below which only completely solid material exists. The melting points for the pure components are denoted by $T_{m\mathrm{A}}$, $T_{m\mathrm{B}}$, and $T_{m\mathrm{C}}$.

Shown on Fig. W13.4 is a cooling path for a liquid with composition (a, b, c). As the temperature is lowered, point 1 is encountered and solid phase A begins to nucleate. Further reduction of the temperature causes an increased growth of phase A and a modification of the composition of the liquid. The liquid composition is determined by the curve 1–2–3–4–5. Along 1–2–3, only solid phase A and a liquid are present. At point 3, phase C begins to nucleate. Along path 3–4–5 (which is the valley between sheets A and C), phases A and C and a liquid of varying composition are present. At point 5 the liquid reaches the ternary eutectic composition. At a lower temperature, only solid phases A, B, and C exist, with the original composition (a, b, c).

Figure W13.5 depicts the same scenario as in Fig. W13.4 but viewed from above. The isothermal contours are not shown but are there implicitly. Note that A–1–2–3 is a straight line. Along line 1–2–3 the composition may be determined by applying the lever rule. Thus at a temperature corresponding to T_1, the liquid will have composition (a_1, b_1, c_1). The amounts of liquid and phase α at $T = T_2$ are in the ratio of the distances d_{A1}/d_{12}. At temperature T_3 the liquid has composition (a_3, b_3, c_3) and the liquid to phase α ratio is d_{A1}/d_{13}. At points 4 and 5 the compositions are such that the center of gravity of points A, C, 4, or 5 lies at the original point 1.

There are numerous other possibilities for drawing the phase diagrams but they will not be covered exhaustively here. The principles of analysis are similar. Several points are worth mentioning, however. Stoichiometric binary compounds (e.g., A_mB_n,

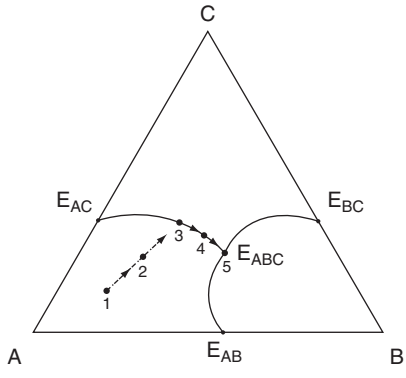

Figure W13.5. Path toward solidification on the ternary phase diagram.

with m and n integers) are represented by points on the appropriate edge (AB in this case). Stoichiometric ternary compounds (e.g., $A_mB_nC_j$, with m, n, and j integers) appear as points in the interior of the triangle. These points are usually surrounded by a domain of influence bounded by a phase boundary. An example of this will be encountered in Section 13.7 of the textbook[†] when the ternary phase diagram for the glass-forming region of $Na_2O \cdot CaO \cdot SiO_2$ is discussed (see Fig. 13.15). The net result is that the ternary phase diagram often has the appearance of a mosaic with numerous phases indicated. Often, there is a definite crystalline order associated with a stoichiometric phase. Points with nearby compositions may be thought of as crystals possessing defects. The farther one goes from the stoichiometric point, the larger the number of defects. When a sufficient number of defects occur, a phase transition to another crystal structure may result.

As mentioned earlier, it is possible to have as many as three distinct phases present at once (i.e., $P = 3$). In that case, the effective number of degrees of freedom for a ternary system is $F = C - P = 0$. Consider the Gibbs triangle depicted in Fig. W13.6, which shows three phases (α, β, γ) to be present. Since $F = 0$, the composition of the material at point O is uniquely determined: the fractions of the various phases present are (f_α, f_β, f_γ), where $f_\alpha + f_\beta + f_\gamma = 1$. For the point O, the composition (a, b, c) will be determined by solving the matrix equation

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a_\alpha & a_\beta & a_\gamma \\ b_\alpha & b_\beta & b_\gamma \\ c_\alpha & c_\beta & c_\gamma \end{bmatrix} \begin{bmatrix} f_\alpha \\ f_\beta \\ f_\gamma \end{bmatrix}. \tag{W13.3}$$

