oersteds), which can be five times greater than observed in the conventional materials discussed earlier. These multilayer structures may consist of a sandwich of ferromagnetic metals such as NiFe, Co, or both, separated by a layer of Cu that can be 2 to 3 nm thick. One of the ferromagnetic layers is magnetically hardened so that its magnetic moment is pinned (i.e., unaffected by any magnetic fields to which it may be exposed in operation). This can be accomplished, for example, by exchange-coupling this layer to a thin antiferromagnetic layer such as MnFe, MnNi, or NiO through the mechanism of exchange biasing. Since the exchange coupling of the ferromagnetic layers through the 2-nm Cu spacer layer is relatively weak, the magnetic moment of the second, magnetically soft ferromagnetic sensing layer can rotate or switch directions in response to the magnetic field of the transition region on the magnetic disk. In this way the resistance of the magnetic sandwich changes, the presence of the bit is read, and the stored data are recovered. This type of magnetic structure is based on the giant magnetoresistance effect and is known as a *spin valve*. A dual-spin-valve structure that employs pinned films on each side of the sensing layer increases the response of the read head.

W17.13 Details on Magnetostrictive Materials

The specific materials with important magnetostrictive applications typically contain at least one magnetic rare earth element and often a magnetic transition metal element as well. Examples include Tb, Dy, and $Tb_{1-x}Dy_x$ alloys, Fe-based intermetallic compounds such as $TbFe_2$, $SmFe_2$, and the pseudobinary compound $Tb_{0.3}Dy_{0.7}Fe_2$, and Fe-based amorphous metallic glasses. Some values of the giant magnetostriction observed in these magnetic materials are presented in Table W17.5. Normal values of the dimensionless magnetostriction λ are in the range 10^{-6} to 10^{-5} for most ferromagnetic and ferrimagnetic materials.

TABLE W17.5 Magnetic Materials with Giant Magnetostrictions[a]

Material	$\frac{3\lambda_s}{2}(10^{-6})$
Dy (78 K)	1400
Tb (78 K)	1250
$TbFe_2$	2630
$SmFe_2$	−2340
$DyFe_2$	650
$Tb_{0.3}Dy_{0.7}Fe_2$ (Terfenol-D)	≈2300

Source: Data from K. B. Hathaway and A. E. Clark, *Mater. Res. Soc. Bull.*, Apr. 1993, p. 36.

[a]These data are for polycrystalline materials at room temperature, unless otherwise noted. The saturation magnetostriction $3\lambda_s/2$ is equal to $\lambda_\parallel - \lambda_\perp$. Here $\lambda_\parallel$ is the magnetostriction measured in the same direction as the applied field $\mathbf{H}$ [i.e., $\delta l(\theta = 0^\circ)/l$] of Eq. (17.29), while $\lambda_\perp$ is the magnetostriction measured in the same direction in the material but with $\mathbf{H}$ rotated by 90° [i.e., $\delta l(\theta = 90^\circ)/l$].

Rare Earth Metals and Alloys. Magnetostrictive strains of up to 10^{-2} have been observed in the rare earth metals Tb and Dy below their Curie temperatures T_C of 237 and 179 K, respectively. The magnetostriction of a $Tb_{0.6}Dy_{0.4}$ alloy is shown in Fig. W17.1 as a function of magnetic field. The magnetic and magnetostrictive behaviors of these lanthanide rare earth metals are determined by their partially filled $4f$ shell. The localized, highly anisotropic wavefunctions of the $4f$ electrons, in which the electron spin and orbital motion are strongly coupled to each other via the spin–orbit interaction, lead to strong magnetic anisotropies and also to high magnetostrictions. Note that the orbital part of the magnetic moment is not quenched (i.e., $L \neq 0$) in the rare earths. Of the $4f$ rare earth ions, Tb^{3+} and Dy^{3+} also have the advantage of having two of the largest observed magnetic moments, $9.5\mu_B$ and $10.6\mu_B$, respectively.

Intermetallic Compounds. Since the rare earth (RE) elements and alloys display giant magnetostrictions only below their T_C values (i.e., well below room temperature), considerable effort has gone into finding materials that have correspondingly high magnetostrictions at ambient temperatures. The most successful materials developed so far have been intermetallic compounds and alloys based on rare earths and Fe [e.g., $TbFe_2$ and $(Tb_{0.3}Dy_{0.7})Fe_2$]. These materials also have the advantage of T_C values, which increase as the rare earth concentration is increased.

At room temperature a giant magnetostriction corresponding to $\delta l/l \approx 10^{-3}$ to 10^{-2} has been observed in high magnetic fields in the magnetically hard cubic Laves-phase C15 intermetallic compound $TbFe_2$ ($T_C = 704$ K). The largest observed magnetostrictions occur in the $TbFe_2$ and $SmFe_2$ compounds in which the rare earth ions are highly anisotropic and also couple strongly to the Fe ions. The magnetostriction itself is highly anisotropic in these $REFe_2$ materials, with $|\lambda_{111}| \gg |\lambda_{100}|$. It follows that the orientation of the grains is very important for obtaining high magnetostrictions in polycrystalline $REFe_2$ alloys.

The ferromagnetic intermetallic compound $Tb_{0.3}Dy_{0.7}Fe_2$ (Terfenol-D) possesses a room-temperature giant magnetostriction of $\lambda \approx 10^{-3}$ even in low magnetic fields. The particular ratio of Dy to Tb chosen in this compound minimizes the magnetic anisotropy. If present, magnetic anisotropy would require high magnetic fields for magnetic saturation and the full magnetostriction to be achieved. This compensation of the magnetic anisotropy is possible because Tb and Dy have uniaxial magnetocrystalline anisotropy coefficients K_{u1} of opposite sign. The magnetic phase diagram for the pseudobinary $Tb_{1-x}Dy_xFe_2$ system is presented in Fig. W17.15. At high temperatures the alloys are cubic in the paramagnetic phase and become trigonal (rhombohedral) with the easy axes along the $\langle 111 \rangle$ directions in the ferrimagnetic phase below T_C. At the composition of Terfenol-D (i.e., $x = 0.7$) a transition to a tetragonal ferrimagnetic phase with spins aligned along the $\langle 100 \rangle$ directions occurs just below room temperature. Choosing a composition where operating at room temperature just above the rhombohedral-to-tetragonal transition is possible allows the alloys to have the desirable attribute of a large magnetostriction in low magnetic fields.

In transductor rods of Terfenol-D the stored magnetoelastic energy density is typically 130 to 200 kJ/m^3 and can be as high as 288 kJ/m^3 in (111) single crystals. These energy densities correspond to maximum strains of 1.6 to 2.4×10^{-3}. The fraction of the magnetic energy that can be converted to mechanical or elastic energy, and vice versa, is about 0.6 for Terfenol-D.

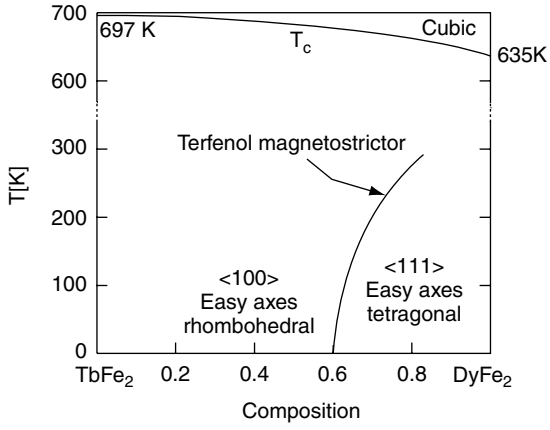

Figure W17.15. Magnetic phase diagram of the pseudobinary system $Tb_{1-x}Dy_xFe_2$. [From R. E. Newnham, *Mater. Res. Soc. Bull.*, **22**(5), 20 (1997). Courtesy of A. E. Clark.]

Terfenol-D can also be used in thin-film form for magnetostrictive sensors and transducers in microelectromechanical system (MEMS) technology. Amorphous films of Terfenol-D are magnetically soft and are preferred over crystalline films because the magnetostriction increases rapidly at low magnetic fields with only small hysteresis observed. Due to the high magnetostriction, the magnetic domain microstructure of these films is controlled by the film stress. When compressively stressed, the magnetization **M** in the domains is perpendicular to the film surface, while under tensile stress **M** lies in the plane of the film.

The mechanical damping in the films can be controlled by external magnetic fields since film stress is closely coupled to the direction of the magnetization **M**, and vice versa. Very high values of damping can be achieved by the application of a perpendicular magnetic field to a film under tensile stress as the direction of the magnetization is rotated from parallel to the film's surface to the perpendicular direction.

Fe-Based Amorphous Metallic Glasses. The conversion of magnetic to mechanical energy in amorphous Fe-based metallic glasses (e.g., a Metglas alloy of composition $Fe_{81}B_{13.5}Si_{3.5}C_2$) can be as high as 90% when the amorphous ribbons are annealed in a transverse magnetic field and then cooled rapidly. In this state the ribbons have an induced tranverse magnetic anisotropy. When placed in a longitudinal magnetic field, the domain magnetizations rotate smoothly from the perpendicular to the parallel direction, with no motion of domain walls. The rotation can be accomplished in very low applied fields due to the low anisotropy fields H_K that can be achieved in these amorphous materials. The ribbons elongate due to their positive magnetostriction.

W17.14 Dilute Magnetic Semiconductors

An interesting class of magnetic materials from a fundamental point of view is the group II–VI semiconductors, such as ZnS, ZnSe, CdS, CdTe, HgS, and HgTe, diluted with Mn atoms which enter these zincblende structures as random substitutional replacements for the divalent Zn or Hg ions. In $Zn_{1-x}Mn_xS$ or $Hg_{1-x}Mn_xTe$, the Mn^{2+} ions with spin $S = \frac{5}{2}$ interact antiferromagnetically with each other via an indirect superexchange

interaction through the bonding electrons associated with the S or Te anions. The Mn^{2+} ions also interact with the conduction-band s and p electrons via the sp–d interaction. This is essentially just the s–d interaction described in Chapter 9, which plays a critical role in the indirect RKKY interaction between pairs of magnetic ions in metals.

The magnetic behavior of these dilute magnetic semiconductors is paramagnetic for low Mn concentrations (e.g., $x \approx 0.15$ to 0.2 for $Cd_{1-x}Mn_xTe$). At higher Mn concentration the behavior corresponds to that of a disordered antiferromagnet (i.e., a type of spin glass in a semiconducting host). The sp–d interaction leads to interesting electrical and optical properties for the s and p conduction-band electrons, including a pronounced magnetoresistance and also a giant Faraday rotation. Potential optoelectronic applications for these materials include their use in display technologies and as infrared detectors, magneto-optical materials, and quantum-well lasers. Other applications of these materials may involve exploiting the spin of the electron in solid-state devices, an area known as *spintronics*. So far it has proven to be difficult to dope these II–VI magnetic semiconductors n- and p-type.

Recently, it has been possible to deposit films of $Ga_{1-x}Mn_xAs$ with Mn concentrations above the solubility limit via low-temperature molecular beam epitaxy. The Mn atoms in these alloys provide both magnetic moments and hole doping.

REFERENCES

Aharoni, A., *Introduction to the Theory of Ferromagnetism*, Clarendon Press, Oxford, 1996.

Chikazumi, S., *Physics of Magnetism*, Wiley, New York, 1964.

Craig, A. E., Optical modulation: magneto-optical devices, in K. Chang, ed., *Handbook of Microwave and Optical Components*, Vol. 4, Wiley, New York, 1991.

PROBLEMS

W17.1 **(a)** Derive the results for the domain width d and energy U given in Eqs. (W17.3) and (W17.4), respectively.

(b) Show also that U given in Eq. (W17.4) for the domain structure shown in Fig. 17.2*b* will be lower than U_m for a single domain given in Eq. (17.4) as long as the thickness t is not too small. Calculate the value of the critical thickness t_c.

(c) Use the parameters appropriate for Fe at $T = 300$ K to calculate t_c. [*Hint:* See the data for Fe at $T = 300$ K given following Eq. (17.6).]

W17.2 **(a)** For the precession of the magnetization vector **M** in a magnetic field **H** in the z direction, as expressed by equation of motion (W17.17) and shown schematically in Fig. W17.5, show that the three components of **M** have the following equations of motion:

$$\frac{dM_x}{dt} = -\gamma\mu_0 M_y H, \qquad \frac{dM_y}{dt} = +\gamma\mu_0 M_x H, \qquad \frac{dM_z}{dt} = 0.$$

(b) Using the trial solutions $M_x(t) = M_\perp \cos\omega t$ and $M_y(t) = M_\perp \sin\omega t$, show that $\omega = \omega_r = \gamma\mu_0 H$.

(c) Calculate ω_r for $g = 2$ and $H = 10^3$ kA/m. To what type of electromagnetic radiation does this correspond?

W17.3 Consider a permanent magnet in the form of a toroid with an air gap, as shown schematically in Fig. W17.6.

(a) If l_g and A_g are the length and cross-sectional area of the air gap, respectively, and l and A are the corresponding values for the magnet, use elementary equations of electromagnetic theory (i.e., $\oint \mathbf{H} \cdot d\mathbf{l} = \mu_0 I$ and $\int \mathbf{B} \cdot d\mathbf{A} = \Phi$) to show that $B/H = -B_g l A_g / H_g l_g A = -\mu_0 l A_g / l_g A$, where $B_g = \mu_0 H_g$ corresponds to the induction in the air gap and $B = \mu H$ corresponds to the induction in the magnet.

(b) By comparing this result with Eq. (W17.23), show that $(1 - N)/N = lA_g/Al_g$.

(c) Show that the limit $N \ll 1$ corresponds to $l_g \ll l$ [e.g., a very narrow air gap (assuming that $A_g \approx A$)].

W17.4 For a certain permanent magnet the demagnetization curve in the second quadrant of the B–H loop can be described approximately by $B(H) = B_r(1 - |H|^2/H_c'^2)$ with $B_r = 1.25$ T and $H_c' = 500$ kA/m.

(a) Calculate the maximum energy product $(BH)_{\max}$ for this material in units of kJ/m^3.

(b) What demagnetization coefficient N should be chosen for this magnet so that in the absence of an external magnetic field, $(BH) = (BH)_{\max}$ at its operating point?

(c) What is the magnetization M in the magnet at this operating point?

CHAPTER W18

Optical Materials

W18.1 Optical Polarizers

A polarizer is basically an optically anisotropic material for which the transmission depends on the direction of polarization of the light relative to the crystal axes. The ability to control the polarization permits one to build such optical elements as modulators and isolators.

Suppose that a plane electromagnetic wave propagates along the z direction. The electric field vector lies in the xy plane and may be characterized by two complex amplitudes: $\mathbf{E}_0 = E_{0_x}\hat{i} + E_{0_y}\hat{j}$. The intensity of the light (i.e., its power per unit area), is written as

$$I = \sqrt{\frac{\epsilon}{\mu}}|E_0|^2 = \sqrt{\frac{\epsilon}{\mu}}(|E_{0_x}|^2 + |E_{0_y}|^2) = I_x + I_y, \tag{W18.1}$$

where I_x and I_y are the intensities of x and y polarized light. If I_x and I_y are the same, the light is said to be *unpolarized*. If they are different, the light may be *linearly polarized*. The degree of linear polarization, P_L, is given by

$$P_L = \frac{I_x - I_y}{I_x + I_y}, \tag{W18.2}$$

where it is assumed that $I_x \geq I_y$ so as to make $0 \leq P_L \leq 1$. If $I_y = 0$, then $P_L = 1$ and there is 100% linear polarization. If $P_L = 0$ the light is unpolarized. If $0 < P_L < 1$, the light is partially linearly polarized.

A more detailed description of the light involves information concerning the relative phases of the electric field components as well as the intensity and degree of polarization. It is convenient to construct the complex column vector

$$\chi_0 = \begin{bmatrix} E_{0_x} \\ E_{0_y} \end{bmatrix} \tag{W18.3}$$

and form the two-dimensional matrix, called the *density matrix*,

$$\chi_0\chi_0^+ = \begin{bmatrix} E_{0_x}E_{0_x}^* & E_{0_x}E_{0_y}^* \\ E_{0_y}E_{0_x}^* & E_{0_y}E_{0_y}^* \end{bmatrix}. \tag{W18.4}$$

(If the light is fluctuating in time, one generally performs a time average and replaces $\chi_0\chi_0^*$ by $\langle\chi_0\chi_0^*\rangle$.) Note that the matrix is Hermitian (i.e., its transpose is equal to

its complex conjugate). A general complex two-dimensional matrix needs eight real numbers to specify its elements, but the Hermitian condition reduces this number to four. This matrix may be expanded in terms of four elementary Hermitian matrices. The Pauli spin matrices (used coincidentally to describe the electron spin operator in Appendix WC) and the identity matrix are chosen for this purpose. Thus multiplying the column vector χ_0 by the row vector χ_0^+ formed from the two complex conjugate elements gives

$$\rho_0 = \chi_0 \chi_0^+ = \tfrac{1}{2}(S_0^0 \mathbf{I} + \mathbf{S}^0 \cdot \boldsymbol{\sigma}), \tag{W18.5}$$

where

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{W18.6}$$

The real numbers $S_i^0 (i = 0, 1, 2, 3)$ are called the *Stokes parameters* and fully characterize the state of polarization, including the relative phase relations. They are given by

$$S_0^0 = |E_{0_x}|^2 + |E_{0_y}|^2, \tag{W18.7a}$$

$$S_3^0 = |E_{0_x}|^2 - |E_{0_y}|^2, \tag{W18.7b}$$

$$S_1^0 = E_{0_x} E_{0_y}^* + E_{0_y} E_{0_x}^*, \tag{W18.7c}$$

$$S_2^0 = i(E_{0_x} E_{0_y}^* - E_{0_y} E_{0_x}^*). \tag{W18.7d}$$

From Eq. (W18.1) one sees that S_0^0 is proportional to the intensity, I. The quantity $P_L = S_3^0/S_0^0$ is the degree of linear polarization and $P_C = S_2^0/S_0^0$ is the degree of circular polarization. The degree of total polarization is given by $P_T = \sqrt{P_C^2 + P_L^2}$. The Stokes parameter S_1^0 contains information concerning the relative phase of the x- and y-polarized light, or equivalently, between the right- and left-circularly polarized light.

Consider the transmission of unpolarized light through a polarizer. Assume for the moment that the principal axes of the polarizer are aligned with the x and y axes. After transmission, the field is changed to $\mathbf{E} = E_x \hat{i} + E_y \hat{j}$, where the new amplitudes are related to the old amplitudes by

$$E_x = E_{0_x} e^{i\phi_x} p_x, \qquad E_y = E_{0_y} e^{i\phi_y} p_y. \tag{W18.8}$$

The parameters p_x and p_y are dimensionless attenuation constants, depending on the absorption coefficients when the electric field is directed along the principal optical axes. Thus $p_x = \exp(-\alpha_x L)$ for a polarizer of thickness L, and similarly for p_y. These coefficients may be frequency dependent, a phenomenon called *dichroism*. Henceforth, as a simplification, it will assumed that the phase factors ϕ_x and ϕ_y are zero.

The Stokes parameters may be arranged as a four-element vector and the effect of the polarizer will then be described by a four-dimensional matrix called the 4×4

Mueller matrix, M,

$$\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} p_x^2 + p_y^2 & 0 & 0 & p_x^2 - p_y^2 \\ 0 & p_x p_y & 0 & 0 \\ 0 & 0 & p_x p_y & 0 \\ p_x^2 - p_y^2 & 0 & 0 & p_x^2 + p_y^2 \end{bmatrix} \begin{bmatrix} S_0^0 \\ S_1^0 \\ S_2^0 \\ S_3^0 \end{bmatrix} \equiv \begin{bmatrix} A & 0 & 0 & B \\ 0 & C & 0 & 0 \\ 0 & 0 & C & 0 \\ B & 0 & 0 & A \end{bmatrix} \begin{bmatrix} S_0^0 \\ S_1^0 \\ S_2^0 \\ S_3^0 \end{bmatrix}. \tag{W18.9}$$

If the principal axes were rotated with respect to the x and y axes by angle θ, this could be described by rotating the M matrix by the rotation matrix T:

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \sin 2\theta & 0 & \cos 2\theta \\ 0 & 0 & 1 & 0 \\ 0 & -\sin 2\theta & 0 & \cos 2\theta \end{bmatrix}, \tag{W18.10}$$

and the Mueller matrix becomes

$$M(\theta) = TMT^{-1} = \begin{bmatrix} A & B\sin 2\theta & 0 & B\cos 2\theta \\ B\sin 2\theta & A\sin^2 2\theta + C\cos^2 2\theta & 0 & (A-C)\sin 2\theta\cos 2\theta \\ 0 & 0 & C & 0 \\ B\cos 2\theta & (A-C)\sin 2\theta\cos 2\theta & 0 & A\cos^2 2\theta + C\sin^2 2\theta \end{bmatrix}. \tag{W18.11}$$

Various types of polarizing sheets have been devised. They are generally based on the use of dichromophore molecules (i.e., molecules that produce dichroism). The *H-sheet*, invented by E. H. Lamb, consists of molecules of polyvinyl alcohol (PVA) stretched along a particular direction, to which an iodine-based dye is added. When light has its electric field parallel to the long axis of the molecules, they become polarized and develop large fluctuating electric-dipole moments. This sets up large local fields near the molecules and their excitation is readily transferred to the iodine-based dye molecules, where the energy is absorbed and thermalized. Light oriented perpendicular to the molecules does not cause as large a polarization and is therefore not transferred to the dye efficiently. Consequently, the perpendicularly polarized light is transmitted with higher efficiency than light oriented parallel to the PVA molecules. The PVA molecules are in laminated sheets consisting of cellulose acetate butyrate for mechanical support and chemical isolation.

Later the *J-sheet* was introduced, consisting of needlelike dichroic crystals of herapathite oriented parallel to each other in a matrix of cellulose acetate. A variation of this is the *K-sheet*, in which rather than achieving dichroism by adding a stain (an additive that absorbs at a particular color or colors), hydrogen and oxygen are removed by a dehydration catalyst. The material is stretched to produce aligned polyvinylene polymers. Another variation, the *L-sheet*, relies on organic dye molecules to achieve the dichroism. Typical dye molecules are aminil black, Erie green, Congo red, and Niagara blue. It is also possible to embed thin parallel metal wires in a substrate to create a polarizer. Typically, fine Al wires are placed in substrates of glass, quartz, or polyethylene.

For a dichromophore molecule or crystallite to be successful, it must exhibit a large anisotropy. In combination with the dye molecule it must be strongly absorbing for one state of polarization and weakly absorbing for the other state.

An example of the spectral dependence of the polarization parameters on wavelength is given in Fig. W18.1, where p_x^2 and p_y^2 are presented for the polarizer KN-36 (a

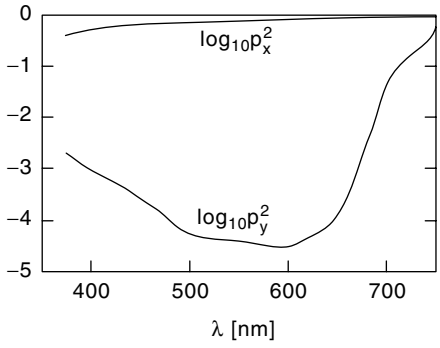

Figure W18.1. Spectral parameters p_x^2 and p_y^2 plotted as a function of the wavelength λ for the polarizer KN-36. (Adapted from E. Collett, *Polarized Light*, Marcel Dekker, New York, 1993.)

commercial polarizer of the K-sheet variety). The filter is called a *neutral polarizer* because these parameters are approximately flat across the visible spectrum.

It should be noted that the concept of a polarizer may be extended to any device that modifies the Stokes parameters of the transmitted light. A large number of physical parameters is associated with the Mueller matrix of the device. Full characterization of a general polarizer is rarely given.

W18.2 Faraday Rotation

In Section W18.1 polarization of light was obtained by means of dichroism. In this section attention is given to how the direction of polarization may be changed with little attenuation. The polarization of an electromagnetic wave is rotated when it propagates through a medium along the direction of a magnetic field, a phenomenon called *Faraday rotation*. The angle of rotation, θ_F, is determined by the magnetic induction or flux density, $\mathbf{B} = B\hat{k} = \mu_0 H\hat{k}$, the length of propagation, z, and the Verdet constant of the material, V:

$$\theta_F = VHz. \tag{W18.12}$$

The process is illustrated in Fig. W18.2.

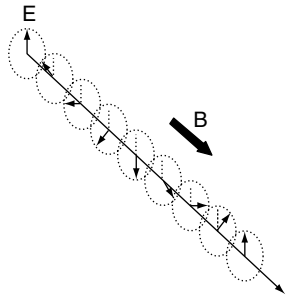

Figure W18.2. Rotation of the electric polarization vector of light propagating along a magnetic field.

To obtain an expression for V, one may model the electrons as a collection of Lorentz oscillators interacting with the light and the magnetic field imposed. The model is general enough to include both bound and free electrons. The classical equation of motion for an oscillator is

$$\left[\frac{d^2}{dt^2} + \gamma\frac{d}{dt} + \omega_0^2\right]\mathbf{r}(t) = -\frac{e}{m_c}\left(\mathbf{E}(t) + \frac{d\mathbf{r}}{dt} \times \mathbf{B}\right), \tag{W18.13}$$

with $\mathbf{B}$ along the positive z direction. For free electrons m_c is the cyclotron effective mass of the electrons (see Problem W18.1), whereas for bound electrons m_c is replaced by the free-electron mass, m. If the electrons are bound, then ω_0 represents an electronic resonance frequency of the medium, while for free electrons it may be taken to be zero. Assuming harmonic variations for $\mathbf{E}(t)$ and $\mathbf{r}(t)$ of the form $\exp(-i\omega t)$, one obtains the following equations for the amplitudes x and y:

$$(\omega_0^2 - \omega^2 - i\omega\gamma)x = -\frac{e}{m_c}(E_x - i\omega By) \tag{W18.14a}$$

$$(\omega_0^2 - \omega^2 - i\omega\gamma)y = -\frac{e}{m_c}(E_y + i\omega Bx). \tag{W18.14b}$$

Letting $x_\pm = x \pm iy$, $E_\pm = E_x \pm iE_y$, and $\omega_c = eB/m_c$ (the cyclotron frequency) gives

$$x_\pm(\omega) = -\frac{e}{m_c}\frac{E_\pm}{\omega_0^2 - \omega^2 - i\omega\gamma \mp \omega\omega_c}. \tag{W18.15}$$

The polarization vector of the medium is expressed similarly as

$$P_\pm = -nex_\pm = \chi_\pm\epsilon_0 E_\pm, \tag{W18.16}$$

where n is the concentration of oscillators. The relative permittivity or dielectric constant is $\epsilon_{r_\pm} = 1 + \chi_\pm$.

The wave vector is different for right- and left-circularly polarized light: $k_\pm = \omega\sqrt{\epsilon_{r_\pm}}/c$. Introducing the dielectric function for zero magnetic field,

$$\epsilon_{r_0} = 1 - \frac{\omega_p^2}{\omega^2 - \omega_0^2 + i\omega\gamma}, \tag{W18.17}$$

where ω_p is the plasma frequency, one finds that

$$\epsilon_{r_\pm} = 1 - \frac{1 - \epsilon_{r_0}}{1 \pm (\omega\omega_c/\omega_p^2)(1 - \epsilon_{r_0})}. \tag{W18.18}$$

To first order in B, the difference in the wave vectors is

$$k_+ - k_- = \frac{\omega_c}{c}\left(\frac{\omega}{\omega_p}\right)^2 \frac{(1 - \epsilon_{r_0})^2}{\sqrt{\epsilon_{r_0}}}. \tag{W18.19}$$

After propagating a distance z through the medium, this leads to a phase-angle difference,

$$\theta_F = (k_+ - k_-)z = \frac{e}{m_c c}\left(\frac{\omega}{\omega_p}\right)^2 \frac{\left(1 - \varepsilon_{r_0}\right)^2}{\sqrt{\epsilon_{r_0}}}Bz. \tag{W18.20}$$

The Verdet constant is therefore

$$V = \frac{e}{m_c c}\left(\frac{\omega}{\omega_p}\right)^2 \frac{(1-\epsilon_{r_0})^2}{\sqrt{\epsilon_{r_0}}} \approx \frac{ne^3}{m_c^2 c\epsilon_0} \frac{\omega^2}{(\omega^2-\omega_0^2)^{3/2}\sqrt{\omega^2-\omega_0^2-\omega_p^2}}, \qquad \text{(W18.21)}$$

where the damping constant is neglected in the last expression.

This formula displays the factors influencing the size of the Verdet constant: the concentration of oscillators, the cyclotron effective mass of the carriers, and the resonance frequency relative to that of the light. In semiconductors, the effective mass could be small and the value of V could be large. In the neighborhood of an electronic resonance, the value of V could likewise become large.

Typical values for the Verdet constant for several nonmagnetic materials are presented in Table W18.1. It is customary to express V in arc-minutes/Oe·m, where $1\ \text{Oe} = 1,000/4\pi$ A/m. A magnetic induction of $B = 4\pi \times 10^{-7}$ T corresponds to a field intensity H of 1 A/m. The Faraday and Kerr effects in magnetic materials are discussed in Chapter 17 of the textbook.† Magneto-optical applications are also given there.

An optical isolator may be constructed from a polarizer and Faraday rotator that rotates the polarization vector by 45°. If light is partially reflected from some interface

TABLE W18.1 Verdet Constants for Several Nonmagnetic Materials

Material	λ (nm)	V (arc-min/Oe·m)	
Diamond	589.3	2.3	
NaCl	589.3	3.5	
KCl	589.3	2.8	
SiO_2	589.3	1.7	
B_2O_3	633	1.0	
Al_2O_3	546.1	2.4	
$SrTiO_3$	620	14	
ZnSe	476	150	
	496	104	
	514	84	
	587	53	
	633	41	
$Tb_2Al_5O_{12}$	520	−103.9	(300 K)
	520	−343	(77 K)
	520	−6480	(4.2 K)
KH_2PO_4 (KDP)	632.8	1.24	

Source: Data from M. J. Weber, *Handbook of Laser Science and Technology*, Vol. 4, CRC Press, Boca Raton, Fla., 1986; and D. R. Lide, ed., *CRC Handbook of Chemistry and Physics*, 75th ed., CRC Press, Boca Raton, Fla., 1994.

† The material on this home pate 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."

after it passes through the isolator, the direction of its electric field vector will be reversed by the reflection. As it propagates backward through the Faraday rotator, the electric field vector will experience a further 45^o rotation. Since the field will then be perpendicular to the polarizer, it will be blocked by it. This prevents the reflected light from propagating backward and possibly causing damage to optical components.

W18.3 Theory of Optical Band Structure

Band-structure engineering may be applied to more complex structures than were considered in Section 18.6. In this section an analysis is given of one such structure, consisting of a one-dimensional periodic array. Each unit cell of the array contains two layers of transparent material with different indices of refraction. The propagation of electron waves in one-dimensional periodic structures is studied in Chapter 7, and it forms the basis for understanding the band theory of solids. Here the concept is extended to the optical case.

Consider the passage of light through two materials in the case where the photon energy is less than the bandgap. Barring any other absorption processes, both materials would, separately, be transparent. Next construct a stratified structure in which alternate layers of the two materials are stacked in a periodic fashion. It will be shown that for some wavelengths, propagation cannot occur and the structure acts as a mirror. Other colors, however, will pass through and the structure therefore acts as a color-selective filter. These effects come about due to the destructive and constructive interference of reflected light waves, in much the same way as electronic band structure results from the interference of scattered electron waves in solids.

Let the indices of refraction for the two materials be n_1 and n_2, and let the thickness of layer n_1 be b and the thickness of layer n_2 be $a - b$. The structure has a periodicity of size a (Fig. W18.3). For transverse waves propagating along the x direction, the problem of wave propagation reduces to solving the Helmholtz equation $[\nabla^2 + k^2(x)]E = 0$, where $k_1 = \omega n_1/c$, $k_2 = \omega n_2/c$, and E is the electric field of the light. The solution in medium 1 is

$$E(x) = A_j e^{ik_1(x-ja)} + B_j e^{-ik_1(x-ja)} \qquad \text{if } ja < x < ja + b, \tag{W18.22a}$$

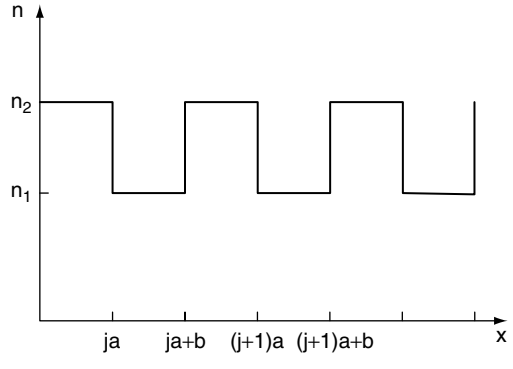

Figure W18.3. Stratified layers of optically transparent materials.

and in medium 2 is

$$E(x) = C_j e^{ik_2(x-ja)} + D_j e^{-ik_2(x-ja)} \qquad \text{if } ja + b < x < ja + a. \tag{W18.22b}$$

Matching E and dE/dx at $x = ja + b$ yields

$$A_j e^{ik_1 b} + B_j e^{-ik_1 b} = C_j e^{ik_2 b} + D_j e^{-ik_2 b}, \tag{W18.23a}$$

$$k_1 A_j e^{ik_1 b} - k_1 B_j e^{-ik_1 b} = k_2 C_j e^{ik_2 b} - k_2 D_j e^{-ik_2 b}. \tag{W18.23b}$$

Repeating the match at $x = (j + 1)a$ yields

$$A_{j+1} + B_{j+1} = C_j e^{ik_2 a} + D_j e^{-ik_2 a}, \tag{W18.24a}$$

$$k_1 A_{j+1} - k_1 B_{j+1} = k_2 C_j e^{ik_2 a} - k_2 D_j e^{-ik_2 a}. \tag{W18.24b}$$

Let

$$\xi_1 = e^{ik_1 a}, \qquad \xi_2 = e^{ik_2 a}, \qquad \eta_1 = e^{ik_1 b}, \qquad \eta_2 = e^{ik_2 b}. \tag{W18.25}$$

After eliminating C_j and D_j from the equations above, one arrives at the recurrence formula

$$\begin{pmatrix} A_{j+1} \\ B_{j+1} \end{pmatrix} = M \begin{pmatrix} A_j \\ B_j \end{pmatrix}, \tag{W18.26}$$

where the 2×2 transfer matrix $\mathbf{M}$ is

$$M = \frac{1}{4k_1 k_2} \begin{pmatrix} (k_1 + k_2)^2 \eta_2^* \eta_1 \xi_2 & (k_2^2 - k_1^2)\eta_1^* \eta_2^* \xi_2 \\ -(k_1 - k_2)^2 \xi_2^* \eta_1 \eta_2 & -(k_2^2 - k_1^2)\xi_2^* \eta_1^* \eta_2 \\ (k_1^2 - k_2^2)\eta_2^* \eta_1 \xi_2 & -(k_2 - k_1)^2 \eta_1^* \eta_2^* \xi_2 \\ -(k_1^2 - k_2^2)\xi_2^* \eta_1 \eta_2 & +(k_1 + k_2)^2 \xi_2^* \eta_1^* \eta_2 \end{pmatrix}. \tag{W18.27}$$

Note that $\mathbf{M}$ is independent of the index j. The sum of the diagonal elements is called the *trace*:

$$\mathrm{Tr}(\mathbf{M}) = \frac{1}{4k_1 k_2}[(k_1 + k_2)^2(\eta_2^* \eta_1 \xi_2 + \eta_2 \eta_1^* \xi_2^*) - (k_1 - k_2)^2(\xi_2^* \eta_1 \eta_2 + \xi_2 \eta_1^* \eta_2^*)]. \tag{W18.28}$$

The determinant of the $\mathbf{M}$ matrix is 1.

The eigenvalues of the $\mathbf{M}$ matrix are defined as the roots of the characteristic equation

$$\begin{vmatrix} M_{11} - \mu & M_{12} \\ M_{21} & M_{22} - \mu \end{vmatrix} = 0 = \mu^2 - \mu\ \mathrm{Tr}(M) + 1, \tag{W18.29}$$

and are

$$\mu_\pm = \frac{1}{2}\mathrm{Tr}(M) \pm \sqrt{\left(\frac{1}{2}\mathrm{Tr}(M)\right)^2 - 1}. \tag{W18.30}$$

The product of the two eigenvalues is equal to 1, the determinant. If both eigenvalues are real, one of them is larger than 1 and the other is smaller than 1. On the other

hand, if one of the eigenvalues is complex, the other is its complex conjugate and each eigenvalue has magnitude 1. If the eigenvalue is real, repeated application of the transfer matrix will cause the amplitudes A_j and B_j to grow exponentially with increasing j, leading to an unphysical situation. Under such circumstances, propagation is not possible. The condition for propagation is therefore that $\mu_\pm$ be complex [i.e., that $(Tr\mathbf{M})^2 < 4$]. This will define what is called a *propagation band*. The condition may be recast as the condition

$$\{(k_1 + k_2)^2 \cos[(k_2 - k_1)b - k_2 a] - (k_1 - k_2)^2 \cos[(k_2 + k_1)b - k_2 a]\}^2 < (4k_1 k_2)^2. \tag{W18.31}$$

In Fig. W18.4 the allowed propagation band for the special case $b = a/2$ is illustrated. Let

$$k = \frac{k_1 + k_2}{2}, \qquad q = \frac{k_2 - k_1}{2}, \qquad x = \frac{ka}{2}, \qquad y = \frac{qa}{2}. \tag{W18.32}$$

Then the propagation-band conditions are

$$y^2 \cos^2 y < x^2 \cos^2 x, \qquad y^2 \sin^2 y < x^2 \sin^2 x. \tag{W18.33}$$

Some wavelengths are able to propagate through the structure and others are blocked.

Typical materials for use in these devices, which may serve as either mirrors or filters, are TiO_2 ($n = 2.4$) and SiO_2 ($n = 1.46$). Other combinations are MgF_2 ($n = 1.38$) and ZnS ($n = 2.35$) or MgF_2 with TiO_2. A one-dimensional array of air holes in a Si strip on top of an SiO_2 substrate has been fabricated[†] which displays a 400-nm gap centered around $\lambda = 1.54$ μm.

To withstand bursts of light that may arise in pulsed lasers, one generally wants matched coefficients of thermal expansion and high thermal conductivity. The reason is that mismatched thermal expansion between successive layers will generate strains upon heating that could produce dislocations at the interface. Repeated thermal expansion may enlarge these dislocations and could eventually crack the material. The high

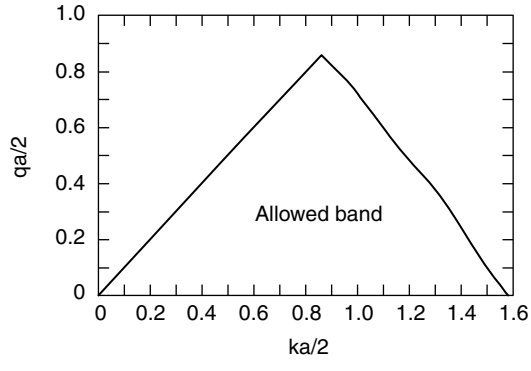

Figure W18.4. Region of parameter space for the propagation band.

[†] J. S. Foresi et al, *Nature*, **390**, 143(1997).

thermal conductivity permits the material to cool rapidly. Optical damage is considered further in Section W18.4.

The extension of the periodic structure to two or three dimensions has led to the construction of what are called *photonic crystals*. By creating an array of holes in a dielectric slab a photonic crystal operating in the microwaves has been built.[†] By stacking Si rods in a face-centered tetragonal array with air filling the interstices, it has been possible to fabricate[‡] a photonic crystal with a bandgap in the infrared (10 to 14.5 μm). Similarly, a periodic array of air-filled spheres in a titania crystal has been fashioned to serve as a photonic crystal in the visible region of the spectrum.[§]

Just as electrons may be localized in a medium with random scatterers, the same is true of electromagnetic radiation. Localization in the microwave region has been demonstrated by using a three-dimensional metal-wire network with random scatterers.[¶] It is clear that band-structure engineering is still at its early stage of development and that new and exciting developments are rapidly emerging in the field.

W18.4 Damage

Laser damage to optical components, such as laser crystals, mirrors, polarizers, fibers, electro-optic crystals, and prisms, is of concern in applications involving high power, in both pulsed and continuous wave (CW) operation. Due to the optical absorption, the materials heat up. Materials with a low heat capacity and low thermal conductivity are more likely to reach high temperatures. In layered structures the mismatch in thermal expansion coefficients can lead to crack formation and propagation. Typically, bulk damage results for 10-ns pulses when the power density is in the range 200 to 4000 TW/m^2.

One of the prime concerns is the phenomenon of self- focusing. This can occur in a medium with a positive value of the nonlinear index of refraction, $n_2 I$. A laser beam generally has a cross-sectional intensity profile with a higher intensity, $I(R)$, near the axis than away from it. A typical form for the profile is Gaussian; that is,

$$I(R) = \frac{2P_0}{\pi f^2} e^{-2(R/f)^2}, \tag{W18.34}$$

where R is the radial distance, P_0 the power in the beam, and f a measure of the beam radius. The nonlinearity causes a larger value for the index of refraction, $n(R) = n_1 + n_2 I(R)$, near the axis, when $n_2 > 0$. The medium behaves as a lens, and this tends to focus the radiation [i.e., make $f(z)$ decrease with increasing propagation distance, z]. However, there is a competing effect due to diffraction, which tends to defocus the radiation. This defocusing effect becomes stronger the smaller the value of f. There exists a critical value of P_0 for which the focusing effect of the nonlinearity dominates over the defocusing effect of diffraction and the beam focuses. When it does so, the focal spot can become as small as a wavelength of light and the intensity can become

[†] E. Yablonovitch et al, *Phys. Rev. Lett.*, **67**, 2295 (1991).
[‡] S. Y. Lin et al, *Nature*, **394**, 251 (1998).
[§] J. Wijnhoven and W. Vos, *Science*, **281**, 803 (1998).
[¶] M. Stoychev and A. Z. Genack, *Phys. Rev. B*, **55**, R8617 (1997).

very large. A crude estimate of the critical power may be obtained by setting $f = 1/k$, where k is the wave vector, and setting $n_1 \approx n_2 I$. This gives $P_{\text{cr}} \sim n_1/n_2 k^2$.

Often, the electric field of the light can exceed the strength of the typical electric fields in the solid and electrons can be accelerated to high energies, causing radiation damage such as atomic displacements. The highly concentrated beam could cause local melting, vaporization or ionization.

The situation is exacerbated when there are preexisting cracks or dislocations in the material. When subjected to the (uniform) electric field of the laser, the local electric field in the vicinity of the defect could be nonuniform, with particularly strong fields being generated near sharp features. The same effects occur near a lightning rod, where the strongest field occurs near the sharpest point. Local breakdown is likely to occur near the defect, often inflicting additional damage there.

Defects are usually introduced into optical components during their fabrication stage. For example, YAG is seen to have edge dislocations, helical dislocations, and zigzag dislocations. Laser crystals are often plagued by secondary phases of crystals mixed in with the primary phase. Bubbles are often present. These larger features can also serve as scattering centers which deplete the laser beam of power and couple their signals to other optical components. For this reason it is important that the optical components be largely free of defects before being used in high-power applications.

REFERENCES

Polarized Light

Collett, E., *Polarized Light: Fundamentals and Applications*, Marcel Dekker, New York, 1993.

Shurcliff, W. A., *Polarized Light*, Harvard University Press, Cambridge, Mass., 1962.

PROBLEM

W18.1 The effective-mass tensor for an electron is diagonal in the xyz-coordinate system and has elements m_1^*, m_2^*, and m_3^*. A magnetic induction $\mathbf{B}$ is directed in an arbitrary direction. If the cyclotron resonance frequency is eB/m_c, find an expression for m_c.

CHAPTER W19

Surfaces

W19.1 Surface States

It is possible to introduce Tamm surface states by adding an attractive delta function potential of strength U to the step potential introduced in Eq. (19.3):[†]

$$V(z) = -V_0\Theta(-z) - U\delta(z). \tag{W19.1}$$

Note that the units of U are J·m and that of V_0 are joules. The independent variables in the Schrödinger equation can be separated with the substitution

$$\psi(\mathbf{r}) = \phi(z)\exp(i\mathbf{k}_\parallel \cdot \mathbf{r}_\parallel) \tag{W19.2}$$

where a solution can be found with

$$\phi(z) = \begin{cases} \exp(-\kappa z) & \text{if} \quad z > 0, \\ \exp(+qz) & \text{if} \quad z < 0. \end{cases} \tag{W19.3}$$

Here

$$\kappa = \sqrt{k_\parallel^2 - \frac{2mE}{\hbar^2}}, \tag{W19.4a}$$

where $E < 0$ and

$$q = \sqrt{k_\parallel^2 - \frac{2m(E+V_0)}{\hbar^2}}. \tag{W19.4b}$$

The function $\phi(z)$ is continuous at $z = 0$. The discontinuity in the derivative is determined by the strength of the delta function:

$$\sqrt{k_\parallel^2 - \frac{2mE}{\hbar^2}} + \sqrt{k_\parallel^2 - \frac{2m(E+V_0)}{\hbar^2}} = \frac{2mU}{\hbar^2}. \tag{W19.5}$$

The solution to this equation gives the dispersion formula for the surface state band, $E(k_\parallel)$. Note that at $k_\parallel = 0$, E must lie below $-V_0$.

[†] 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".

For the Shockley state one may develop a heuristic model to help understand its origin. Consider a semiconductor and look at the states near the top of the valence band at energy E_v. For simplicity's sake the effective mass of the holes will be assumed to be isotropic and the band will be taken to be parabolic. The energy of an *electron* in the valence band is then given by

$$E(k) = E_v - \frac{(\hbar k)^2}{2m_h^*}. \tag{W19.6}$$

One may develop a phenomenological Schrödinger equation based on a spatially dependent mass $m(z)$ with $m(z)$ being the free-electron mass in vacuum and the negative of the hole mass inside, that is,

$$m(z) = \begin{cases} -m_h^* & \text{if} \quad z < 0 \\ +m & \text{if} \quad z > 0. \end{cases} \tag{W19.7}$$

The resulting Schrödinger equation is

$$-\frac{\hbar^2}{2}\nabla \cdot \left[\frac{1}{m(z)}\nabla\phi\right] + E_v\Theta(-z)\phi = E\phi. \tag{W19.8}$$

(The gradient operator is written in this split form so that the probability current perpendicular to the surface may be proven to be continuous.)