In Fig. W13.7 a sequence of four isothermal sections is illustrated, corresponding to the temperatures $T_1 > T_2 > T_3 > T_4$ for an idealized ternary system. Temperature T_1 is above the liquidus surface, so any point in the phase diagram corresponds to a homogeneous liquid. At temperature T_2 it is assumed that part of the liquidus surface is above the isothermal plane and part below. It is assumed that there are compositional ranges for which the phases α, β, and γ coexist with the liquid phase, as illustrated in

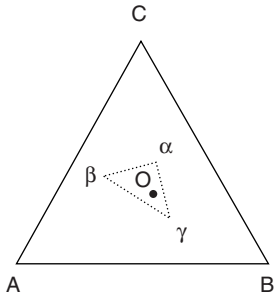

Figure W13.6. Gibbs triangle with a three-phase field. There is a unique admixture of the three phases at point O.

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

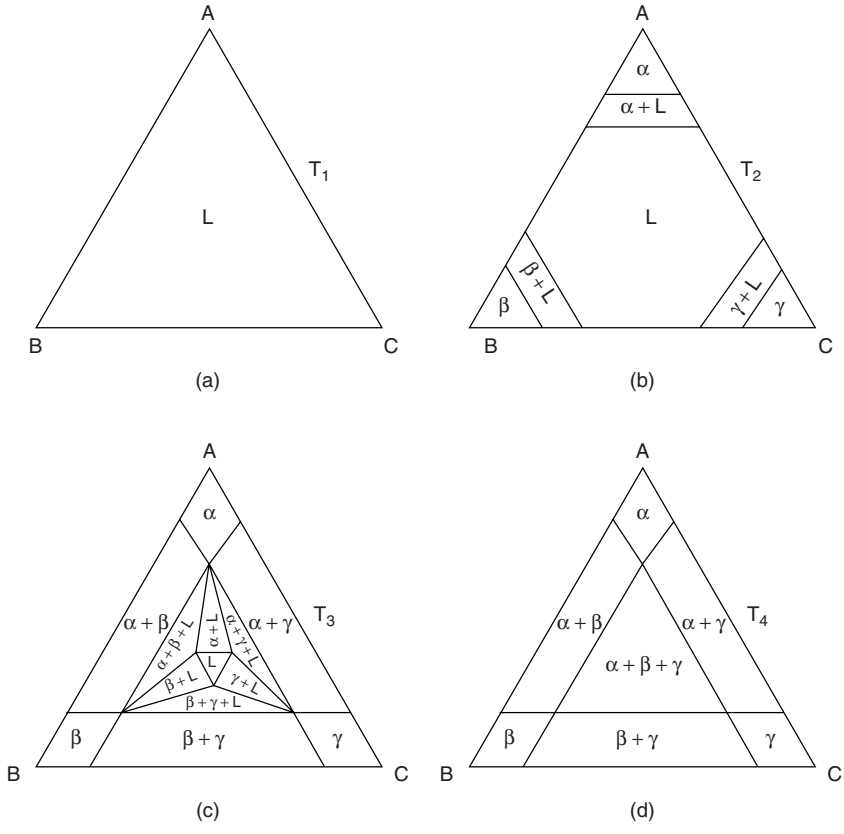

Figure W13.7. Sequence of four isothermal phase diagrams, illustrating the presence of various phases.

the figure. At T_3 the temperature is slightly above the three-phase eutectic temperature. One now finds the coexisting binary solid phases $\alpha + \beta$, $\beta + \gamma$, and $\alpha + \gamma$. There are also regions corresponding to the coexistence of the unary phases with the liquid, $\alpha + L$, $\beta + L$, and $\gamma + L$, as well as regions consisting of the coexistence of the two phases with the liquid, $\alpha + \beta + L$, $\beta + \gamma + L$, and $\alpha + \gamma + L$. At T_4, below the eutectic temperature, only solid phases are present: the unary phases α, β, or γ; the two-phase regions $\alpha + \beta$, $\beta + \gamma$, or $\alpha + \gamma$; and the three-phase region $\alpha + \beta + \gamma$.