As before, look for a solution of the form given by Eqs. (W19.2) and (W19.3). Now

$$q = \sqrt{\frac{2m_h^*}{\hbar^2}(E - E_v) + k_\parallel^2}, \tag{W19.9a}$$

$$\kappa = \sqrt{k_\parallel^2 - \frac{2mE}{\hbar^2}}. \tag{W19.9b}$$

The wavefunction $\phi(z)$ in Eq. (W19.3) is already continuous. The continuity of probability current perpendicular to the surface,

$$-\frac{\hbar}{m_h^*}\mathrm{Im}\left(\phi^*\frac{d\phi}{dz}\right) = \frac{\hbar}{m}\mathrm{Im}\left(\phi^*\frac{d\phi}{dz}\right), \tag{W19.10}$$

which is needed for a valid wavefunction, implies that

$$\frac{q}{m_h^*} = \frac{\kappa}{m}. \tag{W19.11}$$

Thus the condition for the surface-state band is

$$\frac{1}{m_h^*}\sqrt{\frac{2m_h^*}{\hbar^2}(E - E_v) + k_\parallel^2} = \frac{1}{m}\sqrt{k_\parallel^2 - \frac{2mE}{\hbar^2}}. \tag{W19.12}$$

For $k_{\|} = 0$ the surface state lies at an energy above the top of the valence band ($E > -|E_v|$) but below the vacuum level ($E < 0$):

$$E(k_{\|} = 0) = -\frac{|E_v|}{1 + m_h^*/m}. \tag{W19.13}$$

More generally, one often employs a complex band structure in which the bulk energy bands are extended to negative values of k^2. This permits an effective Hamiltonian for the solid to be written which may be solved in conjunction with the Hamiltonian for the electron in vacuum. The procedure of wavefunction matching is similar to what was employed, but the implementation is more computational.

W19.2 Surfactants

Surface-active agents, or *surfactants*, are molecules that can radically alter the surface or interface properties of a system even in small concentrations. The system usually involves the liquid–solid, liquid–liquid, or liquid–gas interface. Sometimes the term *surfactant* is used in reference to adsorbates [e.g., a monolayer of As is used on Si (100) and Ge (100) to aid in Si–Ge heteroepitaxy]. Here, however, the focus is on the liquid–solid interface. The surfactant molecule can consist of a long hydrocarbon chain with an polar unit at one end. In the liquid the hydrocarbon chain must push aside the liquid molecules to make room for the surfactant molecule. This involves reducing the forces responsible for the liquid bonds. In water the surfactant molecule must break apart the hydrogen bonds that exist. Since the hydrocarbon chain has all its valence requirements satisfied by carbon–carbon or carbon–hydrogen bonds, it is fairly inert to chemical or electrical interactions with the liquid. The net result is that the liquid tends to expel the hydrocarbon in order to lower its energy. The hydrocarbon chain is called *hydrophobic*, since it avoids being in water. On the other hand, the polar end can lower its energy by immersing itself in the liquid. There is an electrical attraction between the polar group and the liquid. This end is called *hydrophilic*, due to its affinity for water. In order for the molecule to go into solution, the energy decrease involved in the hydrophilic interaction must be greater than the energy increase due to the hydrophobic interaction. Typical examples of surfactant molecules are $C_{12}H_{25}SO_4{}^-Na^+$ and $C_{12}H_{23}COO^-Na^+$.

The surface or interface provides a region of space where both the hydrophobic and hydrophilic tendencies can be satisfied simultaneously. If the polar group lies in the liquid and the hydrocarbon chemisorbs onto the surface, a doubly low energy can be achieved. The lowest-energy state of the system therefore involves an accumulation of the surfactant molecules at the surface. This means that even in small concentrations the molecules will aggregate at the surface.

The adsorption of the surfactants at the surface or interface lowers the interfacial tension, often significantly. This can radically alter such properties as surface diffusion, chemisorption, and crystal growth. Since the surface atoms are now binding themselves to the surfactant molecules, they have fewer bonding electrons to form the surface bonds, thereby depressing the surface tension.

The surface tension drops monotonically with increasing surfactant concentration until a critical concentration is reached (usually when the surface is completely covered). Beyond that the surface properties no longer change. This curious behavior is traced to an interaction that the surfactant molecules have among themselves. The

surfactant molecules can form a composite unit in solution called a *micelle*. The micelle comes about, for example, by creating a ball of molecules with their hydrocarbon chains directed toward the center of the sphere and the polar groups directed outward into the liquid. Liquid is not present in the interior of the micelle. This also satisfies both the hydrophobic and hydrophilic tendencies of the molecule. Other geometries, involving micellar rods or parallel sheets, are also possible.

To understand why a surfactant molecule would prefer to leave the liquid and adsorb onto a surface, one must compare the energies of the molecule in solution with it being adsorbed on the surface. A crude model for the interaction of the surfactant molecule with the liquid may be obtained by imagining that the polar end is a point dipole that carves out a small spherical cavity around it in the liquid. Let the sphere have a radius equal to a. Denote the strength of the dipole by μ, and the electric permittivity of the liquid by ϵ. The electrostatic potential in all of space is then given by

$$\Phi(r,\theta) = \begin{cases} -E_0 r\cos\theta + \dfrac{\mu\cos\theta}{4\pi\epsilon_0 r^2} & \text{if} \quad r < a, \\ \dfrac{p\cos\theta}{4\pi\epsilon r^2} & \text{if} \quad r > a, \end{cases} \tag{W19.14}$$

where, in order to satisfy the continuity of Φ and the radial component of the electric displacement vector D_r

$$p = \frac{3\mu\epsilon}{\epsilon_0 + 2\epsilon}, \tag{W19.15}$$

$$E_0 = \frac{2\mu}{4\pi\epsilon_0 a^3}\frac{\epsilon - \epsilon_0}{\epsilon_0 + 2\epsilon}. \tag{W19.16}$$

Here E_0 is the electric field in the cavity due to the polarization charges in the liquid. The interaction energy of the dipole with this field, U_s, is called the *solvation energy*:

$$U_s = -\frac{\mu^2}{4\pi\epsilon_0 a^3}\frac{\epsilon - \epsilon_0}{\epsilon_0 + 2\epsilon}. \tag{W19.17}$$

The hydrophobic interaction, U_i, may be estimated by imagining that the hydrocarbon chain carves out a cylindrical cavity with surface area A. This causes a rise in the surface energy given approximately by the product of the surface tension of the liquid and the area

$$U_i = \gamma A. \tag{W19.18}$$

For the molecule to go into solution, the total energy, $U_s + U_i$, must be negative. When chemisorption of the surfactant molecule occurs, there is an additional energy U_c, corresponding to the chemisorption bond. Since $U_c < 0$ it is favorable for the surfactant molecules to go out of solution and adsorb onto the surface.

W19.3 Adsorption

Suppose that a solid is exposed to a monatomic gas at temperature T and pressure P. Atoms will strike the surface and a fraction, s, will stick to it. It is therefore important

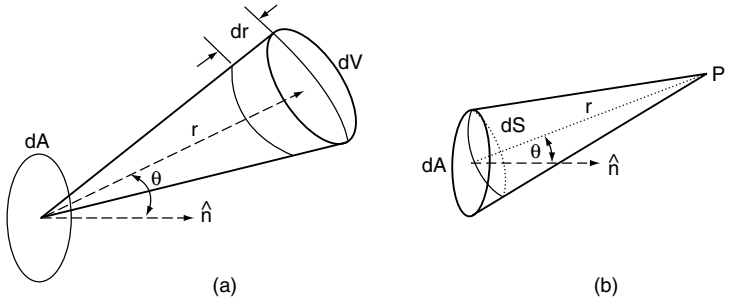

Figure W19.1. An element of area on the surface, dA, and volume element in the gas, dV; particles emanating from a volume element at P strike the element of area dA on the surface.

to determine the impingement flux, F, defined as the number of atoms striking the surface per unit area per unit time. As will be seen, F is determined simply in terms of P, T, and the atomic mass, M.

In Fig. W19.1*a* an element of area dA of the surface is drawn, as well as a volume element dV in the gas a distance r away. The vector joining dA and dV makes an angle θ with the surface normal. The radial extent of dV is dr. The number of atoms in dV is $dN = n\,dV$, where n is the number of atoms per unit volume (number density). For the moment, consider only the subset of atoms moving with a given speed v. These atoms are moving in random directions. Those atoms that are directed approximately at dA will strike it at a time $t = r/v$ later, over a duration lasting $dt = dr/v$. Therefore, the volume element may be expressed as $dV = r^2\,d\Omega v\,dt$, where $d\Omega$ is the solid angle subtended by dV at dA.

The fraction of atoms emanating from dV which strike dA is determined by the solid angle subtended by dA by a typical point in dV, P. Referring to Fig. W19.1*b*, the solid angle is $d\Omega' = dS/r^2$, where dS is the projection of dA onto a plane perpendicular to r, and is given by $dS = dA\cos\theta$. The desired fraction is $df = dA\cos\theta/4\pi r^2$, where the solid angle has been divided by 4π steradians.

The differential flux is

$$dF = \frac{df}{dA}\frac{dN}{dt} = \frac{nv}{4\pi}\cos\theta\,d\Omega. \tag{W19.19}$$

The net flux is obtained by integrating dF over a hemisphere (using $d\Omega' = 2\pi\sin\theta\,d\theta$, where $0 \le \theta \le \pi/2$), that is,

$$F = \frac{n\langle v\rangle}{4}. \tag{W19.20}$$

Here there is finally an average over all speeds.

The kinetic theory of gases provides a means for computing $\langle v\rangle$:

$$\langle v\rangle = \frac{\int d^3v\,v\exp[-\beta(mv^2/2)]}{\int d^3v\exp[-\beta(mv^2/2)]} = \sqrt{\frac{8}{\pi\beta m}}; \tag{W19.21}$$

here $\beta = 1/k_B T$. Finally, employing the ideal gas law, $P = nk_B T$, the desired expression for the impingement flux is obtained:

$$F = \frac{P}{\sqrt{2\pi M k_B T}}. \tag{W19.22}$$

The rate of deposition of adsorbed atoms per unit area, dN_a/dt, is determined by multiplying the impingement flux by the sticking probability, s. The quantity s is the fraction that stick "forever" (or for at least several vibrational periods). Thus

$$\frac{dN_a}{dt} = \frac{sP}{\sqrt{2\pi M k_B T}}. \tag{W19.23}$$

The sticking probability or coefficient can be a complicated function of the surface conditions and the adsorbed atom areal number density, N_a. Often, this areal density is expressed as the coverage, θ, which is the fraction of a monolayer that is adsorbed (i.e., $\theta = N_a/N_{am}$). For example, at low temperatures, s for N_2 on W(110) first rises and then falls as θ increases. For N_2 on W(100), however, s decreases monotonically with increasing coverage. Different faces of the same crystal can have different values of s. For example, for W(100) $s = 0.6$ at $\theta = 0$, whereas $s = 0.4$ for W(411) and $s = 0.08$ for W(111). The existence of steps on the surface often increases the value of s over what it would be for a smooth surface. For example, s for N_2 adsorbing on Pt (110) increases from 0.3 for a smooth surface to 1.0 for a step density of 8×10^8 m^{-1}. This trend is to be expected since steps generally possess dangling bonds which enhance the degree of chemical reactivity.

The impingement flux is rather high at normal atmospheric pressure. For example, for air at room temperature the flux is 3×10^{27} atoms/m^2·s. Taking $s \approx 1$, one sees that a monolayer ($N_a \approx 10^{19}$ m^{-2}) will be deposited on the surface in about 10^{-8} s. To study a clean surface, ultrahigh-vacuum conditions must be maintained, with pressures as low as 10^{-12} torr, 760 torr being 1 atmosphere of pressure. This often requires preparing the sample under ultrahigh-vacuum conditions, as well. The unit of exposure of a surface to a gas is called the *langmuir*; 1 langmuir corresponds to an exposure of 10^{-6} torr·s.

Once the atom strikes the surface and sticks, at least temporarily, it will migrate from place to place by a series of thermally activated jumps. Most of the time, however, will be spent at adsorption sites. These sites correspond to the minima of the potential energy surface. Typical places for these sites are illustrated in Fig. W19.2, which shows the on-top site, T; the bridge site, B; and the centered site, C, for two crystal faces. More complicated sites can exist for other crystal faces. Steps, kinks, and defect sites are also common adsorption sites.

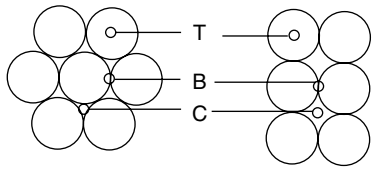

Figure W19.2. The top site, T, the bridge site, B, and the centered site, C for two crystal faces. The left face could be FCC (111) or HCP (0001). The right face could be FCC (100) or BCC (100).

W19.4 Desorption

Desorption is the inverse process to adsorption. Atoms bound in the potential well of the surface vibrate at a characteristic vibrational frequency determined by the atomic mass and the curvature at the bottom of the well. In addition, the atoms interact with the bath of thermal phonons presented by the solid. This causes the energy of the adsorbed atom to fluctuate in time. When the energy fluctuates by an amount sufficient to overcome the binding energy, the atom can dissociate from the surface and be desorbed. The vaporization process is described in terms of desorption in Section 6.3 of the textbook.

A reasonable estimate for the rate of atoms per unit area that desorb may be obtained from the expression

$$\frac{dN_d}{dt} = N_a f \exp\left(-\frac{E_c}{k_B T_s}\right). \tag{W19.24}$$

Here N_a is the number of atoms adsorbed per unit area, f the vibrational frequency of the atoms, and T_s the surface temperature. The probability of the atom achieving the required energy E_c is given by the Boltzmann factor. The factor f represents the "attempt" frequency. In using this expression the situation depicted in Fig. 19.15*a* applies. For the case of a second physisorption well, as in Fig. 19.15*b*, E_p should be used in place of E_c and the density of physisorbed atoms, N_p, should be used rather than the density of chemisorbed atoms, N_a.

In thermal equilibrium the surface and gas temperatures are equal, $T_s = T$, and the adsorption rate equals the desorption rate. Under these conditions it can be shown that

$$N_a(T) = \frac{sP}{f\sqrt{2\pi M k_B T}} \exp\left(\frac{E_c}{k_B T}\right). \tag{W19.25}$$

Thus the number density of adsorbed atoms is proportional to the pressure of adsorbate atoms in the gas.

Now proceed to look at the Langmuir model for adsorption. In this model one regards the surface as having a density of adsorption sites, N_s (denoted by N_{am} in Section W19.3). The sticking probability is modified as these sites are filled with adsorbate atoms. When all the sites are filled, the adsorption process comes to a halt. This model is not general. It applies to a restricted set of adsorption processes, usually corresponding to a strong chemisorption bond formed between the solid and the adsorbate.

Let θ denote the fraction of sites that are occupied, that is,

$$\theta = \frac{N_a}{N_s}. \tag{W19.26}$$

In place of the previous sticking probability, s, one now has $s(1-\theta)$. Thus, equating the adsorption rate to the desorption rate yields

$$\frac{sP(1-\theta)}{\sqrt{2\pi M k_B T}} = N_s \theta f \exp\left(-\frac{E_c}{k_B T}\right). \tag{W19.27}$$

Solving for θ gives the *Langmuir adsorption isotherm*,

$$\theta(P, T) = \frac{aP}{1 + aP}, \tag{W19.28}$$

where

$$a(T) = \frac{s}{N_s f \sqrt{2\pi M k_B T}} \exp\left(\frac{E_c}{k_B T}\right). \tag{W19.29}$$

The formulas above show that the surface coverage saturates to $\theta = 1$ at high gas pressures.

More sophisticated models have been constructed to describe the situation where multilayer adsorption and desorption can occur.

W19.5 Surface Diffusion

The normal state of affairs for adsorbed atoms is for them to move around on the surface at finite temperatures. This is in contrast to the bulk solid, where diffusion occurs via vacancies or interstitials present under equilibrium conditions. Surface diffusion proceeds by a series of thermally activated jumps. In general, no atoms of the substrate have to be "pushed" out of the way to achieve this jump. In this sense it is different from the bulk solid.

Consider a surface that has rectangular symmetry. The diffusion equation for the motion of the adsorbed atoms will be derived. Let the probability for finding an atom in the surface net cell (x, y) at time t be denoted by $F(x, y, t)$. The probability is just the concentration of adsorbed atoms divided by the concentration of available sites, $F = N_a/N_s$. Let p_x be the probability that the atom makes a jump of size d_x in the positive x-direction in a time τ. For the y direction the analogous jump probability involves d_y. Attention will be restricted to the case where there is surface reflection symmetry, so p_x is also the probability for a jump to the point $x - d_x$. At time $t + \tau$ the probability becomes

$$\begin{aligned} F(x, y, t + \tau) = (1 - 2p_x - 2p_y)F(x, y, t) + p_x[F(x + d_x, y, t) + F(x - d_x, y, t)] \\ + p_y[F(x, y + d_y, t) + F(x, y - d_y, t)]. \end{aligned} \tag{W19.30}$$

The first term on the right-hand side represents the probability for the atom originally at (x, y) to have remained on the site. The second and third terms together give the probability that neighboring atoms hop onto the site. Expanding both sides in powers of τ, d_x, and d_y, and retaining lowest-order nonvanishing terms, leads to the diffusion equation

$$\frac{\partial F}{\partial t} = D_x \frac{\partial^2 F}{\partial x^2} + D_y \frac{\partial^2 F}{\partial y^2}, \tag{W19.31}$$

where the diffusion coefficients are

$$D_x = \frac{p_x d_x^2}{\tau}, \qquad D_x = \frac{p_y d_y^2}{\tau}. \tag{W19.32}$$

In the case where there is square symmetry, the two diffusion coefficients become equal to each other and may be replaced by a common symbol, D.

Instead of talking about probabilities, it is more useful to talk about surface concentration, which will now be denoted by C (i.e., $C = N_a = N_s F$). Equation (W19.31) is obeyed by C, since one need only multiply through by N_s. In the derivation above it was assumed that the hopping probabilities are independent of whether or not the site to which it hops is occupied. This is clearly a limitation. It may be remedied by allowing the diffusion constants themselves to be functions of the particle concentration. One may introduce a particle current per unit length, $\mathbf{J}$, defined as the number of adsorbed atoms hopping across a line of unit length per unit time. Suppose, for example, that the surface is horizontal and a line is drawn from south to north. If there is a higher concentration to the east of the line than to the west, there will be a larger number of atoms jumping to the west than to the east. Thus the current will be proportional to the gradient of the probability. Using arguments similar to those used before leads to

$$\mathbf{J} = -\mathbf{D} \cdot \nabla C. \tag{W19.33}$$

Here a diffusion matrix, $\mathbf{D}$, has been introduced and the possibility of having off-diagonal terms must be allowed for.

The continuity equation that governs the flow of particles on the surface is

$$\nabla \cdot \mathbf{J} + \frac{\partial C}{\partial t} = \left(\frac{dC}{dt}\right)_{\text{adsorb}} - \left(\frac{dC}{dt}\right)_{\text{desorb}}. \tag{W19.34}$$

The terms on the right-hand side correspond to the increase or decrease in concentration due to adsorption and desorption, respectively. One thereby obtains the generalized diffusion equation:

$$-\nabla \cdot (\mathbf{D} \cdot \nabla C) + \frac{\partial C}{\partial t} = \left(\frac{dC}{dt}\right)_{\text{adsorb}} - \left(\frac{dC}{dt}\right)_{\text{desorb}}. \tag{W19.35}$$

For pure surface diffusion, the right-hand side of this equation would be zero.

In the diffusion process the probability for making a hop depends on the surface temperature, T_s, and the surface barrier height, E_b;

$$p_x(T_s) = \tau f \exp\left(-\frac{E_b}{k_B T_s}\right). \tag{W19.36}$$

Here f is the attempt frequency, which is essentially the vibrational frequency of the adatom parallel to the surface. In this formula, both the attempt frequency and the barrier height may be different for the x and y directions. For simplicity's sake, attention will henceforth be restricted to the case of square symmetry. Since the hopping probabilities exhibit Arrhenius-type behavior, the diffusion coefficient will also exhibit such behavior. The higher the temperature, the greater will be the rate of surface diffusion.

The solution to the homogeneous diffusion equation, ignoring adsorption and desorption, in two dimensions subject to the initial condition is $C(\mathbf{r}, t = 0) = C_0\delta(\mathbf{r})$ is

$$C(r, t) = \frac{C_0}{4\pi Dt} \exp\left(-\frac{r^2}{4Dt}\right). \tag{W19.37}$$

This may be verified for $t > 0$ by insertion of this formula into the diffusion equation. [Note that $C(r, t)$ and C_0 do not have the same dimensions.] As $t \to 0$ the spatial extent of C becomes narrower and the size of C increases without bound, but the integral over area remains fixed at the value C_0, consistent with the initial condition. This concentration function may be used to compute the mean-square displacement, that is,

$$\langle r^2 \rangle = \frac{\int C(r, t) r^2 dA}{C_0} = 4Dt. \tag{W19.38}$$

The mean-square displacement that a particle travels from its starting point grows as the square root of time for diffusive motion. This is to be contrasted with the case of ballistic motion, where the distance covered grows linearly with t. The presence of surface defects may play an important role in surface diffusion because they often offer paths of high mobility for the diffusing atoms. They may also trap diffusing atoms (e.g., dislocations can pull surface atoms into the bulk or ledges may trap atoms).

One way of observing surface diffusion is by means of the field-ion microscope. Using the atomic-scale resolution capabilities of the microscope permits one to follow the path of a single atom. Usually, the temperature of the tip of the microscope is raised, and the temperature is maintained for some time and then cooled. At elevated temperatures the atom has a chance to hop to an adjacent site. In this way the random walk associated with diffusive motion may be studied. The diffusion coefficient may be extracted from Eq. (19.38) and studied as a function of temperature. From the Arrhenius behavior of D the barrier height E_b may be determined.

W19.6 Catalysis

Surfaces of solids may be used to promote or accelerate particular chemical reactions selectively. Such a catalytic process generally involves the following steps: adsorption of molecules onto the surface; dissociation of the molecules into smaller components (including possibly atoms); diffusion of the components on the surface; reaction of the components to form product molecules; and finally, desorption of the product from the solid. Each of these steps generally involves potential barriers that need to be surmounted, so there are a number of physical parameters governing the overall reaction rate.

Consider, for example, the Haber process for the synthesis of ammonia. Historically, this process has proven to be extremely important because of the role of ammonia as a primary starting material in the manufacture of fertilizers and explosives. The process is illustrated in Fig. W19.3.

The catalyst used is iron. When nitrogen molecules adsorb on iron, the dissociation energy for N_2 is lowered. This is because some of the orbitals that were previously involved in the N–N bond now hybridize with the Fe $3d$ orbitals and serve as the basis for establishing the N_2–Fe bond. At elevated surface temperatures ($\approx 400^\circ$C) the

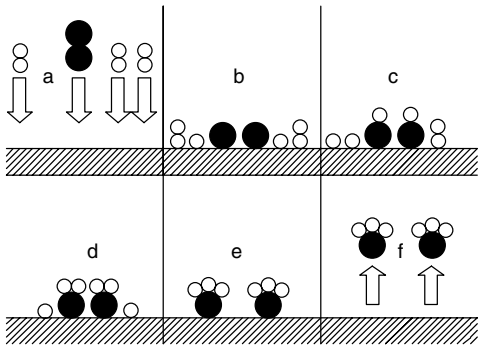

Figure W19.3. Six stages in the Haber process: nitrogen (dark circles) and hydrogen (light circles) combine to form ammonia on iron.

probability for N_2 dissociation increases. The net result is that individual N atoms are bound to the iron and are able to hop from site to site as a result of thermal activation. Hydrogen undergoes a similar dissociation process (i.e., $H_2 \rightarrow H + H$). When a free H and N combine, there is a probability for reacting to form the NH radical, which is still adsorbed. Further hydrogenation results in the formation of NH_2 and ultimately, the saturated NH_3 molecule. Whereas the NH and NH_2 radicals are chemically active, and hence remain chemisorbed to the Fe, the NH_3 is only physisorbed. It is easy for it to desorb. The net result is that Fe has served as the catalyst for the reaction $N_2 + 3H_2 \rightarrow 2NH_3$. Although a number of metals can be used to dissociate N_2 and H_2, Fe is optimal in that it does not attach itself so strongly to N and H so as to prevent their further reacting with each other to reach the desired product, NH_3. What matters is the net turnover rate — how rapidly the overall reaction can be made to proceed per unit area of catalyst.

It is found that some faces of Fe are more catalytically active than others. The Fe (111) and (211) faces are the most active faces, while the (100), (110), and (210) are less active. It is believed that the (111) and (211) faces are special in that they expose an iron ion that is only coordinated to seven other iron atoms (called the C_7 site). It is also found that potassium atoms enhance the sticking coefficient for gas molecules and therefore help promote the catalytic reaction. This is attributed to the lowering of the work function of the surface, which makes it easier for Fe 3*d* orbitals to penetrate into the vacuum so they could form chemical bonds with the adsorbed nitrogen and hydrogen species.

Another example of catalysis is provided by the catalytic convertor used in the automobile industry. Here the problem is to remove carbon monoxide (CO) and nitric oxide (NO) from the exhaust fumes of the internal combustion engine. The catalyst of choice consists of particles of platinum (Pt) and rhodium (Rh) on a (relatively inexpensive) supporting material. An actual catalyst consists of small particles supported on oxide powders. The CO molecule adsorbs on the metal. Some oxygen is present. The O_2 molecules dissociatively adsorb (i.e., $O_2 \rightarrow 2O_{ad}$). Similarly, NO dissociatively adsorbs (i.e., $NO \rightarrow N_{ad} + O_{ad}$). Free N and O atoms diffuse across the surface. When an O atom encounters the CO molecule, the reaction $CO + O \rightarrow CO_2$ is possible. Since the valency requirements of this molecule are fully satisfied, it readily desorbs from the catalyst. The adsorbed N atoms can react similarly to form nitrogen molecules ($N + N \rightarrow N_2$), which also readily desorb.

The morphology of the surface often plays a crucial role in its efficiency as a catalyst. Various crystallographic faces of a given material often have catalytic activities that can vary by orders of magnitude. These large variations reflect the underlying exponential dependence of hopping probability on barrier height. Step sites and other defects often provide locales that favor one or more of the processes needed to transform reactants to products. This is presumably related to the presence of dangling bonds that can be utilized in forming surface-chemical intermediates. Catalysts are frequently used in the form of powders, to maximize the amount of available surface area per unit mass. In some cases coadsorbates are introduced because they provide beneficial surface structures, such as islands, which can play a role similar to that of steps.

W19.7 Friction

The average power generated per unit area by kinetic friction is given by $\mu_k N v/A_a$. This causes an average temperature rise ΔT of the interface. The actual temperature rise will depend on the thermal conductivities κ of the solids and characteristic geometric lengths. One may write the formula as

$$\Delta T = \frac{\mu_k N v}{A_a} \frac{1}{\kappa_1/l_1 + \kappa_2/l_2} = \mu_k P v \frac{1}{\kappa_1/l_1 + \kappa_2/l_2}. \tag{W19.39}$$

where P is the pressure. The lengths l_1 and l_2 correspond to the characteristic distances over which the change ΔT occurs. However, since the actual contact area is much smaller than the apparent contact area, there will be points where the temperature rise is considerably higher. There the temperature rise, to what is called the *flash temperature*, will be given by

$$\Delta T' = \frac{\mu_k N v}{A_t} \frac{1}{\kappa_1/l_1 + \kappa_2/l_2}. \tag{W19.40}$$

This may be a serious problem in ceramics, which generally have low values of κ. The high temperatures produce thermal stresses that lead to brittle fracture. This may be eliminated by depositing a good thermally conducting layer, such as Ag, which serves to dissipate the frictional heat.

A possible explanation for the velocity dependence of μ_k, noted above, is due to the melting of surface asperities. When v becomes sufficiently large, $\Delta T'$ given by Eq. (W19.40) may be large enough to melt the surface asperities.

An interesting case arises if two atomically flat surfaces with different lattice spacings are brought into contact and slide past each other. If the ratio of the lattice spacings is an irrational number, the lattices are said to be *incommensurate*. In that case simulations show that one surface may slowly slide relative to the other without the need to change the number of bonds between them. Furthermore, the energy released by forming a new bond may be resonantly transferred to open a nearby existing bond. There is no static friction predicted in such a case, only viscous friction.

One interesting result of nanotribology is that the kinetic friction force is actually velocity dependent. The force is proportional to the relative velocity at the true contact points. Of course, this velocity may be quite different than the macroscopic velocity due to the local deformations that occur. The kinetic friction force, on a microscopic

level, is actually a viscouslike friction force. The characteristic relaxation time is given by the sliptime.

Lubrication involves attempting to control friction and wear by interposing a third material between the two contacting surfaces. Commonly used solid-state lubricants include the layered materials graphite and MoS_2. Here lubrication is achieved by having weakly bound layers slough off the crystals as shear stress is applied. Liquid lubricants include such organic compounds as paraffins, diethyl phosphonate, chlorinated fatty acids, and diphenyl disulfide. Spherical molecules, such as fullerene, or cylindrical molecules such as carbon nanotubes, behave in much the same way as ball bearings in reducing friction. Lubricants can also carry heat away from flash points or can serve to equalize stress on asperities.

Molecular-dynamics (MD) simulations are often used in conjunction with nanotribology experiments to obtain a more complete understanding of the physics of friction. An example involves the jump-to-contact instability, in which atoms from a surface (such as Au) will be attracted toward an approaching tip of a solid (such as Ni) when the separation is less than 1 nm. At a separation of 0.4 nm, the two metals will actually come into contact by means of this instability.

In another example it was recently found that the amount of slip at a liquid–solid interface is a nonlinear function of the shear rate, $\dot{\gamma}$. If Δv is the relative velocity of the fluid and solid at the interface, Navier had postulated that $\Delta v = L_s \dot{\gamma}$, with L_s being a slip length characteristic of the solid and liquid. The MD simulations[†] show that $L_s = L_s^0(1 - \dot{\gamma}/\dot{\gamma}_c)^{-1/2}$.

The interplay between triboelectricity and friction is not yet completely understood, although there is evidence that the sudden stick-slip motion does produce electrification. When two different materials are brought into contact, a charge transfer will occur to equalize the chemical potential for the electrons. The resulting difference in potential is called the *contact potential*. If the materials are slowly separated from each other the charge transfer is reversed and no electrification occurs. However, for sudden separation, as occurs in a slip, there is incomplete reverse charge transfer and the materials become electrified. It is possible that this accounts for the picosecond bursts of light seen at the moving meniscus of the Hg–glass interface[‡].

Appendix W19A: Construction of the Surface Net

Let $\{\mathbf{R}\}$ be a set of lattice vectors and $\{\mathbf{G}\}$ the corresponding set of reciprocal lattice vectors for a Bravais lattice. The lattice vectors are expressed in terms of the primitive lattice vectors $\{\mathbf{u}_i\}$ $(i = 1, 2, 3)$ by

$$\mathbf{R} = n_1\mathbf{u}_1 + n_2\mathbf{u}_2 + n_3\mathbf{u}_3, \tag{W19A.1}$$

where $\{n_1, n_2, n_3\}$ are a set of integers. Similarly, the reciprocal lattice vectors may be expanded in terms of the basis set $\{\mathbf{g}_j\}$ by

$$\mathbf{G} = j_1\mathbf{g}_1 + j_2\mathbf{g}_2 + j_3\mathbf{g}_3, \tag{W19A.2}$$

[†] P. A. Thomson and S. M. Troian, *Nature*, **389**, 360 (1997).
[‡] R. Budakian et al, *Nature*, **391**, 266 (1998).

where $\{j_1, j_2, j_3\}$ are also a set of integers. The primitive and basis vectors obey the relations

$$\mathbf{u}_i \cdot \mathbf{g}_j = 2\pi\delta_{ij}. \tag{W19A.3}$$

Select an atom at point **O** in the interior of the solid as the origin. Let the surface plane be perpendicular to a particular vector **G** and a distance h from **O**. If the displacement vector **r** from **O** to a point on the surface plane is projected along **G**, the magnitude of this projection is constant. Thus the plane is described by the equation

$$\mathbf{r} \cdot \hat{G} = h \tag{W19A.4}$$

where $\hat{G}$ is a unit vector. This is illustrated in Fig. W19A.1.

Inserting a lattice vector for **r** leads to the formula

$$2\pi(j_1 n_1 + j_2 n_2 + j_3 n_3) = hG. \tag{W19A.5}$$

This equation may be used to eliminate one of the numbers n_1, n_2, or n_3. Which can be eliminated depends on the numbers j_1, j_2, and j_3. If j_1 is nonzero, n_1 may be eliminated and

$$\mathbf{R} = \frac{\mathbf{u}_1}{j_1}\left(\frac{h}{2\pi}G - n_2 j_2 - n_3 j_3\right) + n_2\mathbf{u}_2 + n_3\mathbf{u}_3 \tag{W19A.6}$$

If j_1 is zero, either n_2 can be eliminated (assuming that j_2 is nonzero) or n_3 can be eliminated (assuming that j_3 is nonzero), with analogous formulas for **R** following accordingly. In the following it will be assumed that j_1 is nonzero.

The atoms of the ideal surface plane lie on a regular two-dimensional lattice called the *surface net*. To study this net more closely, project the vector **r** onto the surface lattice plane. Referring to Fig. W19A.2 shows that for a general vector **r** the projected vector is

$$\mathbf{r}' = \mathbf{r} - \mathbf{r} \cdot \hat{G}\hat{G} = \hat{G} \times (\mathbf{r} \times \hat{G}). \tag{W19A.7}$$

Thus a set of projected primitive lattice vectors $\{\mathbf{u}_\mathbf{i}'\}$ can be constructed:

$$\mathbf{u}_1' = \hat{G} \times (\mathbf{u}_1 \times \hat{G}), \tag{W19A.8a}$$

$$\mathbf{u}_2' = \hat{G} \times (\mathbf{u}_2 \times \hat{G}), \tag{W19A.8b}$$

$$\mathbf{u}_3' = \hat{G} \times (\mathbf{u}_3 \times \hat{G}). \tag{W19A.8c}$$

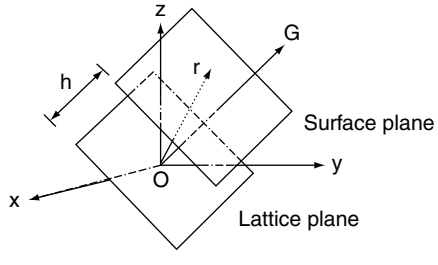

Figure W19A.1. Ideal surface plane defined in terms of the direction of the reciprocal lattice vector, **G**, and h, the distance of an atom at **O**.

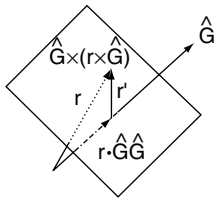

Figure W19A.2. Projecting a vector **r** onto the lattice plane defined by vector **G**.

The projected lattice vector is therefore

$$\mathbf{R}'_{mn} = \frac{h\mathbf{u}'_1}{2\pi j_1} G + n_2\mathbf{v}_2 + n_3\mathbf{v}_3, \tag{W19A.9}$$

where $\mathbf{v}_2$ and $\mathbf{v}_3$ are the primitive surface net vectors, defined by

$$\mathbf{v}_2 = \mathbf{u}'_2 - \frac{j_2}{j_1}\mathbf{u}'_1, \tag{W19A.10a}$$

$$\mathbf{v}_3 = \mathbf{u}'_3 - \frac{j_3}{j_1}\mathbf{u}'_1. \tag{W19A.10b}$$

Note that the projected vector $\mathbf{R}'_{mn}$ is defined by only two subscripts, m and n. The angle between the primitive surface net vectors is determined by the formula

$$\cos\theta = \frac{\mathbf{v}_2 \cdot \mathbf{v}_3}{v_2 v_3}. \tag{W19A.11}$$

(It is convenient to relabel the net vectors so that v_1 and v_2 define the surface net. This is accomplished by making the cyclic permutation $3 \to 2 \to 1 \to 3$.)

In many cases the surface net that results from cutting the lattice by a surface plane is easy to visualize, so one might argue that the mathematical machinery above is superfluous. However, when attempting to automate the procedure, the analytic approach has decided advantages. After all, a computer is not adept at visualization.

Example. Suppose that a simple cubic crystal is sliced by a plane perpendicular to the [111] direction. Take this plane to pass through an atom at the origin. In this case, $j_1, j_2, j_3 = (1, 1, 1)$ and $h = 0$. Thus

$$\hat{G} = \frac{\hat{i} + \hat{j} + \hat{k}}{\sqrt{3}}. \tag{W19A.12}$$

The projected primitive lattice vectors are

$$\mathbf{u}'_1 = \frac{a}{3}(2\hat{i} - \hat{j} - \hat{k}), \tag{W19A.13a}$$

$$\mathbf{u}'_2 = \frac{a}{3}(-\hat{i} + 2\hat{j} - \hat{k}), \tag{W19A.13b}$$

$$\mathbf{u}'_3 = \frac{a}{3}(-\hat{i} - \hat{j} + 2\hat{k}). \tag{W19A.13c}$$

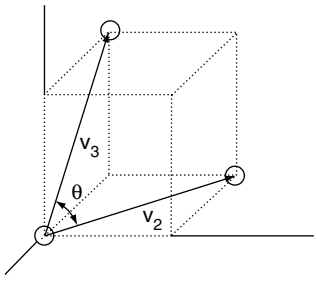

Figure W19A.3. Simple cubic lattice being sliced by a (111) plane passing through the origin.

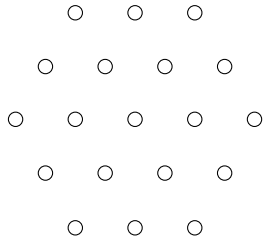

Figure W19A.4. The (111) surface of a simple cubic crystal.

The surface net vectors are

$$\mathbf{v}_2 = a(-\hat{i} + \hat{j}), \tag{W19A.14a}$$

$$\mathbf{v}_3 = a(-\hat{i} + \hat{k}). \tag{W19A.14b}$$

The surface-projected lattice vector is

$$\mathbf{R}'_{mn} = ma(-\hat{i} + \hat{j}) + na(-\hat{i} + \hat{k}). \tag{W19A.15}$$

Figure W19A.3 shows three of the atoms that lie in the surface plane. Figure W19A.4 depicts the layout of the corresponding surface net. It must be emphasized that these two-dimensional nets are the analogs of the Bravais lattices in three dimensions. Just as the lattice in three dimensions may be endowed with a basis of atoms, the same is true in two dimensions.

Applying the formalism above allows one to obtain a precise picture of the surface that results by taking an arbitrary slice through any crystalline structure.

Appendix W19B: Fowler–Nordheim Formula

In this appendix the Fowler–Nordheim formula for the current density produced in field emission is derived. An electric field E_0 is applied normal to a flat metal surface. The potential energy experienced by the electrons is given by

$$V(z) = \begin{cases} 0 & \text{if} \quad z < 0, \\ V_0 - Fz & \text{if} \quad z > 0, \end{cases} \tag{W19B.1}$$

where $F = eE_0$, as illustrated in Fig. 19.11. The Schrödinger equation governing the tunneling process is

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + V(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r}). \tag{W19B.2}$$

The transverse motion is decoupled by writing $\psi(\mathbf{r}) = \phi(z)\exp(i\mathbf{k}_\parallel \cdot \mathbf{R})$. In the region $z < 0$ the Schrödinger equation becomes

$$\left(\frac{\partial^2}{\partial z^2} + k_z^2\right)\phi(z) = 0, \tag{W19B.3}$$

where

$$k_z = \sqrt{\frac{2mE}{\hbar^2} - k_\parallel^2}. \tag{W19B.4}$$

The solution of Eq. (W19B.3) is given by

$$\phi(z) = e^{ik_z z} + re^{-ik_z z}, \tag{W19B.5}$$

with r being interpreted as a reflection amplitude.

For $z > 0$ the Schrödinger equation is

$$-\frac{\hbar^2}{2m}\frac{d^2\phi}{dz^2} + (V_0 - Fz)\phi = \frac{\hbar^2 k_z^2}{2m}\phi. \tag{W19B.6}$$

With the substitution

$$u = \left(\frac{2m}{\hbar^2 F^2}\right)^{1/3}\left(V_0 - Fz - \frac{\hbar^2 k_z^2}{2m}\right), \tag{W19B.7}$$

the Schrödinger equation becomes Airy's differential equation:

$$\frac{d^2\phi}{du^2} - u\phi = 0. \tag{W19B.8}$$

The solution may be expressed as a linear combination of the two Airy functions. The coefficients are chosen so that for large x, ϕ represents a wave traveling to the right. Asymptotic expansions of the Airy functions are presented in Table W19B.1. Thus

$$\phi(u) = N[Bi(u) + iAi(u)], \tag{W19B.9}$$

where N is a normalization constant. The current density carried by this wave is given by

$$J_z = \frac{e\hbar}{m}\mathrm{Im}\left(\phi^*\frac{d\phi}{dx}\right) = \frac{e\hbar|N|^2}{m\pi}\left(\frac{2m}{\hbar^2 F^2}\right)^{1/3}F. \tag{W19B.10}$$

TABLE W19B.1 Asymptotic Expansion of the Airy Functions[a]

$Ai(u) \to \frac{1}{2\sqrt{\pi}u^{1/4}}e^{-\zeta}$,	$Ai'(u) \to -\frac{1}{2\sqrt{\pi}}u^{1/4}e^{-\zeta}$
$Bi(u) \to \frac{1}{\sqrt{\pi}u^{1/4}}e^{\zeta}$,	$Bi'(u) \to \frac{1}{\sqrt{\pi}}u^{1/4}e^{\zeta}$
$Ai(-u) \to \frac{1}{\sqrt{\pi}u^{1/4}}\sin\left(\zeta+\frac{\pi}{4}\right)$,	$Ai'(-u) \to -\frac{1}{\sqrt{\pi}}u^{1/4}\cos\left(\zeta+\frac{\pi}{4}\right)$
$Bi(-u) \to \frac{1}{\sqrt{\pi}u^{1/4}}\cos\left(\zeta+\frac{\pi}{4}\right)$,	$Bi'(-u) \to \frac{1}{\sqrt{\pi}}u^{1/4}\sin\left(\zeta+\frac{\pi}{4}\right)$.

Source: Data from M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C., 1964.

[a] $\zeta = \frac{2}{3}u^{3/2}$.

The wavefunction given by Eq. (W19B.9) and its first derivative at $z = 0$ are set equal to the corresponding quantities given by Eq. (W19B.5). Solving these equations for N yields

$$N = \frac{2ik_z\sqrt{\pi}\ e^{-\zeta_0}L^{-3/2}}{ik_z/u_0^{1/4} - Fu_0^{1/4}(2m/\hbar^2F^2)^{1/3}}, \tag{W19B.11}$$

where $u_0 = (2m/\hbar^2F^2)^{1/3}(V_0 - \hbar^2k_z^2/2m)$, $\zeta_0 = \frac{2}{3}u_0^{3/2}$, and L^3 is the volume of the metal.

The current density is obtained by integrating Eq. (19B.10) over the Fermi sphere:

$$J = \sum_s \sum_{\mathbf{k}} J_z\Theta(E_F - E) = 2\int \frac{d^3kL^3}{(2\pi)^3}J_z\Theta(E_F - E). \tag{W19B.12}$$

The integration over transverse coordinates leads to

$$\int d^2k_\parallel\Theta(E_F - E) = \pi\left(\frac{2mE_F}{\hbar^2} - k_z^2\right)\Theta\left(\frac{2mE_F}{\hbar^2} - k_z^2\right). \tag{W19B.13}$$

Thus one obtains

$$J = \frac{2me}{\pi^2\hbar^3V_0}\int_0^{E_F} dE'(E_F - E')\sqrt{E'(V_0 - E')}\exp\left[-\frac{4\sqrt{2m}}{3F\hbar}(V_0 - E')^{3/2}\right]. \tag{W19B.14}$$

The major contribution to the integral comes from the region $E' = E_F$. Thus one may make the replacements $(V_0 - E')^{3/2} \approx W^{3/2} + \frac{3}{2}\sqrt{W}(E_F - E')$, $E'(V_0 - E') \approx E_FW$ and extend the lower limit of the integral to $-\infty$. Here W is the work function. One finally obtains the Fowler–Nordheim formula:

$$J = \frac{e^3E_0^2}{4\pi^2\hbar V_0}\sqrt{\frac{E_F}{W}}\exp\left(-\frac{4}{3eE_0\hbar}\sqrt{2mW^3}\right). \tag{W19B.15}$$

An additional correction may be included to account for the image potential that the charge experiences when it is in the vacuum region, but it will not be included here.

Appendix W19C: Photoemission Yields

In this appendix theoretical expressions for the photoelectric yield will be derived for an idealized solid whose surface consists of a potential step. The Sommerfeld model will be used to describe the electrons.

First, the simplifying assumption that the potential is only a function of the normal coordinate, z, will be made. The wavefunctions are then of the form

$$\psi_f(\mathbf{r}) = \phi_f(z)\exp(i\mathbf{k}'_\parallel \cdot \mathbf{r}_\parallel), \tag{W19C.1a}$$

$$\psi_i(\mathbf{r}) = \phi_i(z)\exp(i\mathbf{k}_\parallel \cdot \mathbf{r}_\parallel), \tag{W19C.1b}$$

where the subscripts f and i refer to the final and initial states, respectively, and $\mathbf{k}_\parallel$ and $\mathbf{k}'_\parallel$ refer to propagation vectors along the surface.

Write the matrix element in Eq. (19.29) as

$$\langle\psi_f|\boldsymbol{\mu}\cdot\mathbf{E}|\psi_i\rangle = -e\langle\psi_f|\mathbf{r}_\parallel\cdot\mathbf{E}_\parallel|\psi_i\rangle - e\langle\psi_f|zE_z|\psi_i\rangle. \tag{W19C.2}$$

By introducing the Hamiltonian, H, the first term can be shown to vanish:

$$\begin{aligned}\langle\psi_f|\mathbf{r}_\parallel\cdot\mathbf{E}_\parallel|\psi_i\rangle &= \frac{1}{E_f - E_i}\langle\psi_f|[H, \mathbf{r}_\parallel\cdot\mathbf{E}_\parallel]|\psi_i\rangle \\ &= -\frac{i}{m\omega}\langle\psi_f|\mathbf{p}_\parallel\cdot\mathbf{E}_\parallel|\psi_i\rangle = -\frac{i\hbar}{m\omega}\mathbf{k}_\parallel\cdot\mathbf{E}_\parallel\langle\psi_f|\psi_i\rangle = 0.\end{aligned} \tag{W19C.3}$$

In this model it is only the normal component of the electric field that is capable of exciting the electron gas and of causing photoemission. Any photoemission observed at normal incidence, in which case the electric field would be tangent to the surface, would be considered volume photoemission and beyond the scope of the model.