It is important to stress that the phase diagram applies only for thermal equilibrium. Nevertheless, for rapid cooling, the diagram may be used as an intuitive guide to understanding solidification. The composition of the microstructure that will form may be estimated in much the same way as in the study of metals (see Section 6.5 and Figs. 6.9 and 6.10). The faster the material passes through a given phase domain as the sample is cooled, the less time there is available for nucleation and growth of that equilibrium phase to occur.

W13.2 Silicates

Silicon and oxygen are the two most abundant elements in Earth's crust. There is a broad class of minerals based on combinations of Si and O and other elements called

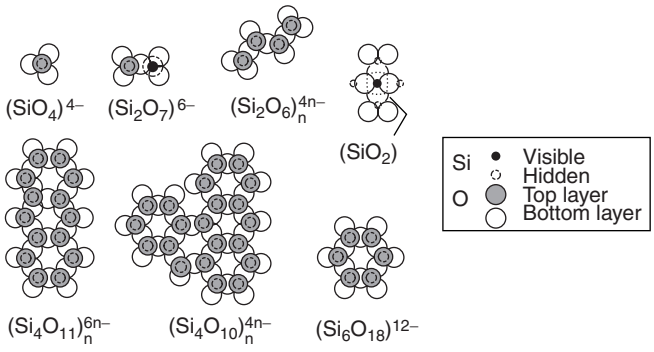

Figure W13.8. Schematic representation of the seven classes of silicate ions. There are O^{2-} ions residing at the corners of the tetrahedra and Si^{4+} ions at their centers. (Adapted from H. W. Jaffe, *Crystal Chemistry and Refractivity*, Dover, Mineola, N.Y., 1996.)

silicates. An appreciation of the various ions formed from Si and O permit one to understand more complex structures in which other cations, such as Al, substitute for the Si ions.

The valence of Si is +4 and that of O is −2. A basic ion formed is the $(SiO_4)^{4-}$ ion. The Si^{4+} resides at the center of a tetrahedron, and the O^{2-} ions are at the vertices. The bond is about equally covalent and ionic and is very strong. The tetrahedra may be connected in a variety of ways to form complex ions. Figure W13.8 depicts the basic structures. There are seven principal classes of silicates. Orthosilicates (also known as nesosilicates or island silicates), such as forsterite (Mg_2SiO_4), olivine ($Mg_xFe_{2-x}SiO_4$), and zircon ($ZrSiO_4$), are based on independent $(SiO_4)^{4-}$ tetrahedra linked by divalent cations. In place of the $(SiO_4)^{4-}$ ion, there could be substituted the $(AlO_4)^{5-}$ ion. An example of this is the synthetic crystal YAG [yttrium aluminum garnet, $Y_3Al_2(AlO_4)_3$], used as a laser crystal. In the sorosilicates there are two tetrahedra joined vertex to vertex, sharing a common oxygen to form the $(Si_2O_7)^{6-}$ ion. An example is the mineral thortveitite [$Sc_2(Si_2O_7)$]. The structure with a triad of tetrahedra corner-sharing one oxygen ion to form the $(Si_3O_9)^{6-}$ ion does not seem to be found in nature. In the cyclosilicates, such as the gemstone beryl ($Be_3Al_2Si_6O_{18}$), the tetrahedra are arranged in hexagonal rings corner-sharing six oxygens to create $(Si_6O_{18})^{12-}$ ions. In the inosilicates, such as the mineral jadeite [$NaAl(Si_2O_6)$], tetrahedra form a linear chain with corner-shared oxygens to produce an ion of the form $(SiO_3)_n^{2n-}$. In the phyllosilicates, such as mica or talc [$Mg_3(Si_2O_5)_2(OH)_2$], the basic ionic unit is the $(Si_2O_5)^{2-}$ ion. In the amphiboles (or double-chain silicates) two parallel inosilicate chains link together so that every second tetrahedron has a corner-shared oxygen, producing the ion $(Si_4O_{11})_n^{6n-}$. An example is the mineral tremolite [$Ca_2Mg_5(Si_4O_{11})_2(OH)_2$]. The final class of silicate is the tektosilicate, based on the neutral SiO_2 subunit. An example of this is quartz itself, with the composition SiO_2, or anorthite [$CaOAl_2O_3(SiO_2)_2$]. The results are summarized in Table W13.1.

An oxygen shared by two tetrahedra is called a *bridging oxygen*. One that is not shared is called a *nonbridging oxygen* (NBO). One may classify the structures according to the number of nonbridging oxygens that the tetrahedra possess, as shown in Table W13.1. Tektosilicates have no NBOs, or equivalently, four shared corners. The structural unit is neutral and is based on SiO_2. Disilicates have only one NBO

TABLE W13.1 Seven Principal Classes of Silicates

Class	Ion	Shared Corners	Nonbridging Oxygens
Nesosilicate	$(SiO_4)^{4-}$	0	4
Sorosilicate	$(Si_2O_7)^{6-}$	1	3
Cyclosilicate	$(Si_6O_{18})^{12-}$	2	2
Inosilicate	$(SiO_3)_n^{2n-}$	2	2
Amphibole	$(Si_4O_{11})_n^{6n-}$	2, 3	2, 1
Phyllosilicate	$(Si_2O_5)^{2-}$	3	1
Tektosilicate	(SiO_2)	4	0

Source: Data from H. W. Jaffe, *Crystal Chemistry and Refractivity*, Dover, Mineola, N.Y., 1996.

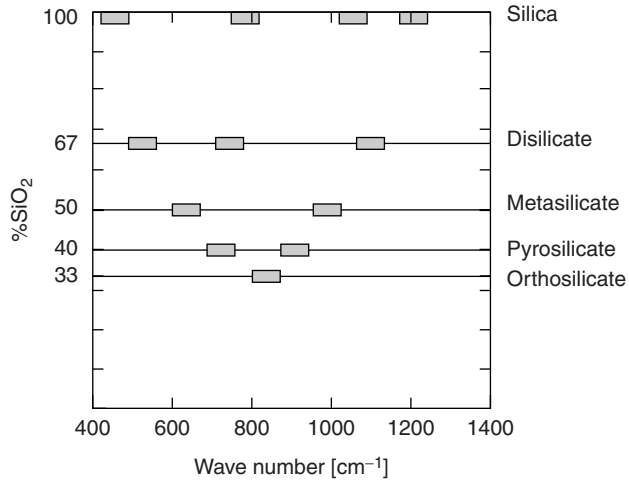

Figure W13.9. Ranges of Raman shifts for various silicates. [Adapted from P. F. McMillan, *Am. Mineral.*, **69**, 622 (1984).]

or, equivalently, three shared corners, and the ion is $(Si_2O_5)^{2-}$. Metasilicates have two NBOs (i.e., two shared corners) and the ion is $(SiO_3)^{2-}$. Pyrosilicates have three NBOs (i.e., one shared corner) and the ion is $(Si_2O_7)^{6-}$. Orthosilicates have four NBOs, hence no shared corners, and are based on the $(SiO_4)^{4-}$ ion.