The full Hamiltonian governing the interaction of the electron with the light is

$$H = H_0 + H_\gamma = \frac{p^2}{2m} + V(z) + eE_z z[\exp(\lambda z)\Theta(-z) + \Theta(z)] + e\mathbf{E}_\parallel\cdot\mathbf{r}_\parallel. \tag{W19C.4}$$

The last term is the interaction of the electron with the component of the field parallel to the surface, and can be dropped. The third term is the perturbation, H_γ. For the initial state the unperturbed Schrödinger equation becomes

$$\left[\frac{p_z^2}{2m} + V(z) - \varepsilon_i\right]\phi_i(z) = 0, \tag{W19C.5a}$$

and for the final state,

$$\left[\frac{p_z^2}{2m} + V(z) - \varepsilon_f\right]\phi_f(z) = 0, \tag{W19C.5b}$$

where

$$\varepsilon_i = E_i - \frac{\hbar^2 k_\parallel^2}{2m}, \tag{W19C.6a}$$

$$\varepsilon_f = E_f - \frac{\hbar^2 k_\parallel^2}{2m}. \tag{W19C.6b}$$

The Schrödinger equation will be solved for the simple step potential:

$$V(z) = \begin{cases} 0 & \text{if} \quad z > 0 \\ -V_0 & \text{if} \quad z < 0. \end{cases} \tag{W19C.7}$$

(The effect of a finite electron mean free path could, in principle, be included by making V_0 complex.)

For the initial state the solution was found in Eq. (19.8) in the discussion of relaxation of metals. Thus

$$\phi_i(z) = \begin{cases} B\exp(-\kappa z) & \text{if} \quad z > 0 \\ \dfrac{B\sin(qz+\delta)}{\sin\delta} & \text{if} \quad z < 0 \end{cases} \tag{W19C.8}$$

where

$$\kappa = \frac{1}{\hbar}\sqrt{-2m\varepsilon_i}, \tag{W19C.9a}$$

$$q = \frac{1}{\hbar}\sqrt{2m(V_0+\varepsilon_i)}. \tag{W19C.9b}$$

For the final state one has an *out-state*, an outgoing wave with unit amplitude in the vacuum supplemented with incoming waves in both the vacuum and the metal. (A packet constructed out of such states will evolve into a purely outgoing packet for long times.) Thus

$$\phi_f = \begin{cases} \exp(ikz) + r\exp(-ikz) & \text{if} \quad z > 0, \\ t\exp(iq'z) & \text{if} \quad z < 0, \end{cases} \tag{W19C.10}$$

where

$$k = \frac{1}{\hbar}\sqrt{2m\varepsilon_f}, \tag{W19C.11a}$$

$$q' = \frac{1}{\hbar}\sqrt{2m(\varepsilon_f + V_0)} \tag{W19C.11b}$$

Matching the wavefunction and the derivative at $z = 0$ yields

$$t = 1 + r, \tag{W19C.12a}$$

$$q't = k(1 - r). \tag{W19C.12b}$$

Then the reflection amplitude is

$$r = \frac{k - q'}{k + q'}, \tag{W19C.13a}$$

and the transmission amplitude is

$$t = \frac{2k}{k + q'}. \tag{W19C.13b}$$

The matrix element of the perturbation is

$$\langle \psi_f | H_\gamma | \psi_i \rangle = \int d^2 r_\parallel \exp\lceil i(\mathbf{k}_\parallel - \mathbf{k}'_\parallel) \cdot \mathbf{r}_\parallel \rceil \times eE_z \int_{-\infty}^{\infty} dz\, \phi_f^*(z) z [\exp(\lambda z)\Theta(-z) + \Theta(z)]\phi_i(z), \tag{W19C.14}$$

which may be written as

$$\langle \psi_f | H_\gamma | \psi_i \rangle = eE_z (2\pi)^2 \delta(\mathbf{k}'_\parallel - \mathbf{k}_\parallel)(I_1 + I_2). \tag{W19C.15}$$

The first integral is

$$I_1 = \frac{t^* B}{\sin\delta} \int_{-\infty}^{0} dz z\ \exp[z(\lambda - iq')] \sin(qz + \delta) - \frac{t^* B}{2i \sin\delta} \left[\frac{\exp(i\delta)}{[\lambda + i(q - q')]^2} - \frac{\exp(-i\delta)}{[\lambda - i(q + q')]^2} \right], \tag{W19C.16a}$$

and the second integral is

$$I_2 = \int_0^\infty [\exp(-ikz) + r^* \exp(ikz)] z B\ \exp(-\kappa z) dz = B \left[\frac{1}{(\kappa + ik)^2} + \frac{r^*}{(\kappa - ik)^2} \right]. \tag{W19C.16b}$$

Plugging this into Fermi's golden rule gives the transition rate per unit area:

$$\frac{d\Gamma}{dA} = \frac{2\pi}{\hbar} \sum_s \int \frac{d^2 k_\parallel}{(2\pi)^2} \int_0^\infty \frac{dq}{\pi} \int \frac{d^2 k'_\parallel}{(2\pi)^2} \int_{-\infty}^{\infty} \frac{dk}{2\pi} 2\ \sin^2\delta (eE_z)^2 (2\pi)^2 \delta(\mathbf{k}'_\parallel - \mathbf{k}_\parallel) |M|^2 \times \delta(E_i + \hbar\omega - E_f)\Theta(k)\Theta(E_F - E_i)\Theta(E_f - E_F). \tag{W19C.17}$$

where E_F is the Fermi energy level and

$$M = -\frac{t^* \exp(i\delta)}{2i \sin\delta [\lambda + i(q - q')]^2} + \frac{t^* \exp(-i\delta)}{2i \sin\delta [\lambda - i(q + q')]^2} + \frac{1}{(\kappa + ik)^2} + \frac{r^*}{(\kappa - ik)^2}. \tag{W19C.18}$$

The photoelectric yield is obtained by dividing this by the incident number of photons per unit area:

$$Y = \frac{d\Gamma / dA}{I / \hbar\omega} = \frac{8\pi\hbar\omega}{cE_0^2} \frac{d\Gamma}{dA}. \tag{W19C.19}$$

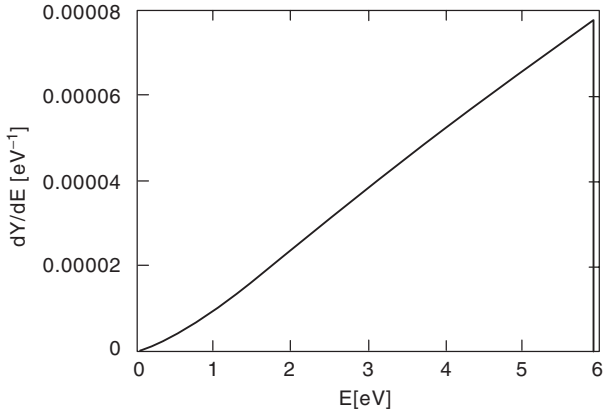

Figure W19C.1. Theoretical differential photoelectric yield of emitted electrons for Al irradiated with 10.2-eV photons. The quantity $dY/d\varepsilon_F$ is defined in Eq. (W19C.22).

The transverse wave-vector integral is

$$\int \frac{d\mathbf{k}_\parallel}{(2\pi)^2}\Theta\left(E_F-\varepsilon_i-\frac{\hbar^2k_\parallel^2}{2m}\right)\Theta\left(-E_F+\varepsilon_f+\frac{\hbar^2k_\parallel^2}{2m}\right)$$
$$=\frac{m}{2\pi\hbar^2}[E_F-\varepsilon_i-\max(0,E_F-\varepsilon_f)]\Theta(E_F-\varepsilon_i-\max(0,E_F-\varepsilon_f)). \quad \text{(W19C.20)}$$

After evaluating the remaining integrals, one finds that

$$Y=\frac{16m\omega e^2}{\pi\hbar^2c}\sin^2\theta\int_0^\infty dq\int_0^\infty dk|M|^2\sin^2\delta\;\delta(\varepsilon_f-\varepsilon_i-\hbar\omega)$$
$$\times\,[E_F-\varepsilon_i-\max(0,E_F-\varepsilon_f)]\Theta(E_F-\varepsilon_i-\max(0,E_F-\varepsilon_f)), \quad \text{(W19C.21)}$$

where θ is the angle of incidence relative to the surface normal.

The *energy distribution curve* (EDC) is obtained by omitting the integration over the variable k and using the energy-conserving delta function to do the q integration. The result is expressed in terms of ε_f:

$$\frac{dY}{d\varepsilon_f}=\frac{8}{\pi}\frac{m^2e^2\omega}{\hbar^4C}\sin^2\theta\frac{|M^2|\sin^2\delta}{\sqrt{\varepsilon_f(V_0+\varepsilon_f-\hbar\omega)}}[E_F-\varepsilon_f+\hbar\omega-\max(0,E_F-\varepsilon_f)]$$
$$\times\,\Theta[E_F-\varepsilon_f+\hbar\omega-\max(0,E_F-\varepsilon_f)]\Theta(\varepsilon_f+V_0-\hbar\omega). \quad \text{(W19C.22)}$$

It is straightforward to show that near threshold the matrix element M is proportional to k.

A theoretical electron EDC is presented for Al in Fig. W19C.1. This is to be compared with experimental results, as shown in Fig. 19.13. In both cases one notes a rise in the photoyield with increasing energy followed by a precipitous drop at high energy, corresponding to electrons emerging from the Fermi surface, giving rise to those with maximum kinetic energy, $(mv^2/2)_{\max}$. There is evidence for band-structure features in the experimental data. Band-structure effects are not included in the simple Sommerfeld model used here.

CHAPTER W20

Thin Films, Interfaces, and Multilayers

W20.1 Strength and Toughness

Having seen how a film adheres to the surface, attention now turns to a study of its mechanical strength. The strength of the bond of a thin film to a substrate may be determined by comparing the surface energies before and after separation. Let $\gamma_{SS'}$ denote the surface tension between the film and the substrate. In delaminating the film from the substrate new solid–vapor interfaces are created, so the change in surface energy per unit area, called the *intrinsic toughness*, is given by the *Dupré formula*:

$$\delta u = \gamma_{SV} + \gamma_{S'V} - \gamma_{SS'}. \tag{W20.1}$$

This is a positive number because it takes energy to create a cleavage.

If sufficient stress is applied to a film in the direction normal to the interface, the film will separate from the surface. The maximum stress the interface can withstand will be denoted by $\sigma_{\max}$. Let $\sigma_{zz}(z)$ denote the stress needed to separate the film a distance z from the equilibrium position, taken to be $z = 0$. Then

$$\delta u = \int_0^\infty \sigma_{zz}(z)\, dz. \tag{W20.2}$$

In the case of metal films on metal substrates, it has been found that the stress may be obtained by taking the derivative of a potential energy per unit area of the empirical form

$$u(z) = F\left(\frac{z}{a}\right) \Delta E, \tag{W20.3}$$

where ΔE and a are parameters that depend on the metals and F is the universal function:

$$F(t) = -(1 + t)e^{-t}. \tag{W20.4}$$

It is believed that this form results from the formation of bond charge at the interface and depends on the exponential falloff of the wavefunctions into vacuum. It is also believed that this formula applies as well to covalent bonds. The stress is therefore

$$\sigma_{zz} = \frac{\Delta E}{a^2} z e^{-z/a}. \tag{W20.5}$$

It rises from zero at the surface, goes through a maximum at $z = a$, and falls off with further increase in z. At the maximum it has the value

$$\sigma_{\max} = \frac{\Delta E}{ae}, \tag{W20.6}$$

where $e = 2.718$. Integrating the analytical formula for the stress results in the expression

$$\sigma_{\max} = \frac{\delta u}{ae} = \frac{\gamma_{SV} + \gamma_{S'V} - \gamma_{SS'}}{ae}. \tag{W20.7}$$

W20.2 Critical Thickness

If a crystalline film grows epitaxially on a substrate in such a way that both are constrained to be flat, there is a critical film thickness beyond which misfit dislocations will develop. This often leads to degradation of the mechanical and electrical properties of the film. The theory of Freund and Nix[†] generalizes earlier work by Matthews and Blakeslee[‡], who analyzed this phenomenon for the case of a thin film on a thick substrate. This critical thickness is determined by the condition that the work needed to produce a dislocation be equal to the strain energy recovered from the system. Letting a_f and a_s be the stress-free lattice constants for the film and substrate, respectively, and ε_f and ε_s be the corresponding strains, one has

$$\varepsilon_m = \frac{a_s - a_f}{a_f} \approx \varepsilon_f - \varepsilon_s \tag{W20.8}$$

for the mismatch strain.

It will be convenient to assume that the film and substrate are both isotropic materials and that they have identical mechanical properties, such as G, the shear modulus, and ν, the Poisson ratio. The film and substrate are subjected to a biaxial stress. The components of the stress tensor may be expressed as $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6) = (P, P, 0, 0, 0, 0)$, where P is the in-plane pressure. The compliance tensor S_{ij} will be of the same form as Eq. (10.18) in the textbook[§] with S_{ij} elements replacing C_{ij} elements. Using Eq. (10.14*b*), the elements of the strain tensor are $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6) = (P(S_{11} + S_{12}), P(S_{11} + S_{12}), 2S_{12}P, 0, 0, 0)$. Note that $\varepsilon_1 = \varepsilon_2 = \varepsilon_m$. The *biaxial modulus* M common to both the substrate and the film is defined by the relation $\varepsilon_1 = P/M$. From Table 10.4, using $S_{11} - S_{12} = 1/(2G)$ and $S_{12} = -\nu S_{11}$, one obtains an expression for the biaxial modulus:

$$M = 2G\frac{1 + \nu}{1 - \nu}. \tag{W20.9}$$

[†] L. B. Freund and W. D. Nix, *Appl. Phys. Lett.*, **69**, 173 (1996).

[‡] J. W. Matthews and A. E. Blakeslee, *J. Cryst. Growth*, **27**, 118 (1974).

[§] The material on this home page is supplemental to the *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."

The net force per unit length on a plane perpendicular to the interface must vanish, so

$$M\varepsilon_f t_f + M\varepsilon_s t_s = 0, \tag{W20.10}$$

where t_f and t_s are the corresponding thicknesses of the film and substrate. Thus

$$\varepsilon_s = -\varepsilon_m \frac{t_f}{t_f + t_s}, \qquad \varepsilon_f = \varepsilon_m \frac{t_s}{t_f + t_s} \tag{W20.11}$$

before any dislocations are generated.

The geometry is illustrated in Fig. W20.1 both before and after the dislocation is formed in the substrate. Let **b** be the Burgers vector of the dislocation, b_x and b_y its components parallel to the interface, and b_z the perpendicular component. From elasticity theory, the long-range attractive force per unit length on the edge dislocation from both free surfaces is estimated to be

$$F(z) = \frac{G[b_x^2 + b_y^2 + (1-\nu)b_z^2]}{4\pi(1-\nu)} \left(\frac{1}{z} - \frac{1}{t_s + t_f - z} \right). \tag{W20.12}$$

The direction of the force is shown in Fig. W20.1. The energy released per unit thickness when the strain in the substrate is relaxed is $\Delta U = M\varepsilon_s t_s b_x$. The work per unit thickness needed to cause a migration of the edge dislocation from the bottom of the substrate to the interface is

$$\begin{aligned} W &= -\int_{r_0}^{t_s} F(z)\, dz = -\frac{G[b_x^2 + b_y^2 + (1-\nu)b_z^2]}{4\pi(1-\nu)} \int_{r_0}^{t_s} \left(\frac{1}{z} - \frac{1}{t_s + t_f - z} \right) dz \\ &= -\frac{G[b_x^2 + b_y^2 + (1-\nu)b_z^2]}{4\pi\ (1-\nu)} \ln \frac{t_s t_f}{r_0(t_s + t_f)}. \end{aligned} \tag{W20.13}$$

where r_0 is a cutoff parameter of atomic dimensions at which macroscopic elasticity theory breaks down. The bottom of the substrate is at $z = 0$. Equating W and ΔU

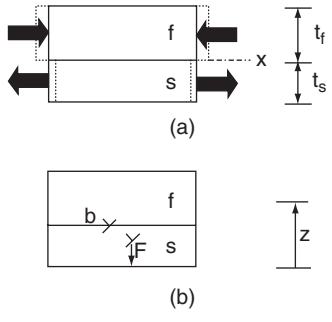

Figure W20.1. (*a*) Film on a substrate subjected to stresses due to lattice mismatch for the case $a_f > a_s$; (*b*) an edge dislocation migrates from a surface to the interface. [From L. B. Freund and W. D. Nix, *Appl. Phys. Lett.*, **69**, 173 (1996). Copyright 1996, American Institute of Physics.]

results in the formula

$$\varepsilon_m = \frac{b_x^2 + b_y^2 + (1-\nu)b_z^2}{8\pi(1+\nu)b_x t_c} \ln \frac{t_c}{r_0}, \tag{W20.14}$$

where a reduced critical thickness is defined by $1/t_c \equiv 1/t_{fc} + 1/t_{sc}$. Equation (W20.14) expresses ε_m in terms of t_c, but this may be inverted numerically to give t_c in terms of ε_m. Note that if the substrate is thick, t_c gives the film thickness t_{fc} directly.

Typical experimental data for Ge_xSi_{1-x} films deposited on a thick Si substrate[†] give the critical thickness as approximately 1000, 100, 10, and 1 nm for $x = 0.1$, 0.3, 0.5, and 1.0, respectively.

W20.3 Ionic Solutions

The description of an ionic solution involves specifying the ionic densities, $n_\pm(\mathbf{r})$, the solvent density, $n_s(\mathbf{r})$, and the potential, $\phi(\mathbf{r})$, as functions of the spatial position $\mathbf{r}$. The presence of a solid such as a metal or semiconductor is likely to introduce spatial inhomogeneities in these quantities. Far from the solid one may expect these variables to reach the limiting values $n_\pm^\infty$, n_s^∞, and ϕ^∞, respectively. It is convenient to take $\phi^\infty \equiv 0$. If the ionic charges are z_+e and $-z_-e$, then bulk neutrality requires that $z_+n_+^\infty = z_-n_-^\infty$. Near the solid deviations from neutrality occur and electric fields are present. In this section the relationship between these quantities is studied.

It is convenient to use a variational principle to derive these equations[‡]. At $T = 0$ K the familiar Poisson equation may be derived from the energy functional:

$$U = \int d\mathbf{r}\, u = \int d\mathbf{r} \left[-\frac{\epsilon}{2}(\nabla\phi)^2 + z_+ e n_+ \phi - z_- e n_- \phi \right]. \tag{W20.15}$$

By using the Euler–Lagrange equation

$$\nabla \cdot \left(\frac{\partial u}{\partial \nabla \phi} \right) = \frac{\partial u}{\partial \phi}, \tag{W20.16}$$

one obtains

$$\nabla^2 \phi = -\frac{e}{\epsilon}(z_+ n_+ - z_- n_-), \tag{W20.17}$$

where ϵ is the electric permittivity of the solvent.

For $T > 0$ K one constructs a quantity analogous to the Helmholtz free energy:

$$F = \int d\mathbf{r}\, f = U - TS, \tag{W20.18}$$

[†] J. C. Bean et al., *J. Vac. Sci. Technol.*, **A2**, 436 (1984).

[‡] The approach is similar to that of I. Borukhov, D. Andelman, and H. Orland, *Phys. Rev. Lett.*, **79**, 435 (1997).

where S is the entropy, defined in terms of an entropy density, s,

$$S = \int d\mathbf{r}\, s. \tag{W20.19}$$

To obtain s imagine partitioning the volume of the solvent into boxes of size V. The number of ions of a given type in a box is $N_\pm = n_\pm V$, and the number of solvent molecules is $N_s = n_s V$. Idealize the situation by imagining that each particle (positive ion, negative ion, or solvent molecule) occupies the same volume. Let N be the number of sites available in volume V. Then $N = N_+ + N_- + N_s$. The number of ways of distributing the particles among the N sites is $W = N!/(N_+!\,N_-!\,N_s!)$. The entropy for the box is given by $S = sV = k_B \ln(W)$. Use of Stirling's approximation results in the expression

$$S = -k_B \int d\mathbf{r} \left(n_+ \ln \frac{n_+}{n} + n_- \ln \frac{n_-}{n} + n_s \ln \frac{n_s}{n} \right), \tag{W20.20}$$

where $n = N/V$. The total numbers of positive and negative ions are fixed. One varies F subject to these constraints

$$\delta \left(F - \mu_+ \int d\mathbf{r}\, n_+(\mathbf{r}) - \mu_- \int d\mathbf{r}\, n_-(\mathbf{r}) \right) = 0, \tag{W20.21}$$

where the chemical potentials $\mu_\pm$ are Lagrange multipliers. Variation with respect to $n_\pm$ and ϕ leads to the Poisson equation, as before, and

$$n_\pm(\mathbf{r}) = (n - n_+(\mathbf{r}) - n_-(\mathbf{r})) \exp[-\beta(\pm z_\pm e\phi(\mathbf{r}) - \mu_\pm)], \tag{W20.22}$$

where $\beta = 1/k_B T$ and use has been made of the fact that $n_s + n_+ + n_- = n$. Evaluating this far from the solid, where $\phi(\mathbf{r}) \to 0$, yields

$$\mu_\pm = k_B T \ln \frac{n_\pm^\infty}{n - n_\pm^\infty - n_\pm^\infty (z_\pm / z_\mp)}. \tag{W20.23}$$

The Poisson equation becomes

$$\nabla^2 \phi = -\frac{ne}{\epsilon} \frac{z_+ n_+^\infty \exp(-\beta z_+ e\phi) - z_- n_-^\infty \exp(\beta z_- e\phi)}{n_s^\infty + n_+^\infty \exp(-\beta z_+ e\phi) + n_-^\infty \exp(\beta z_- e\phi)}. \tag{W20.24}$$

At high charge densities on an interface the right-hand side saturates at a maximum value. Thus, if $\mp\beta z_\pm e\phi \gg 1$,

$$\nabla^2 \phi = \mp \frac{ne}{\epsilon} z_\pm. \tag{W20.25}$$

In the limit where $n_\pm \ll n$ the denominator simplifies and Eq. (W20.24) reduces to what is called the *Poisson–Boltzmann equation*:

$$\nabla^2 \phi = -\frac{e}{\epsilon} [z_+ n_+^\infty \exp(-\beta z_+ e\phi) - z_- n_-^\infty \exp(\beta z_- e\phi)]. \tag{W20.26}$$

In the limit where $|\beta z_{\pm} e\phi| \ll 1$, this reduces further to the *Debye–Hückel* equation:

$$\nabla^2\phi = \frac{1}{\lambda_D^2}\phi, \tag{W20.27}$$

where λ_D is the Debye screening length, given by

$$\frac{1}{\lambda_D^2} = \frac{e^2}{\epsilon k_B T}(z_+^2 n_+^\infty + z_-^2 n_-^\infty). \tag{W20.28}$$

In this case the potential will fall off exponentially with distance as $\phi(z) \propto \exp(-z/\lambda_D)$. The distance λ_D determines the range over which the charge neutrality condition is violated and an electric field exists.