Raman scattering may be used to identify the various ions. In Fig. W13.9 the ranges of the Raman bands for the various ions in silicate glasses are depicted by the shaded areas. In silicates there are cations present in addition to the silicate ions, so that one may regard the materials as part silica and part foreign cations. The ordinate of Fig. W13.9 gives the percentage of the material that is SiO_2. Silica, of course, is 100% SiO_2. The 400-cm^{-1} peak is associated with a rocking motion in which the Si–O–Si angle remains fixed but the oxygen rocks back and forth perpendicular to the initial Si–O–Si plane. The 800-cm^{-1} peak corresponds to a bending motion of the Si–O–Si bond angle. The peak at 1100 to 1200 cm^{-1} is due to a stretching motion of the Si–O bond. In the orthosilicates, the bending motion of the Si–O–Si bond is responsible for the 800-cm^{-1} peak. In the pyrosilicates two tetrahedra are joined together. The bending motions could be either in phase or out of phase. As a result, the 800-cm^{-1}

peak is split into two peaks, one at a higher frequency and the other at a lower one. A normal-mode analysis of the silicate ions leads to a more detailed description of the correlation of peak location with ion type.

W13.3 Clay

Shards of pottery excavated in scattered archeological sites around the world testify to the role that clay has played since antiquity as a primary technological material. Clays are layered aluminosilicates, being composed primarily of Al, Si, O, and H with varying degrees of alkali, alkaline earths, or Fe. Some common clays found in nature include kaolinite, pyrophyllite, and talc. They are members of a mineral family called phyllosilicates that include micas, such as muscovite, as well as serpentines and chlorites. Clays are crystalline materials that have a small particle size. When combined with water they become hydroplastic (i.e., they are readily moldable). When heated, the particles fuse together while the overall macroscopic shape is retained. Upon cooling, the molded shape becomes the desired object.

There are two types of primary layers in the clay structure. One is a 0.22-nm layer composed of SiO_4 tetrahedra joined by their corners in a hexagonal array (Fig. W13.10*a*). The bases are coplanar and the tips of the tetrahedra all point in the same direction. At the vertices are either O atoms or OH radicals. The second primary layer is a 0.22-nm sheet of octahedra containing Al at the center which are sixfold coordinated with O atoms or OH radicals at the vertices (Fig. W13.10*b*). [In the case where there are only hydroxyl radicals, it is the mineral gibbsite, $Al_2(OH)_6$]. The various types of clay differ from each other in the number of these sheets, the

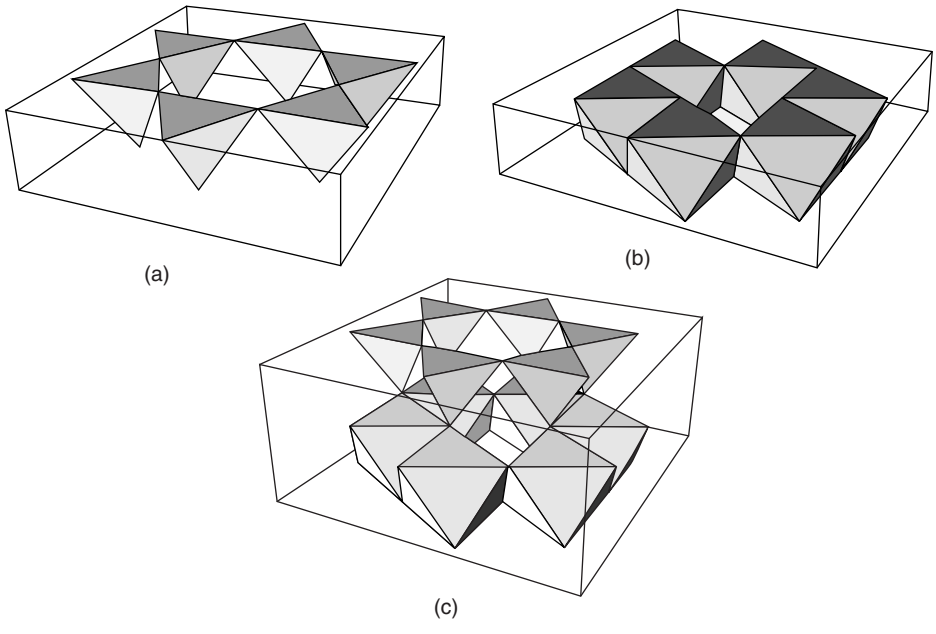

Figure W13.10. (*a*) Silica layer; (*b*) gibbsite layer; (*c*) kaolinite layer.

replacement of some Al or Si by other elements, or by the presence of sheets of water between the layers.

Kaolinite [$Al_2Si_2O_5(OH)_4$] has a 1:1 structure (i.e., the bilayer consists of one silica layer and one gibbsite layer). The overall thickness is 0.716 nm (0.22 nm for the tetrahedra + 0.22 nm for the octahedra + 0.276-nm spacing). The silica tetrahedra (SiO_4) point toward the gibbsite sheet, with the oxygens on the basal plane of the silica forming one outer surface and the hydroxyls of the gibbsite forming the second outer surface. The Al ions lie on a hexagonal lattice with two-thirds of the possible sites filled. Successive bilayers have the same orientation and are bound to each other by hydrogen bonding. A schematic of this arrangement (with the two sheets separated from each other for illustration purposes) is drawn in Fig. W13.10*c*. The atomic positions in the successive layers are sketched in Fig. W13.11. Figure W13.11*a* shows the basal O^{2-} plane with Si^{4+} atop the midpoint of the triangles formed by the oxygens; Fig. W13.11*b* shows O^{2-} ions above the Si^{4+} ions, completing the *tetrahedral layer* (T layer); Fig. W13.11*c* shows the positions of the Al^{3+} ions and OH^- ions in the same layer as the aforementioned O^{2-} ions. The OH^- layers lie above the voids in the basal layer. Finally, Fig. W13.11*d* shows a top layer with OH^- ions. Each Al^{3+} ion is surrounded by six negative ions. Below each Al^{3+} is a triangle with two O^{2-} ions and one OH^- ion. Above each Al^{3+} is a triangle of three OH^- ions. The orientation of the upper triangle is opposite to that of the lower triangle. The net result is that each Al^{3+} ion sits at the center of an octahedron. The layer is referred to as the *O* layer. The protons of the top OH^- layer are directed away from preceding *O* layer, ready to hydrogen-bond with the next T layer. Thus the stacking sequence in kaolinite may be denoted by TO–TO–TO–$\cdots$. The actual crystal structure is not orthorhombic, as in the sketch, but is slightly triclinic, with parallelipiped unit cell dimensions $(a, b, c) = (0.51, 0.89, 0.72)$ nm and angles $(\alpha, \beta, \gamma) = (91.8^\circ, 104.5^\circ, 90^\circ)$.

The lattice spacings in isolated gibbsite do not precisely match the lattice spacings in silica. When the two layers are brought into registry, one layer is compressed and the

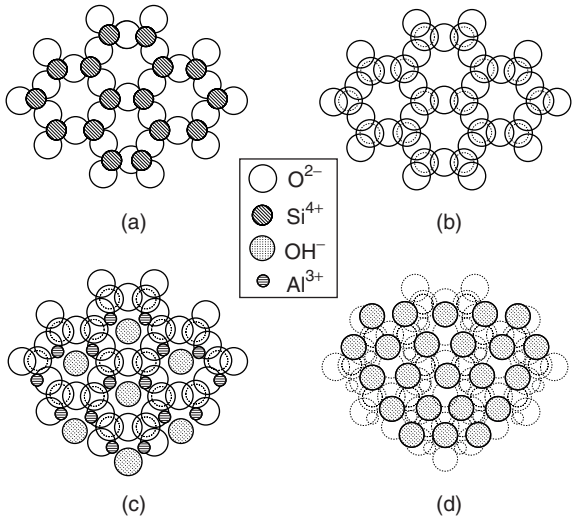

Figure W13.11. Layer-by-layer assembly of a kaolinite sheet. (Adapted from H. W. Jaffe, *Crystal Chemistry and Refractivity*, Dover, Mineola, N.Y., 1996.)

other is stretched. The resulting strain energy grows as the area of the layer increases. Eventually, the layers crack to relieve the strain energy. This limits the extent of the clay particles to a small size.