Returning to Eq. (W20.24), in the one-dimensional case, let the solid occupy the half-space $z < 0$. One may obtain a first integral by multiplying through by $d\phi/dz$ and integrating from 0 to ∞:

$$\frac{\beta\epsilon}{2}\left(\frac{d\phi}{dz}\right)^2\bigg|_{z=0} = n \ln \frac{n_s^\infty + n_+^\infty \exp(-\beta z_+ e\phi_0) + n_-^\infty \exp(\beta z_- e\phi_0)}{n_s^\infty + n_+^\infty + n_-^\infty} \tag{W20.29}$$

where ϕ_0 is the solid-surface potential. The quantity $d\phi/dz$ is the negative of the electric field and is related to the charge density on the surface through the boundary condition that D_z is continuous. This is also partly determined by solving the Poisson equation inside the solid and linking the two solutions across the surface. The interface between a semiconductor and an ionic solution is considered in Section W20.4.

W20.4 Solid–Electrolyte Interface

Having considered both the semiconductor and the ionic solution in isolation, we are now in a position to combine them and to study their interface. Some aspects of solid–ionic solution systems have been encountered in Section W12.4 in the discussion of corrosion and oxidation, and in Section 19.11 concerning anodization. To be somewhat general, imagine that both a metal surface and a semiconductor surface are involved (Fig. W20.2). In thermal equilibrium the chemical potential of the electrons is constant throughout the system. Furthermore, there has to be net charge neutrality. Consider what happens when an electrochemical reaction occurs involving an exchange of electrons with the solids. An example is the reduction–oxidation reaction (redox couple) $H_2 \rightleftharpoons 2H^+ + 2e^-$. In the forward direction the reaction is the oxidation of H_2. In the backward direction it is the reduction of H^+. Each species is characterized by its own unique chemical potential in the electrolyte. To dissociate and ionize the H_2 molecule, energy must be supplied equal to the difference in energy between the two species. For the moment, any complications caused by the realignment of the solvation shell of solvent molecules are ignored. The solvation shell consists of those water molecules in the immediate vicinity of the ion whose dipole moments are somewhat aligned by the electric field of the ion.

More generally, consider the redox couple between two hypothetical ionic species labeled A_1 and A_2, of ionic charges $z_1 e$ and $z_2 e$, respectively:

$$n_1 A_1 \rightleftharpoons n_2 A_2 + n e^-. \tag{W20.30}$$

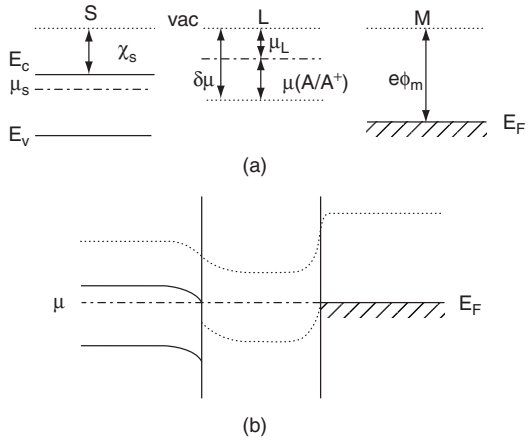

Figure W20.2. Band bending and equalization of Fermi levels in the semiconductor–electrolyte–metal system: (*a*) semiconductor (S), electrolyte (L), and metal (M) in isolation, sharing a common vacuum level; (*b*) band-bending and electrostatic-potential profile when the materials are brought in contact.

The chemical potentials obey the relation

$$n_1(\mu_1 + z_1 e\phi) = n_2(\mu_2 + z_2 e\phi) + n(\mu - e\phi), \tag{W20.31}$$

where the energy shift due to the local electrostatic potential is included. The chemical potentials in solution are given in terms of the activities by the Nernst equation:

$$\mu_i \equiv -ez_i\varepsilon_i = -ez_i\varepsilon_i^0 + k_B T \ln a_i, \tag{W20.32}$$

where ε_i^0 and a_i are the standard electrode potentials and activities of species A_i, respectively. To a first approximation the activities are often set equal to the fractional concentrations, c_i:

$$\mu_i \approx -ez_i\varepsilon_i^0 + k_B T \ln c_i. \tag{W20.33}$$

Charge conservation gives

$$z_1 n_1 = z_2 n_2 - n. \tag{W20.34}$$

Therefore, μ is a sensitive function of the ionic concentrations:

$$\begin{aligned} \mu &= \frac{n_1\mu_1 - n_2\mu_2}{n} \\ &= e\varepsilon - \frac{k_B T}{n} \ln \frac{(c_2)^{n_2}}{(c_1)^{n_1}}. \end{aligned} \tag{W20.35}$$

Here

$$\varepsilon = \frac{n_2 z_2 \varepsilon_2^0 - n_1 z_1 \varepsilon_1^0}{n} \tag{W20.36}$$

is called the *standard redox potential* of the couple. At any given point in the electrolyte the redox reaction is driven backwards or forwards, allowing concentrations of species 1 and 2 to adjust so as to maintain the chemical potentials at constant levels.

In the description above, the energy of reduction of a positive ion (i.e., the energy needed to add an electron to the ion) equals the energy of oxidation (i.e., the energy needed to remove an electron from an atom to create a positive ion). However, when the response of the solvent is included, these energies no longer coincide. The solvent molecules adjust themselves so as to minimize the Coulomb energy of the system. Since charge-exchange reactions alter the net ionic charge, there is a solvent shift of the energy levels. Thermal fluctuations in the solvent cause the energy levels to fluctuate in time. Whenever the energy balance condition is satisfied, a resonant charge exchange process can occur.

The convention is to take the hydrogen couple $H_2 \rightleftharpoons 2H^+ + 2e^-$ as the reference level by which to measure the *redox potentials* (the standard electrode potentials) of other redox couples. Typical couples are presented in Table W20.1 along with their standard redox potentials. The entries are arranged according to how good a reducing agent the atoms are. Thus Li is a strong reducing agent (i.e., it readily donates electrons to a solid). F_2 is a strong oxidizing agent, readily accepting electrons from a solid.

Equation (W20.35) must be modified for use in describing the solid–electrolyte interface. The problem arises because of the arbitrariness of the choice of the hydrogen couple in defining the zero of the standard redox potential. For use in describing the solid–electrolyte interface, both chemical potentials must be referred to the same reference level (e.g., vacuum). It is therefore necessary to find the difference between the standard redox potentials and the energies relative to vacuum, $\delta\mu$ (see Fig. W20.2). Thus Eq. (W20.35) should be replaced by

$$\mu = e\varepsilon + \delta\mu - \frac{k_B T}{n} \ln \frac{(c_2)^{n_2}}{(c_1)^{n_1}}. \tag{W20.37}$$

The value of the offset energy $\delta\mu$ is obtained by looking at the Gibbs free-energy changes (i.e., $\Delta_r G^o$) for a series of reactions (Morrison, 1980) and comparing the result to the value quoted for the standard redox potential:

$Ag^+(g) + e^-$	$= Ag(g)$	-7.57 eV
$Ag(g)$	$= Ag(s)$	-2.95 eV
$Ag^+(aq)$	$= Ag^+(g)$	$+5.00$ eV
$Ag^+(aq) + e^-$	$= Ag(s)$	-5.52 eV

The first line corresponds to the free-space ionization of a silver atom. The second line introduces the cohesive energy of silver. The third line utilizes a calculated value for the solvation energy of a silver ion in water. The solvation energy is the difference in electrostatic energy of an ion of charge +e at the center of a spherical cavity in the water and the electrostatic energy of the ion in free space:

$$U = -\frac{e^2}{8\pi\epsilon_0 a}\left(1 - \frac{1}{\epsilon_r}\right). \tag{W20.38}$$

Here a is the metallic radius of Ag^+ (0.145 nm) and $\epsilon_r(0) = 80$ is the static dielectric constant for H_2O at $T = 27°C$. The value of the standard redox potential for the reaction $Ag^+(aq) + e^- = Ag(s)$ (Table W20.1) is 0.800 eV. Thus $\delta\mu = -5.52 + 0.80 = -4.72$ eV. However, this value must be regarded as being only approximate. It disregards the solvation energy of the electron and underestimates the radius of the solvation shell. Typically, values for $\delta\mu$ in the range -4.5 to -4.8 eV are employed in the literature.

Electrons in an isolated semiconductor will, in general, have a chemical potential which is different from that of an electron in an electrolyte. This is illustrated in Fig. W20.2. The upper half of the diagram shows the semiconductor (S), electrolyte (L), and metal (M) isolated from each other, sharing a common vacuum level. Note that the chemical potential of an electron in the electrolyte, μ_L, is determined by subtracting the chemical potential for the redox couple, $\mu(A/A^+)$ [given by Eq. (W20.37)], from the offset energy $\delta\mu$, as in Fig. W20.2.

When the two are brought into contact, as in the lower half of Fig. W20.2, there will be a charge transfer and the chemical potentials will equilibrate. This will cause band bending in the semiconductor in much the same way that it was caused in the p-n junction. At the two interfaces there is not charge neutrality and electric fields exist due to the dipole double layers.

W20.5 Multilayer Materials

One rather simple use of multilayers is to fabricate optical materials with interpolated gross physical characteristics. For example, one could achieve an interpolated index of refraction n by alternating sufficiently thin layers of indices n_1 and n_2. The linear interpolation formula, $n = (1 - f)n_1 + fn_2$, where f is the fraction of space occupied by material 2, would only give a crude approximation to n and is not physically

TABLE W20.1 Standard Redox Potential Energies at $T = 25°C$

Redox Couple	ε (V)
$Li = Li^+ + e^-$	3.045
$Rb = Rb^+ + e^-$	2.925
$K = K^+ + e^-$	2.924
$Cs = Cs^+ + e^-$	2.923
$Na = Na^+ + e^-$	2.711
$Mn = Mn^{2+} + 2e^-$	1.029
$Zn = Zn^{2+} + 2e^-$	0.763
$Cu = Cu^{2+} + 2e^-$	0.34
$Pb = Pb^{2+} + 2e^-$	0.126
$H_2 = 2H^+ + 2e^-$	0.000
$Cu^+ = Cu^{2+} + e^-$	−0.153
$Fe^{2+} = Fe^{3+} + e^-$	−0.770
$Ag = Ag^+ + e^-$	−0.800
$2Br^- = Br_2 + 2e^-$	−1.065
$2Cl^- = Cl_2 + 2e^-$	−1.358
$2F^- = F_2 + 2e^-$	−2.870

motivated. A better interpolation could be obtained by recalling that $n_i = \sqrt{\epsilon_{r_i}}$ and making use of the Clausius–Mossotti formula, Eq. (8.40). That formula showed that the ratio $(n^2 - 1)/(n^2 + 2)$ may be expressed as a linear combination of polarizability contributions from each of the materials present in a composite medium. Thus an appropriate interpolation formula would be

$$\frac{n^2 - 1}{n^2 + 2} = (1 - f)\frac{n_1^2 - 1}{n_1^2 + 2} + f\frac{n_2^2 - 1}{n_2^2 + 2}. \tag{W20.39}$$

The design is valid provided that the length scale of the periodicity is small compared with the wavelength of light.

The linear interpolation formula $\kappa = (1 - f)\kappa_1 + f\kappa_2$ could be used to fabricate materials with interpolated thermal conductivities. However, this is only approximate, since the interface region between two media often has different physical properties from either medium, including its own thermal resistance due to phonon scattering.

As another example of linear interpolation, suppose that there are two physical properties, denoted by n and p, that one would like to obtain. Assume that there are three materials, with values (n_1, n_2, n_3) and (p_1, p_2, p_3), respectively. Construct the multilayer by taking lengths (a_1, a_2, a_3) such that the superlattice has periodicity

$$a_1 + a_2 + a_3 = D. \tag{W20.40}$$

Then, assuming simple additivity of the properties, one has

$$a_1 n_1 + a_2 n_2 + a_3 n_3 = Dn, \tag{W20.41a}$$

$$a_1 p_1 + a_2 p_2 + a_3 p_3 = Dp. \tag{W20.41b}$$

These three linear equations may be solved for the lengths a_1, a_2, and a_3. One finds that

$$\frac{a_1}{D} = \frac{1}{\Delta}[(n_2 p_3 - p_2 n_3) + (p_2 - p_3)n + (n_2 - n_3)p], \tag{W20.42a}$$

$$\frac{a_2}{D} = \frac{1}{\Delta}[(n_3 p_1 - p_3 n_1) + (p_3 - p_1)n + (n_3 - n_1)p], \tag{W20.42b}$$

$$\frac{a_3}{D} = \frac{1}{\Delta}[(n_1 p_2 - p_1 n_2) + (p_1 - p_2)n + (n_1 - n_2)p], \tag{W20.42c}$$

where

$$\Delta = n_2 p_3 + n_3 p_1 + n_1 p_2 - p_2 n_3 - p_3 n_1 - p_1 n_2. \tag{W20.43}$$

The extension to a higher number of variables is obvious.

W20.6 Second-Harmonic Generation in Phase-Matched Multilayers

Nonlinear polarization is introduced in Section 8.9 and discussed further in Section 18.6. For efficient second-harmonic generation one needs two things: a material with a large nonlinear electrical susceptibility and birefringence. The latter is needed so that

phase matching between the primary beam at frequency ω and the secondary beam at frequency 2ω can be obtained over a long coherence length. The semiconductor GaAs has a large $\chi^{(2)}$ (240 pm/V) but is a cubic crystal, so is optically isotropic and not birefringent. By constructing a multilayer structure with interspersed thin layers of oxidized AlAs (Alox), artificial birefringence is obtained[†].

Here one uses the approximate additivity of the dielectric function for the TE mode of propagation:

$$\epsilon_{\mathrm{TE}} = (1 - f)\epsilon_{r_1} + f\epsilon_{r_2}. \tag{W20.44}$$

The TE mode of a waveguide has the electric field perpendicular to the direction of propagation, but the magnetic field need not be. Similarly, the approximate additivity of the inverse of the dielectric function for the TM mode of propagation yields

$$\frac{1}{\epsilon_{\mathrm{TM}}} = \frac{1 - f}{\epsilon_{r_1}} + \frac{f}{\epsilon_{r_2}}. \tag{W20.45}$$

The TM mode has a magnetic field perpendicular to the propagation direction. In Eqs. (W20.44) and (W20.45), ϵ_{r_1} and ϵ_{r_2} are the respective dielectric functions of the materials and f is the filling fraction. The respective indices of refraction for GaAs and Alox are $n_1 = \sqrt{\epsilon_{r_1}} = 3.6$ and $n_2 = \sqrt{\epsilon_{r_2}} = 1.6$. The net birefringence is determined by the difference in the indices of refraction for the TE and TM modes:

$$\Delta n = \sqrt{\epsilon_{\mathrm{TE}}} - \sqrt{\epsilon_{\mathrm{TM}}}. \tag{W20.46}$$

This, in turn, is a function of the filling fraction and may therefore be engineered to specifications.

The same concept may be used to the advantage of another nonlinear process, *difference frequency generation* (DFG). In this process, photons of frequencies ω_1 and ω_2 are mixed together to produce a photon of frequency $|\omega_1 - \omega_2|$.

W20.7 Organic Light-Emitting Diodes

Recently, a structure composed partly of stacked organic films was designed to act as a tunable three-color transparent organic light-emitting diode (TOLED). Since the additive primary colors are red, blue, and green, this device can function as a universal light-emitting diode. The structure is illustrated in Fig. W20.3. Electron injection into the upper organic layer is through the low work function Mg:Ag cathode. The transparent conductor indium tin oxide (ITO) serves as the anodes. The organic molecules used are 4,4′-bis[*N*-(1-napthyl)-*N*-phenylamino]biphenyl (α-NPD), which is a hole conductor, bis(8-hydroxy)quinaldine aluminum phenoxide ($\mathrm{Alq}_2'\mathrm{Oph}$), which fluoresces in the blue, and tris(8-hydroxyquinoline aluminum) (Alq_3), which is an electron conductor and fluoresces in the green. By doping Alq_3 with 3% 5,10,15,20-tetraphenyl-21*H*,23*H*-porphine (TPP), the fluorescent band is pulled down to the red. A layer of crystalline 3,4,9,10-perylenetetracarboxylic dianhydride (PTCDA) serves as a transparent hole conductor and shields the sensitive organic layer against ITO sputtering. One of the

[†] A. Fiory et al., *Nature*, **391**, 463 (1998).

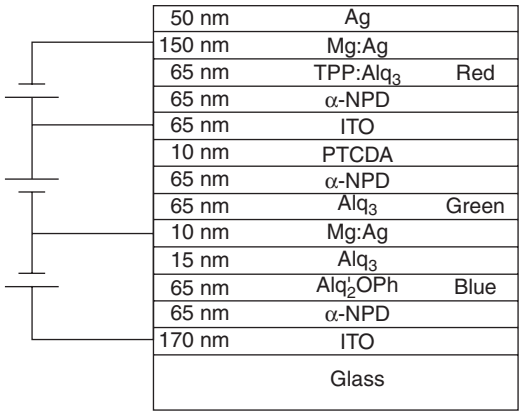

Figure W20.3. Three-color tunable organic light-emitting device. [Reprinted with permission from Z. Shen et al., *Science*, **276**, 2009 (1997). Copyright 1997, American Association for the Advancement of science.]

keys to success in fabricating this device is that amorphous and organic films tend not to be tied down by the need to satisfy lattice-matching constraints.

W20.8 Quasiperiodic Nonlinear Optical Crystals

A recent application of multilayer structures to the field of nonlinear optics involves the construction of a periodic superlattice. For example, to carry out second-harmonic generation efficiently, phase matching is required (i.e., the material must be able to simultaneously satisfy momentum and energy conservation). However, $\mathbf{k}(2\omega) - 2\mathbf{k}(\omega) = \mathbf{K}_{21} \neq 0$, in general. Similarly, for third-harmonic generation, $\mathbf{k}(3\omega) - 3\mathbf{k}(\omega) = \mathbf{K}_{31} \neq 0$. By constructing a superlattice with the periodicity $2\pi/K_{21}$ or $2\pi/K_{31}$, the index of refraction will possess this periodicity and will be able to supply the missing wave vector. The strength of the scattering amplitude will involve the Fourier component of the index of refraction at that wave vector. This scheme has been applied to such nonlinear crystals as $LiNbO_3$.

It is also possible to construct a quasiperiodic lattice (one-dimensional quasicrystal) which can supply K_{21} and K_{31} simultaneously. It is assumed that these wave vectors are such that K_{31}/K_{21} is not a rational number. Such a structure can be based on the Fibonacci sequence of layers ABAABABAABAAB.... Such a crystal using $LiTaO_3$ has been built[†]. In that scheme the A and B layers each had a pair of antiparallel ferroelectric domains. The thicknesses of the domains were L_{A1} and L_{A2} in layer A and L_{B1} and L_{B2} in layer B. Let $L_A = L_{A1} + L_{A2}$ and $L_B = L_{B1} + L_{B2}$ and assume that $L_{A1} = L_{B1} = L$. Let $L_{A2} = L(1 + \eta)$ and $L_{B2} = L(1 - \eta\tau)$, with $\tau = (1 + \sqrt{5})/2$ and η a small number. Let $D = \tau L_A + L_B$ be a characteristic distance. Then the vectors $G_{m,n}$ serve as quasiperiodic reciprocal-lattice vectors

$$G_{m,n} = \frac{2\pi}{D}(m + n\tau). \tag{W20.47}$$

[†] S. Zhu et al., *Science*, **278**, 843(1997).

There exist a set of numbers (m_1, n_1) that make $G_{m_1,n_1} \approx K_{21}$ and another pair (m_2, n_2) that make $G_{m_2,n_2} \approx K_{31}$. Thus both K_{21} and K_{31} are provided by the structure. In the reference cited above, the values used for the structural parameters were $L = 10.7$ μm and $\eta = 0.23$.

W20.9 Graphite Intercalated Compounds

Graphite consists of graphene layers of sp^2-bonded carbon rings arranged in the stacking sequence ABAB... and separated by 0.335 nm, which is substantially larger than the nearest-neighbor distance of 0.142 nm. The in-plane lattice constant of the hexagonal sheet is 0.246 nm. The layers are only weakly bound together by van der Waals forces. It is possible to insert foreign atoms and molecules in the interlayer region to form graphite intercalated compounds (GICs). It is found that the atoms intercalate in well-defined stoichiometric ratios, forming compounds such as KC_{24}. In one type of arrangement one layer of intercalate is followed by n graphene layers, as illustrated in Fig. W20.4*a*. This is called an *n-stage GIC*. For example, KC_{24} can exist as a two-stage compound $KC_{12\times2}$ or a three-stage compound $KC_{8\times3}$. Values of n up to 8, or higher, are not uncommon. In other compounds there may be several intercalate layers, followed by n graphene layers. In still other situations the intercalates may form islands arranged in an array interspersed in the graphite structure (the Daumas–Herold domain structure). This is illustrated in Fig. W20.4*b*.

The distance between successive intercalate layers, d_c, depends on the degree of staging. Different forms of ordering are found in the GICs. The intercalated layers could either be commensurate or incommensurate with the host lattice. The graphene layers could either maintain the ABAB.... stacking sequence or adopt some other sequence, such as AB/BA/AB/BA/... (where a slash denotes an intercalated layer). The intercalate could exist as an ordered two-dimensional crystal, a disordered glass, or even a liquid.

The intercalated atoms and molecules may act as either donors or acceptors. In either case, carriers are injected into the π bands of the graphene sheet. Typical donors are the alkali metals, which form GICs such as LiC_6, LiC_{12}, LiC_{18}, KC_8, KC_{24}, ... ,

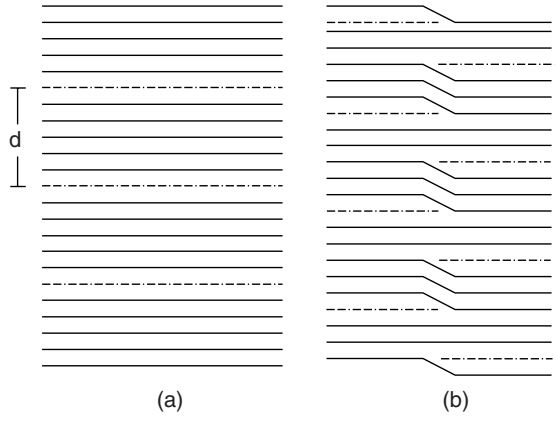

Figure W20.4. Graphite intercalated compounds: (*a*) $n = 5$ stage compound; (*b*) island intercalation.

KC_{72}, RbC_8, RbC_{24} or CsC_8, and CsC_{24}. Acceptor compounds are $C_{10}HNO_3$, $C_{14}Br$, or $C_{16}AsF_5$. Note the convention of placing the chemical symbol for the donors to the left of the carbon and the symbol for acceptors to the right.

Staging results from the interplay of various microscopic forces. Charge transfer is brought about by the difference in chemical potentials between the graphite and the intercalate. This, by itself, lowers the energy of the system. The Coulomb interaction between the layers, partially screened by the mobile carriers in the graphite, is important in establishing the staging. Elastic interactions are also involved, since the layer spacing of the host lattice is altered to accommodate the intercalated layer. One of the early attempts[†] at describing the system theoretically involved the introduction of the model internal energy:

$$\frac{U}{N_0} = t\sum_i \sigma_i - \frac{u}{2}\sum_i \sigma_i^2 + \frac{1}{2}{\sum_{ij}}' V_{ij}\sigma_i\sigma_j, \tag{W20.48}$$

where N_0 is the number of intercalation sites in a layer and σ_i is the fractional occupancy of the ith layer, a number between 0 and 1. The first two terms represent the interaction of the intercalate with the host, and the bonding of the intercalate to form a two-dimensional solid, respectively. The third term describes the screened Coulomb energy and is positive. The parameters V_{ij} are taken to be of the form $V_{ij} = (V/2)|z_{ij}|^{-\alpha}$, where z_{ij} is the interplanar distance. This form is suggested by making a Thomas–Fermi analysis of the screening for large n. The quantities t, u, V, and $\alpha(\approx 5)$ parametrize the theory.