Pyrophyllite $[Al_2(Si_2O_5)_2(OH)_2]$ differs from kaolinite in that it contains two silica sheets instead of one (i.e., it has a 2:1 composition). The tetrahedra in the silica layers point inward toward the gibbsite core layer, so the outer surface of the trilayer structure consists of oxygen planes. Additional trilayers bond to this by weak van der Waals bonds. The unit cell is monoclinic with dimensions $(a, b, c) = (0.52, 0.89, 1.86)$ nm and angles $\alpha = \beta = 90^\circ$ and $\gamma = 99.9^\circ$.

Talc $[Mg_3(Si_2O_5)_2(OH)_2]$ has the same 2:1 structure as pyrophyllite, with the exception that the two Al^{3+} ions are replaced by three Mg^{2+} ions to maintain the valence requirements. Thus all the sites of the hexagonal lattice are now filled with Mg atoms, as opposed to the two-thirds occupancy for Al. Talc may be thought of as being based on the mineral brucite $[Mg_3(OH)_6]$ rather than on gibbsite, as before. It forms a monoclinic crystal with unit cell dimensions (0.53, 0.91, 1.89) nm and $\beta = 100^\circ$. Closely related is the clay montmorillonite, in which only some of the Al^{3+} are replaced by Mg^{2+} ions. Because of the valence mismatch, additional ions, such as Na^+, must also be incorporated, giving the composition $Al_{2-x}Mg_xNa_x(Si_2O_5)_2(OH)_2$. In the clay illite, some of the Si^{4+} ions are replaced by Al^{3+} ions. The valence mismatch is now compensated by adding K^+ ions to the hexagonal voids of the O layers. The structure is thus $Al_2(Si_{2-x}Al_xK_xO_5)_2(OH)_2$. In the special case where $x = 0.5$, the mica muscovite $[KAl_3Si_3O_{10}(OH)_2]$ is obtained. The K^+ ion serves to ionically bind adjacent trilayers tightly, thereby giving considerable rigidity to the structure.

W13.4 Cement

If limestone (calcite) is heated to 900°C, the reaction $CaCO_3 \rightarrow CaO + CO_2$ occurs and CaO (*quick lime*) is produced. When placed in contact with water, the CaO becomes hydrated and the product is called *slaked lime*. Heat is released, and the material swells and eventually hardens (sets). Mortar is a mixture of quick lime and sand (silica), which, when hydrated, forms a composite material that is used to bind bricks together.

Concrete, a composite material, is the primary structural material in use today. It consists of pebbles and sand bound together by cement.

In this section the focus will be on the most common type of cement, called *Portland cement*. The composition is 60 to 66% CaO (lime), 19 to 25% SiO_2 (silica), 3 to 8% Al_2O_3 (alumina), 1 to 5% Fe_2O_3 (ferrite), up to 5% MgO (magnesia) and 1 to 3% SO_3. When heated, four primary compounds are formed: dicalcium silicate (DCS) ($2CaO{\cdot}SiO_2$), tricalcium silicate (TCS) ($3CaO{\cdot}SiO_2$), tetracalcium aluminoferrite (TCAF) ($4CaO{\cdot}Al_2O_3{\cdot}Fe_2O_3$), and tricalcium aluminate (TCA) ($3CaO{\cdot}Al_2O_3$). Portland cement is, on average (by wt %), 46% TCS, 28% DCS, 8% TCAF, and 11% TCA. In addition, there is 3% gypsum ($CaSO_4{\cdot}2H_2O$), 3% magnesia, 0.5% K_2O or Na_2O, and 0.5% CaO. When water is added, a hydration reaction occurs and heat is generated. The hydrated particles conglomerate and a gel is formed. The cement sets in the course of time.

The four compounds provide various attributes to the cement. Thus DCS hardens slowly and improves the cement's strength after a considerable time (a week). TCS hardens more rapidly, gives the initial set, and provides early strength. TCA also provides early strength and dissipates early heat. TCAF reduces the "clinkering"