The entropy for a given layer is determined by partitioning $N_0\sigma_i$ intercalate atoms among N_0 sites. Since there are $W_i = N_0!/[(N_0\sigma_i)!(N_0 - N_0\sigma_i)!]$ ways of doing this, the layer entropy is, by Stirling's approximation,

$$S_i = k_B \ln W_i = -k_B N_0[\sigma_i \ln \sigma_i + (1-\sigma_i)\ln(1-\sigma_i)]. \tag{W20.49}$$

The Helmholtz free energy for the system is

$$\frac{F}{N_0} = t\sum_i \sigma_i - \frac{u}{2}\sum_i \sigma_i^2 + \frac{1}{2}{\sum_{ij}}' V_{ij}\sigma_i\sigma_j + k_BT\sum_i[\sigma_i \ln \sigma_i + (1-\sigma_i)\ln(1-\sigma_i)]. \tag{W20.50}$$

Only the layers with nonzero σ_i contribute to F. The chemical potential for the ith layer is given by

$$\mu_i = \frac{1}{N_0}\frac{\partial F}{\partial \sigma_i} = t - u\sigma_i + {\sum_j}' V_{ij}\sigma_j + k_BT\left[\ln \sigma_i - \ln(1-\sigma_i)\right]. \tag{W20.51}$$

Setting all the chemical potentials equal to μ leads to the set of coupled equations

$$\sigma_i = \frac{1}{1 + e^{\beta(t - u\sigma_i + {\sum_j}' V_{ij}\sigma_j - \mu)}}. \tag{W20.52}$$

[†] S. A. Safran, Stage ordering in intercalation compounds, H. Ehrenreich and D. Turnbull, eds., *Solid State Physics*, Vol. 40, Academic Press, San Diego, Calif., 1987, p. 183.

For a given set of staging occupancies it is possible to obtain $\mu(T)$, F, and the other thermodynamic variables.

Further refinements in the theory have evolved over the years. Interest in GICs stems largely from the fact that their electrical conductivity is high and may be varied in a controlled way by changing the stoichiometry.

Graphite fluorides $(CF)_n$ have been used as cathodes in lithium batteries. By itself, $(CF)_n$ is a poor electrical conductor, so it is often combined with a good electrical conductor such as graphite. The anode is made of lithium. Such lithium batteries have high specific energy (360 W·h/kg) and a high voltage (3 V). The material $(CF)_n$ is a stage 1 compound with every C atom bonded to a fluorine. The layers alternate in the sequence CFCFCF. . . . The lattice constants are $a = 0.257$ nm and $c = 0.585$ nm.

Other GICs that may potentially be used as cathodes have intercalant anions such as PF_6^-, AsF_6^-, and SbF_6^-. The obstacle to their use is the lack of a suitable electrolyte. Superconductivity is also observed in GICs (see Chapter W16).

REFERENCES

Critical Thickness

Freund, L B., and W. D. Nix, *Appl. Phys. Lett.*, **69**, 173 (1996).

Ionic Solutions

Borukhov, I., D. Andelman, and H. Orland, *Phys. Rev. Letters.*, **79**, 435 (1997).

Solid-Electrolyte Interface

Morrison, S. R., *Electrochemistry at Semiconductor and Oxide Metal Electrodes*, Plenum Press, New York, 1980.

Second-Harmonic Generation in Phase-Matched Multilayers

Fiory, A., et al., *Nature*, **391**, 463 (1998).

Organic Light-Emitting Diodes

Shen, Z., et al., *Science*, **276**, 2009 (1997).

Quasi-periodic Nonlinear Optical Crystals

Zhu, S., et al., *Science*, **278**, 843 (1997).

Graphite Intercalated Compounds

Zabel, H., and S. A. Solin, eds., *Graphite Intercalation Compounds*, Springer-Verlag, New York, 1990.

PROBLEM

W20.1 Consider the case of a thin film deposited on a thick substrate ($t_f \ll t_s$).

(a) Show that the resulting strains in the substrate and film are $\epsilon_s \approx 0$ and $\epsilon_f \approx (a_{s0} - a_{f0})/a_{f0}$, respectively, where a_{s0} and a_{f0} are the stress-free lattice constants of the substrate and film.

(b) Show that the strain in the film can be relieved completely if the misfit dislocations at the film/substrate interface are, on the average, separated by a distance $d = a_{s0}/|\epsilon_m|$, where ϵ_m is the misfit strain defined by Eq. (W20.8).

CHAPTER W21

Synthesis and Processing of Materials

W21.1 Synthesis and Processing Procedures

The various procedures used in the synthesis and processing of materials can be grouped into a few general classes. Specific examples of many of these procedures are given in Chapter 21 of the textbook[†] and in this chapter. Important classes of *synthesis* include those that produce materials in bulk form or in forms with reduced dimensionality (e.g., powders, fibers, and thin films or layers and surface coatings). Bulk materials and larger powders often require further processing to produce materials with the final desired shape or form. *Processing* that changes only the form and not the microstructure of a material is not stressed here. Smaller powders, fibers, and thin films are more often prepared in essentially their final form but may still require further processing to achieve the desired microstructure.

Important classes of materials synthesis and processing procedures are listed in Table W21.1. Specific examples discussed here and in the textbook are also indicated.

A wide range of energy sources are used in the synthesis and processing of materials, depending on the specific procedure involved and the products desired. Some important examples are listed in Table W21.2.

W21.2 Heteroepitaxial Growth

Consider the case where atoms of type A, with lattice constant a in the solid state, are deposited on a flat substrate consisting of atoms of type B, with lattice constant b, where $b > a$. Assume that the symmetries of the two crystals are the same. At first the A atoms may form a monolayer in registry with the substrate. As additional layers are deposited, however, the bulk strain energy in A builds up since there is a lattice-mismatch strain given by $(b - a)/a$ [see Eq. (W20.8)]. The strain may be relieved by having misfit dislocations form at the interface or, alternatively, by having the surface of the A crystal warp. These possibilities are illustrated in Fig. W21.1. Misfit dislocations are discussed in Section W20.2.

If the surface warps, an undulating pattern appears that may be observed using such high-resolution instruments as the transmission electron microscope or the atomic force microscope. The condition for warping is that the additional surface energy needed to curve the surface be less than the bulk strain energy relieved by allowing the adsorbate

[†] 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."

TABLE W21.1 Important Classes of Materials Synthesis and Processing Procedures

- Synthesis of bulk samples
  - Synthesis from the liquid phase
    - Czochralski method for growth of single-crystal Si (Section 21.6)
    - Liquid-phase epitaxy (LPE): GaAs
    - Bridgman method
    - Sol-gel synthesis (Section 21.12)
    - Rapid solidification (Section W21.12)
    - Flux growth of ceramics using oxide fluxes
    - Arc melting of metallic alloys
    - Hydrothermal growth: crystalline quartz, TGS, ADP, KDP
  - Synthesis from solid powders or bulk material
    - Sintering of powders (Section 21.11)
    - Catalysis (Section 21.14)
    - Polymers (Section 21.13 to 21.15 and W21.21 to W21.25)
    - High pressure–high temperature synthesis of diamond crystals
  - Synthesis from the vapor phase
    - Modified Lely process (SiC platelets): PVD (Section W21.17)
- Synthesis of fine particles or powders
  - Grinding (Section 21.11)
  - Plasma spraying
  - Gas condensation: carbon nanotubes (Section 21.15)
  - Nucleation from a saturated liquid phase
- Synthesis of fibers
  - Drawing from the melt: silica fibers
- Synthesis of thin films and surface coatings
  - Synthesis from the vapor phase
  - Chemical vapor deposition (CVD) (Section W21.5)
    - Molecular beam epitaxy (MBE) (Section W21.6)
    - Metal–organic CVD (MOCVD), also known as metal–organic vapor-phase epitaxy (MOVPE)
    - Plasma-enhanced CVD (PECVD) (Section W21.7)
  - Physical vapor deposition (PVD)
    - Sputter deposition (reactive versus nonreactive) (Section W21.3)
    - Ion beam deposition
    - Thermal evaporation (electron beam or hot filament)
    - Thermal spraying
  - Synthesis from the liquid phase
    - Chemical deposition (surface plating via immersion)
    - Electrochemical deposition or electroplating (surface plating via passage of a current through a solution)
  - Synthesis via chemical reactions
    - Reaction between a vapor or a liquid and the surface
      - Thermal oxidation: $Si(s) + O_2(g) \rightarrow SiO_2(s)$ (Section 21.7)
- Processing
  - Annealing
    - Rapid thermal annealing
  - Oxidation
    - a-SiO_2 via thermal oxidation or SIMOX (Section 21.7)
  - Doping
    - Via diffusion or ion implantation

TABLE W21.1 (*Continued*)

Ion implantation (Section W21.3)
For surface modification (e.g., carburizing, nitriding, etc.) (Section W21.13)
Etching (Section W21.8)
Plasma treatments (Section W21.8)
Float-zone purification (Section W21.4)
Lithography (Section W21.8)
Mechanical processing (Section W21.10)
Work hardening

TABLE W21.2 Sources of Energy Used in Synthesis and Processing

Thermal (heating due to contact with hot gases and/or thermal radiation)
Annealing
Rapid thermal processing
Pressure and temperature
Sintering
Shock compression
Plasma (heating due to energy absorbed from accelerated electrons and ions, emitted light, also the direct effects of Joule heating)
Electromagnetic radiation
Laser beams
Electric fields and the kinetic energy of accelerated ions
Sputtering

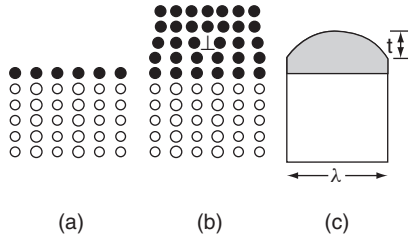

Figure W21.1. Epitaxial growth: (*a*) monolayer of atoms in registry with the substrate; (*b*) formation of a misfit dislocation; (*c*) warping of an adsorbed thick layer of atoms.

to relax its strain. The condition for this may be estimated by assuming a parabolic profile for the warp $y = 4tx(\lambda - x)/\lambda^2$, where t is the height of the warp and λ is the periodicity. If $t \ll \lambda$, the change in surface area is $\Delta A = 8wt^2/3\lambda$ and the volume of the warp is $\Delta V = 4tw\lambda/6$, where w is the surface dimension transverse to the warp. The strain energy relieved is approximately $E\varepsilon^2 \Delta V/2$, where the mismatch strain is given by $\varepsilon = (b/a) - 1$, and E is the Young's modulus of the adsorbate. The increase in surface energy is $\sigma\,\Delta A$, where σ is the energy per unit area at the vacuum interface. This leads to the condition

$$\frac{t}{\lambda^2} < \frac{E}{8\sigma}\left(1 - \frac{b}{a}\right)^2 \tag{W21.1}$$

for the development of an undulating surface pattern rather than misfit dislocations.

Recently, a lattice-engineered compliant substrate has been invented which does not cause the adsorbate to develop misfit locations or to warp.[†] This is important, because it permits epitaxial growth of badly mismatched materials without sacrificing crystal quality.

The compliant substrate is a bilayer substrate that is created by having an adsorbed layer bonded to a substrate of the same material but at a twisted angle, as illustrated in Fig. W21.2. The two layers interact, go into partial registry in a domainwise fashion, and form domain walls consisting of screw dislocations, as is shown in Fig. W21.3. This embeds an intrinsic strain into the bilayer substrate. Since the interatomic forces are anharmonic, with the spring constants becoming substantially weaker as the bonds are stretched, the effective spring constants for the substrate are less stiff than they would be for a fully periodic substrate. The compliant substrate is therefore able to deform readily to accommodate an adsorbate with a different lattice constant.

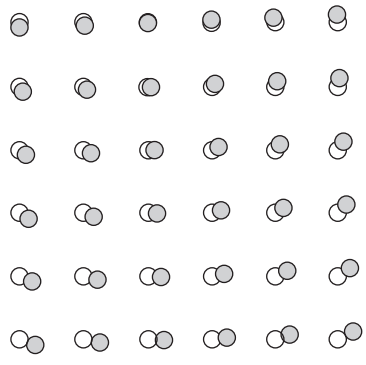

Figure W21.2. Bilayer substrate consisting of a base layer bonded to a twisted overlayer.

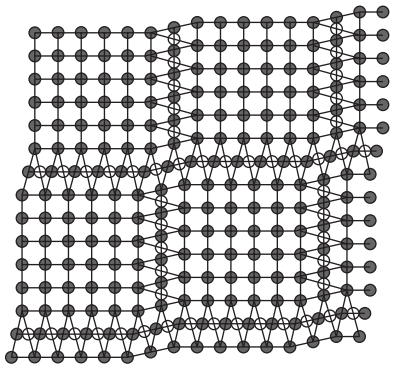

Figure W21.3. Accommodation of the bilayer by the formation of registered domains with domain walls formed by screw dislocations. [Adapted from F. E. Ejeckam et al., *Appl. Phys. Lett.*, **70**, 1685 (1997).]

[†] F. E. Ejeckam et al., *Appl. Phys. Lett.*, **70**, 1685 (1997).

Thin-Film Growth Modes. The *nucleation* and *growth* of thin films on solid surfaces can involve a variety of atomic processes, including adsorption, surface diffusion, and the formation of chemical bonds between adatoms and also between adatoms and atoms of the surface at specific surface sites. These surface processes are discussed in detail in Chapters 19 and W19. Three main modes of thin-film crystal growth are believed to occur at surfaces, at least in those cases in which interdiffusion or chemical reaction between the adsorbing species and the substrate does not lead to the formation of an alloy, chemical compound, or intermetallic compound and in which surface defects such as steps or dislocations do not play a dominant role in the nucleation stage of film growth. Other important modes of thin-film growth include, for example, processes such as the reaction of O_2 with the surface of Si at high temperatures leading to the growth of an amorphous SiO_2 layer or the formation of silicides when metals such as Cu, Au, Ni, Pd, and Pt are deposited on Si.

The three thin-film growth modes to be described here are the *island growth mode*, also known as the *Volmer–Weber mode*, the *layer growth mode*, also known as the *Frank–van der Merwe mode*, and the *layer-plus-island growth mode*, also known as the *Stranski–Krastanov mode*. These growth modes are illustrated schematically in Fig. W21.4. To aid in their description, use will be made of the *surface free energies* σ_A and σ_B of the growing film and the substrate, respectively, as well as the *free energy* σ_{AB} of the A–B *interface*. Examples of thin films growing in each growth mode will also be given. It is, of course, doubtful that concepts such as surface energies can be applied to thin films which nucleate on surfaces as single atoms. In such cases, an atomistic point of view that focuses on individual atomic processes and the potential energies of interaction of adsorbate atoms with the substrate and with each other must be employed. The nucleation of the new phase, whether it be in the form of a cluster or a monolayer, is often a rate-determining step in thin-film growth and, in general, must be understood as resulting from atomic interactions.

Useful reviews of the processes involved in the nucleation and growth of thin films and also of the three growth modes discussed here can be found in Venables et al. (1984) and Venables (1994). Another approach that describes the deposition of thin films from thermal beams and focuses on four different types of atom/molecule-surface interactions has been given by Voorhoeve (1976). A variety of techniques are used to monitor thin-film growth, either in situ or ex situ. These include transmission and

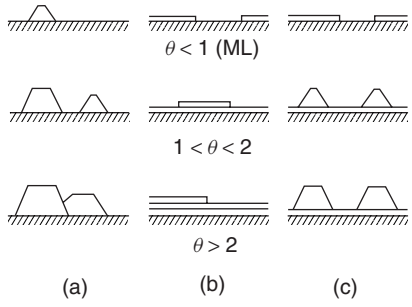

Figure W21.4. Three main thin-film growth modes (ML = monolayer): (*a*) island growth mode, also known as the Volmer–Weber mode; (*b*) layer growth mode, also known as the Frank–van der Merwe mode; (*c*) layer-plus-island growth mode, also known as the Stranski–Krastanov mode.

scanning electron microscopies (TEM and SEM, respectively), reflection high-energy electron diffraction (RHEED), Auger electron spectroscopy (AES), and, more recently, various forms of scanning tunneling microscopy (STM).

Island Growth Mode (Volmer–Weber). In this growth mode, small clusters of adsorbing atoms (or molecules) nucleate on the substrate surface and, if they are stable, continue growing as islands until they coalesce. The islands grow by incorporating atoms that reach the island directly from the vapor phase or by diffusing across the surface. This growth mode is believed to occur when the atoms or molecules of the growing film are more strongly bonded to each other than to the substrate or, in terms of the surface and interface free energies, when $\sigma_A + \sigma_{AB} > \sigma_B$. This inequality is only qualitatively correct since it does not take into account the free energy of A atoms within the bulk of the film when the deposited islands are more than one monolayer thick. Island growth is also expected when the lattice parameters of the film and substrate are very different and when the two lattices cannot be brought into some form of epitaxial alignment by rotation.

Examples of this growth mode include metal films deposited on insulating substrates such as the alkali halides (e.g., NaCl), on the basal plane of graphite and other layered materials, such as MoS_2 and mica, and on insulators such as MgO. By measuring the densities and sizes of stable Au or Ag clusters on the (100) surfaces of alkali halides and comparing with existing theoretical models, researchers have been able to determine that the size of a *stable nucleus* is usually just one metal atom. In addition, values for the exponential prefactors and activation energies associated with desorption and surface diffusion have been determined. Effects associated with cluster mobility at high temperatures can play important roles in this mode of thin-film growth and are therefore often included in the growth models.

Surface reconstructions are common on semiconductor surfaces and can complicate thin-film growth due to the resulting surface anisotropy and possibly to steps with different heights on the same surface. The presence of surface impurities such as carbon or oxygen or of defects such as dislocations can lead to island growth and defective films. In the case of heteroepitaxy [e.g., Si on SiO_2 or on Al_2O_3 (sapphire)], island growth is typically observed, with critical nucleus sizes in the range of one to four atoms.

Layer Growth Mode (Frank–van der Merwe). In this growth mode the adsorbing atoms form a monolayer on the substrate, and additional nucleation and layer growth can occur simultaneously on the substrate and also on the previously deposited layers. The growth in this mode can appear complex, for kinetic reasons (Fig. W21.5), when the thickness of the region in which growth is occurring corresponds to several monolayers. The actual structure of this growth zone or interface transition region will depend

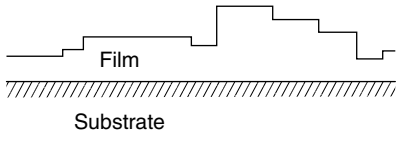

Figure W21.5. Layer growth mode showing nucleation occurring within a multilayer growth zone.

on the relative rates of the nucleation and growth processes. When the nucleation rate is high and monolayer growth is slow, the growth zone will be wider than when nucleation is slow and layer growth is fast. When the growth rate is high enough, deposition will occur monolayer by monolayer (i.e., each monolayer will be essentially completed before nucleation of a new monolayer occurs).

Monolayer-by-monolayer growth can readily be monitored via RHEED, in which case regular oscillations of the RHEED intensity occur with the same period as the monolayer growth. These oscillations are observed when nucleation of each new monolayer occurs on the terraces of existing monolayers but not when growth occurs by step flow (i.e., by the addition of adatoms to existing steps on an off-axis substrate). Decay of the RHEED oscillations can provide evidence for the development of surface roughness due to widening of the growth zone from a single monolayer to several monolayers.

Nucleation will be enhanced at high supersaturations (i.e., high incoming fluxes of growth species) while growth will be enhanced at high temperatures, as long as the temperature is not so high that the growth species tend to be desorbed from the surface before they are incorporated into the growing monolayer. In the limits of very high supersaturation and low temperature, the growing film can be quite disordered and may even be amorphous.

This layer growth mode is believed to occur when the atoms or molecules in each monolayer are more tightly bonded to the substrate than to each other or, in terms of surface and interface free energies, when $\sigma_A + \sigma_{AB} < \sigma_B$. This condition is analogous to that presented in Section W20.1 for the wetting of liquids on surfaces. In some cases the second monolayer to be formed in this growth mode may be less tightly bonded to the first monolayer than the first monolayer is to the substrate.

Examples of this growth mode include inert gases on graphite, some alkali halides, and metal-on-metal [e.g., Ni on Cu(100) or Cu(111) and Ag on W(110)] and semiconductor-on-semiconductor growth systems. Interesting examples include FCC Fe on Ni, Cu, and Au, where the normal BCC crystal structure of α-Fe (ferrite) is not stable due to the strain imposed by the substrate. Misfit dislocations often appear at finite thicknesses in the case of the *heteroepitaxial* growth of metals on metals due to strain in the growing film.

The epitaxial growth of the semiconductors Si, Ge, GaAs, $Ga_{1-x}Al_xAs$, and other compound and alloy semiconductors has been studied widely. In the case of *homoepitaxy* [e.g., Si on Si(100)] the layer growth mode is observed under the ideal conditions of clean substrate surfaces and the high temperatures required for the adatom surface mobility that is necessary to allow crystalline films to be formed. Growth is often carried out on vicinal surfaces that are slightly off-axis ($\approx 1^\circ$ to 4°), in order to have available regular arrays of surface steps at which growth can occur via the layer mode. In this way the difficult initial step involving nucleation of growth on perfectly flat terraces can be avoided.

Layer-Plus-Island Growth Mode (Stranski–Krastanov). As the name suggests, this growth mode is intermediate between the island and the layer growth modes just described in that a strained monolayer (or several monolayers) of growth occurs first, with additional growth occurring in the form of islands nucleating on the growing film. As a result, there is a transition from two- to three-dimensional growth. This growth mode can apparently occur for a variety of reasons: for example, the first monolayer of

the growing film assumes the surface structure of the substrate, which is different from that of the bulk film. This is called *pseudomorphic growth*. In this case layer growth occurs initially when

$$E_{sA'}d + \sigma_{A'} + \sigma_{A'B} < \sigma_B, \tag{W21.2}$$

where A′ refers to the growing film, which is strained when it takes on the structure of the substrate. The term $E_{sA'}d$ represents the *elastic energy* per unit area associated with the strain in the growing film, with $E_{sA'}$ the *strain energy* per unit volume and d the film thickness. As d increases, the left-hand side of Eq. (W21.2) will eventually exceed the right-hand side at a certain *critical thickness*. When this occurs, either misfit dislocations will appear in the film to relieve the strain, as discussed in Section W20.2, or the island growth mode will take over. When island growth that is essentially unstrained takes over, it follows that $\sigma_A + \sigma_{AA'} > \sigma_{A'}$.

The critical nucleus size, ≈ 10 to 100 atoms, for the second, or island, phase of the Stranski–Krastanov growth mode is much larger than in the case of island (Volmer–Weber) growth, where typically a single atom is the critical nucleus. The need for a larger critical nucleus in the Stranski–Krastanov growth mode is likely due to the rather small preference for island growth over layer growth.

Examples of this growth mode include the growth of some metals on metals and on semiconductors [e.g., the Pb/W(110), Au/Mo(110), Ag/W(110), Ag/Si(111), and Ag/Ge(111) systems, among others]. The growth of Ge on Si(100) and Si(111) can also occur via this mode, with a uniformly strained Ge film initially growing to about three monolayers. This is followed by a transition to the growth of three-dimensional Ge nanocrystals on top of the initial strained Ge film, which is often called a *wetting layer*.

W21.3 Processing Using Ion Beams

Ions provide a versatile means for processing solids. They provide a directed source of energy that couples to the ions of a solid via collisions or via excitation of the electrons. Ions play a triple role in the processing of materials. First, an ion beam may be used to sputter material off the surface, thereby cleaning or etching it. Second, ion beams are used to implant ions into surfaces, such as dopants into semiconductors. Third, ion beams may be used to deposit material from another target onto the surface, a process known as *sputter deposition*.

In cleaning or etching via sputtering one generally employs relatively low-energy ions (1 to 10 keV) of an inert gas, such as Ar^+, to deposit energy in the surface region. A collision cascade results in which the ion energy is shared among many atoms, much as when a cue ball strikes an array of billiard balls. When the kinetic energy of an excited surface atom exceeds its binding energy, it will leave the solid. Atomic layers of the solid are thereby removed. The sputtering yield Y is the number of sputtered atoms per incident ion. This number is typically between 0.01 and 10 and depends on the energy of the beam and the material being sputtered.

In the ion-implantation process, a low-flux energetic ion beam (10 to 500 keV) penetrates the solid to a depth of ≈ 10 nm to ≈ 10 μm. For example, 200-keV As^+ ions penetrate 20 μm in Si before coming to rest. Some ions are able to penetrate much deeper if the direction of the beam is nearly parallel to a crystal axis through a process called *channeling*. Boron is used almost exclusively as an acceptor. Donor ions include Sb, As, and P. The ions slow down due to collisions with the nuclei and

the electrons and eventually come to rest some distance below the surface. There are a range of penetration depths that occur, with the net result that the solid is doped by the ions. Essentially, any element may be injected and the absolute concentration as well as the concentration profile may be controlled precisely. Since the technique is not thermodynamic in nature, it permits one to build up high concentrations of dopants, beyond the limits imposed by solubility constraints. By subsequent annealing, much of the radiation damage may be removed and the result can be a supersaturated solid solution of the dopant atoms in the host crystal. Precipitation or segregation may also occur. As the incident ion slows down by nuclear collisions, it leaves a trail of radiation damage in its wake. This consists largely of interstitial ions and vacancies. The concentration of displaced ions, N_d, is proportional to the fluence, ψ (the number of incident ions per square meter), and is given approximately by the formula

$$N_d = \frac{4000\psi F_d}{E_d}, \tag{W21.3}$$

where F_d is the energy deposition per unit length of penetration and E_d is the energy needed to displace an ion (10 to 25 eV). In some circumstances the radiation damage may be annealed out by elevating the temperature. In other cases it may be used to create amorphous material. Typical values of ψ are in the range 10^{16} to 10^{19} ions/m^2.

In the ion sputtering process, ion beams are directed at various target materials with different chemical compositions to create a vapor of varying chemical composition. Atoms or molecules from the vapor strike the substrate of interest and stick to it. For example, ion-beam deposition of highly tetrahedral amorphous C is produced with C ions of energy 10 to 100 eV. Layers as thin as a monolayer may be deposited on a substrate. Ion deposition is frequently used for metallization or for coating disks with magnetic material. In some cases the ion beam can assist in the deposition of a chemical vapor directly on the surface by activating the vapor of the material to be deposited.

The path of an incident ion as it penetrates the solid is a directed random walk. In characterizing the penetration of the ion beam, various moments of the distribution of final resting places are employed. Assuming the beam to be directed in the z direction, there is the mean projectile range or penetration depth

$$R_z = \langle z \rangle = \frac{1}{N}\sum_{n=1}^{N} z_n, \tag{W21.4}$$

where N is the number of ions striking the sample and z_n is the penetration depth of the nth ion. The mean radial displacement is given by

$$R_r = \langle \sqrt{x^2 + y^2} \rangle = \frac{1}{N}\sum_{n=1}^{N} \sqrt{x_n^2 + y_n^2}. \tag{W21.5}$$

Higher moments include the straggling distance,

$$\sigma_z = \sqrt{\langle (z - R_z)^2 \rangle} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (z_n - R_z)^2}, \tag{W21.6}$$

the radial straggling distance,

$$\sigma_r = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n^2 + y_n^2) - R_r^2}, \tag{W21.7}$$

and still higher statistical moments of the distribution, such as the skewness (asymmetry) and kurtosis (sharpness of falloff in the wings). Calculations of the spatial distribution, as well as the statistical moments, may be performed by resorting to numerical simulations in which a large number of trajectories is analyzed.

The physical parameters controlling the ion processes are the atomic numbers and masses of the projectile and target, Z_1, Z_2 and M_1, M_2, respectively, the Thomas–Fermi screening constant of the solid, k_{TF} (which curtails the long-range nature of the Coulomb interaction), the incident current, I_1, the beam area, A, and the kinetic energy of the projectile, E. Two energy loss processes are of importance, nuclear stopping and electronic stopping. In the nuclear-stopping process the projectile and target nuclei make a close collision, interacting via the screened Coulomb interaction. Energy and momenta are shared between the two nuclei. In the electronic-stopping process the electric field pulse of a passing projectile ion excites the electrons in the conduction band or upper valence band of the solid. Both interband and intraband excitations may occur. The gain of energy of the electrons is offset by the loss of energy of the projectile, so that energy is always conserved. By the energy-time uncertainty principle, the shorter the duration of the pulse, the wider is the spread of excitation energies. Thus $\Delta E \Delta t \approx h$ with $\Delta t \approx b/v$, where b is the impact parameter (perpendicular distance between the line of approach of the incident ion and the target nucleus) and v is the projectile speed. Hence electronic stopping is expected to dominate at high energies, where a wider range of excitation energy is available due to the shortness of the pulse.

In the nuclear-stopping process the incident ion is deflected from a target ion through an angle ϕ and therefore transfers an amount of energy T to the recoiling target nucleus, where

$$T = \frac{4M_1M_2E}{(M_1 + M_2)^2}\sin^2\frac{\phi}{2}. \tag{W21.8}$$

Maximum energy transfer for a given M_1 and M_2 occurs during backscattering, when $\phi = \pi$. Furthermore, when $M_1 = M_2$ there will be a maximum energy transfer for a given ϕ.

In discussing the energy-loss processes it is convenient to introduce a dimensionless energy, ϵ, defined as the ratio of an effective Bohr radius to the distance of closest approach in a head-on Coulomb collision. The effective Bohr radius is given empirically by $a \approx 0.8854a_1(Z_1^{2/3} + Z_2^{2/3})^{-1/2}$, where a_1 is the Bohr radius, 0.0529 nm. The distance of closest approach is $r_0 = e^2Z_1Z_2(M_1 + M_2)/4\pi\epsilon_0EM_2$. The dimensionless energy is

$$\epsilon = E(\mathrm{keV}) \times \frac{32.53M_2}{(M_1 + M_2)Z_1Z_2\sqrt{Z_1^{2/3} + Z_2^{2/3}}}. \tag{W21.9}$$

A comparison of the nuclear and electronic-stopping powers, $d\epsilon/d\rho$, is given in Fig. W21.6. The scaled penetration distance ρ is the distance in units of a, the effective Bohr radius. The nuclear and electronic stopping powers become equal at some energy.

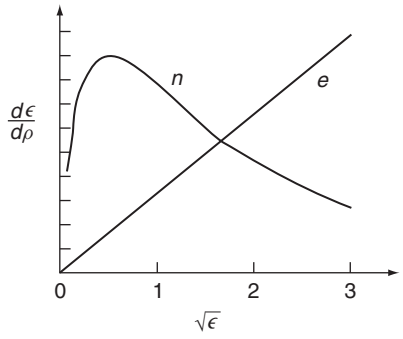

Figure W21.6. Stopping power for nuclear (n) and electronic (e) processes as a function of the parameter ϵ. In Si $\epsilon = 1$ corresponds to $E = 9$ keV for ^{11}Be ions or $E = 1.5$ MeV for Bi ions. (Adapted from J. A. Davies, *Mater. Res. Soc. Bull.*, **17**(6), 26 (1992).

For As, B, and P in Si, this energy is 700, 10, and 130 keV, respectively. Sputtering processes generally occur in the realm $\epsilon < 10$. For $Z_1 > Z_2$ the electronic stopping power is given approximately by the formula $(d\epsilon/d\rho)_e = 0.15\sqrt{\epsilon}$.

The mean projectile range is given by

$$R_z = a \int_0^{\epsilon_{\text{in}}} \frac{1}{(d\epsilon/d\rho)_e + (d\epsilon/d\rho)_n} d\epsilon, \tag{W21.10}$$

where ϵ_{in} corresponds to the incident energy E. An approximate formula for the mean range is

$$R_z(\text{nm}) = E(\text{keV}) \times 13{,}000 \frac{1 + M_2/M_1}{\rho_s Z_1^{1/3}}, \tag{W21.11}$$

with ρ_s being the mass density of the solid (in kg/m^3). The straggling in average total path length R is given approximately for small ϵ by

$$\frac{\Delta R}{R} = 0.7 \frac{\sqrt{M_1 M_2}}{M_1 + M_2}. \tag{W21.12}$$

In reactive-ion etching (RIE) the surface of a solid is exposed to a chemical etchant in the presence of an ion beam. The ion beam serves to excite the reactants, thereby enhancing the chemical reaction rate. The system behaves as if its temperature were elevated. Examples include the etching of Si by F_2, Cl_2, or Br_2 in the presence of an Ar^+ beam. The ion beam also serves to create steps on the surface with dangling bonds available for chemical reaction.

Recently, it has been shown that ion implantation, combined with annealing and recrystallization, can be used to fabricate semiconductor nanocrystals.[†] Alumina substrates were bombarded with semiconductor ion doses up to 10^{21} ions/m^2. If the substrate is kept at a high temperature during bombardment, then cooled and annealed at a relatively low temperature, the substrate retains the α-alumina structure and the

[†] J.D. Budal et al., *Nature*, **390**, 384 (1997).

semiconductor nanocrystals that precipitate align themselves relative to the substrate. If the substrate is bombarded at low temperatures with a high dose of ions, the substrate is amorphized. A low-temperature anneal then leads to the substrate forming γ-alumina. This leads to a different orientation of the nanocrystals than above.

Ion implantation may be combined with etching to produce thin slices of crystals in a technique called *ion slicing*. He^{2+} ions, with an energy of ≈ 4 MeV, impinge on a crystal. The implanted ions deposit a high percentage of their energy near the penetration depth (≈ 10 μm), creating a damage layer. This layer may be attacked with an etching solution and the resulting crystal slice may be delaminated from the rest of the crystal. Subsequently, it could be placed on the surface of a different crystal. This circumvents the need for epitaxial growth of thin films and extends the ability to obtain films on substrates to cases where epitaxial growth may not be possible.

W21.4 Float-Zone Purification of Single-Crystal Si

The purest single crystals of Si are currently grown from the liquid phase using a method in which the molten Si is not in contact with any container, thereby eliminating the main source of impurities. This is the *float-zone* (FZ) method, illustrated schematically in Fig. W21.7, and is a type of zone refining. The starting material is a cylindrical rod of pure, polycrystalline Si which is mounted vertically and held at both ends, either under vacuum or in an inert atmosphere. In this method only a short section of the Si rod away from the ends is molten at any given time. The molten section is heated via radio-frequency induction using a coil surrounding the container and is held in place by surface tension forces. To initiate the growth of a single crystal, a small single-crystal Si seed is placed in contact with the molten end of the rod. A necking process similar to that used in the CZ growth method, described in Chapter 21, is then used to remove any dislocations from the growing crystal.

The external heating coil and the molten Si zone are moved slowly along the Si rod several times in the same direction until the desired purity and crystallinity are obtained. Rotation of the cylindrical rod is also used in this method, to promote cylindrical uniformity of the material. Single crystals of FZ Si of up to 15 cm in diameter

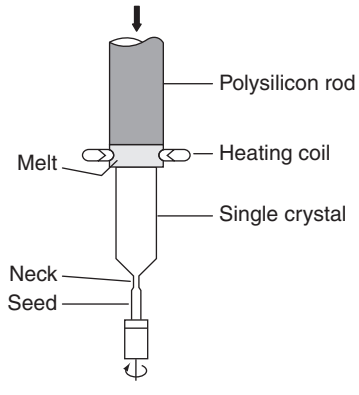

Figure W21.7. Float-zone method used for the growth of extremely pure single crystals of Si and other materials.

can be grown and purified by this technique. The use of FZ Si in Si microelectronic devices is limited due to its low oxygen content, $\approx 10^{22}$ atoms/m^3, a factor of 100 less than in CZ Si. As a result, the beneficial effects of internal gettering and of mechanical strengthening due to oxygen precipitation are not available in FZ Si.

The attainment of extremely high purities in the single-crystal Si rod, corresponding to impurity fractions of $\approx 10^{-10}$ (i.e., 99.99999999% pure Si), results from the much lower solubility of most atoms in solid Si than in liquid Si. This difference in solubility is due to the much more restrictive conditions for the bonding of atoms in solid Si as compared to liquid Si and is expressed in terms of the *equilibrium distribution* or *segregation coefficient* K_{A} for a given atom A. The coefficient K_{A} is the ratio of the equilibrium concentrations of atom A in the two phases:

$$K_{\mathrm{A}} = \frac{c_{\mathrm{A}}(\text{solid})}{c_{\mathrm{A}}(\text{liquid})}. \tag{W21.13}$$

If the fractional concentrations c_{A}(solid) and c_{A}(liquid) are both $\ll 1$, K_{A} is also given by the ratio of the thermodynamic activities of atom A in the two phases. The coefficient K_{A} can be determined experimentally from the equilibrium phase diagram for the Si–A system. If the liquidus and solidus curves are nearly straight lines for low concentrations of A in Si and have negative slopes s_L and s_S, respectively (Fig. W21.8),

$$K_{\mathrm{A}} = \frac{s_L}{s_S} < 1. \tag{W21.14}$$

Solutes that depress the melting temperature of Si have $K_{\mathrm{A}} < 1$, while those that raise T_m have $K_{\mathrm{A}} > 1$.

The distribution coefficient K_{A} for dilute concentrations of A atoms in a solid such as Si can be related to the enthalpy change ΔH_m associated with melting of the solid and to the change of T_m as a function of the A-atom concentration in the solid. The appropriate expression, obtained by equating the chemical potentials of A atoms in the

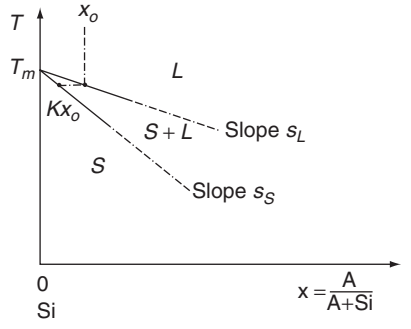

Figure W21.8. Equilibrium phase diagram for the Si–A system. The liquidus and solidus curves are nearly straight lines for low A-atom concentrations and have negative slopes s_L and s_S, respectively.

liquid and solid phases,[†] is

$$K_{\mathrm{A}} = 1 + \frac{\Delta H_{m0}}{RT_{m0}^2} \frac{T_m - T_{m0}}{c_{\mathrm{A}}(\text{liquid})}. \tag{W21.15}$$

Here ΔH_{m0} and T_{m0} correspond to pure Si. For dilute solutions [i.e., c_{A}(liquid) and c_{A}(solid) both $\ll 1$], the ratio $(T_m - T_{m0})/c_{\mathrm{A}}$(liquid) is essentially independent of temperature and so, therefore, is K_{A}. It can be seen from Eq. (W21.15) that, as stated earlier, $K_A < 1$ when $\Delta T_m = T_m - T_{m0}$ is negative, and vice versa.

To illustrate the connection between distribution coefficients and phase diagrams, consider the case of solid-solution Si–Ge alloys whose phase diagram is shown in Fig. W21.9. The distribution coefficients for Ge in Si, K_{Ge}(Si), and for Si in Ge, K_{Si}(Ge), can be obtained from this diagram using the slopes s_L and s_S as the concentrations of Ge and Si tend to zero. The following results are obtained:

$$K_{\mathrm{Ge}}(\mathrm{Si}) \approx 0.3 \quad \text{and} \quad K_{\mathrm{Si}}(\mathrm{Ge}) \approx 5.5. \tag{W21.16}$$

Thus Si atoms have a greater tendency than Ge atoms to enter the solid phase in Si–Ge alloys and actually prefer the solid phase to the liquid phase. The solid phase in equilibrium Si–Ge alloys will therefore always be enriched in Si relative to the liquid phase, as indicated in Fig. W21.9. This follows from the fact that the melting temperature of Si, $T_m = 1414°\mathrm{C}$, is greater than that of Ge, $T_m = 938°\mathrm{C}$. As discussed in Chapter 6, this behavior is also observed for solid-solution Cu–Ni alloys, which are always Ni-rich in the solid phase, Ni having the higher melting point.

Values of c_{A}(solid) obtained experimentally can deviate from those expected from the equilibrium value of K_{A} when the growth process deviates from equilibrium conditions. As an example, K_{A} is observed to depend on the growth rate. It is reasonable to expect that $K_{\mathrm{A}} \to 1$ as the growth rate approaches infinity since A atoms at the growth interface will be trapped in the solid phase due to lack of time to diffuse away.

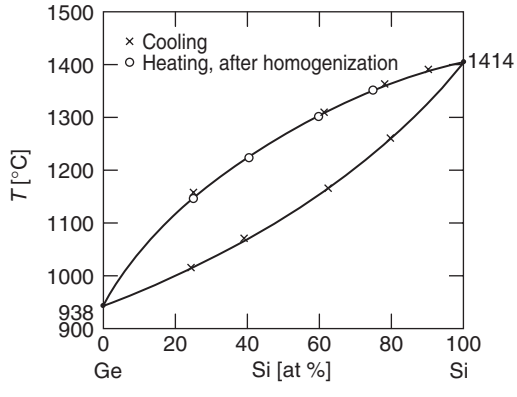

Figure W21.9. Equilibrium phase diagram for solid-solution Si–Ge alloys. (Adapted from M. Hansen, *Constitution of Binary Alloys*, McGraw-Hill, New York, 1958.)

[†] P. Gordon, *Principles of Phase Diagrams in Material Systems*, McGraw-Hill, New York, 1968, p. 140.

TABLE W21.3 Distribution Coefficients K of Elements in Si Near $T_m = 1414°C$

Column III K		Column IV K		Column V K		Column VI K	
B	0.8	C	0.07	N[a]	$<10^{-7}$	O	0.5
Al	0.002	Si	1	P	0.35		
Ga	0.008	Ge	0.3	As	0.3		
In	0.0004	Sn	0.016	Sb	0.023		

Source: Most values are from F. A. Trumbore, *Bell Syst. Tech. J.*, **39**, 221 (1960).

[a]The value for N is uncertain.

In the FZ method if a given dilute impurity with distribution coefficient $K < 1$ has an initial concentration c_0 in the solid Si rod, the first portion of the Si rod that is melted and then allowed to resolidify slowly will have the lower impurity concentration $Kc_0 < c_0$. The same level of purification will not, however, be achieved in the rest of the Si rod since the concentration of the impurity in the molten zone will slowly increase above c_0. The impurity concentration in the first segment of the Si rod will therefore be reduced by the factor K each time the molten zone is passed slowly through it. Since typically $K \ll 1$ for many unwanted impurities, an extremely low concentration $c \approx K^n c_0$ can in principle be achieved in the first segment of the Si rod after n passes of the molten zone. The opposite end of the Si rod in which the impurities have become concentrated is cut off after the purification process is completed. Since the impurity concentration, while low, will still be nonuniform along the length of the Si rod, homogenizing treatments that involve passing the molten zone repeatedly along the rod in both directions are employed to obtain a uniform impurity concentration.

Values of the equilibrium distribution coefficients for several elements in Si are given in Table W21.3. The only elements with distribution coefficients in solid Si which are greater than 0.05 are from groups III, IV, V, and VI of the periodic table (e.g., B, C, Ge, P, As, and O). The elements B, P, and As are substitutional impurity atoms which are often used for doping Si. Unwanted metallic impurities such as Cu, Au, and Zn have very low values of $K \approx 10^{-7}$ to 10^{-4}. The coefficient K is observed to be temperature dependent, falling rapidly with decreasing T.

In addition to its use for Si, the FZ technique remains the preferred method for obtaining highly purified crystals of a wide variety of semiconducting, metallic, and ceramic materials, including single crystals of the high-T_c superconductor La–Sr–Cu–O.

W21.5 Epitaxial Growth of Single-Crystal Si Layers via CVD

The *homoepitaxial* growth of single-crystal layers (*epilayers*) of Si on Si substrates as carried out via *chemical vapor deposition* (CVD) is the preferred method of growth for the layers used in the fabrication of Si-based electronic devices. The CVD of Si employs a wide variety of deposition systems and conditions and so is a very versatile growth procedure. The CVD process involves the thermal decomposition (pyrolysis) of gaseous precursor molecules, with both vapor-phase (*homogeneous*) and surface (*heterogeneous*) chemical reactions playing important roles. It is desirable, in general, to suppress vapor-phase chemistry to avoid powder formation and the defects that

would result from particle incorporation in the films. The Si epitaxial layers deposited undergo further processing when used in Si-based electronic devices. These additional processing steps are discussed in Section W21.8, where the fabrication of Si-based integrated circuits is described.

The growth of Si from the vapor phase at substrate temperatures in the range $T_s = 500$ to 1150°C has several advantages relative to the Czochralski and float-zone methods, which involve growth from the melt at $T_m = 1414$°C. The advantages include reduced diffusion of both dopant and unwanted impurity atoms and reduced thermal stresses in the film and substrate. Reduced dopant diffusion allows the fabrication of abrupt interfaces between regions of different doping levels, an important factor in the development of smaller and faster devices.

The single-crystal Si wafers used as substrates for the epitaxial growth of Si layers are grown via the Czochralski method and are required to be as defect-free as possible since dislocations and other structural defects present in the substrate can propagate into the growing film. The surface of the substrate must also be smooth and clean (i.e., free from impurities such as carbon and oxygen), to prevent the nucleation of stacking faults and the appearance of other defects, such as dislocations, voids, inclusions, and precipitates in the growing film. There exist well-developed polishing and cleaning procedures, both ex situ and in situ, for the preparation of Si wafers for use as substrates. Ex situ chemical cleaning, which results in an air-stable, oxide-free Si surface, involves an H_2O_2-based chemical cleaning procedure, the *RCA clean*,[†] followed by a 10-s dip in a 10:1 H_2O:HF solution. This treatment generates a hydrophobic Si surface which is chemically stabilized by a surface layer of strong Si–H bonds. In situ cleaning methods include high-temperature treatments, often in H_2, to remove any SiO_2 present on the surface as volatile SiO molecules and also to remove C from the surface via its diffusion into the bulk or by the evaporation of the surface layer of Si.

A typical cold-wall Si CVD system is shown in Fig. W21.10. It consists of a water-cooled fused-quartz tube surrounded by radio-frequency heating coils into which the Si wafer substrates are placed in a susceptor made of graphite, SiC-coated graphite, or quartz. The deposition can be carried out at atmospheric pressure (APCVD) or at reduced pressures (RPCVD), $P \approx 0.01$ to 0.1 atm. The current standard epitaxial growth method is RPCVD, which has the advantage of minimizing autodoping (i.e., the doping of the growing Si layer by dopant atoms originating from the Si substrate).

Film growth from the vapor phase is a very general method of materials synthesis and typically involves the following steps, each of which may in fact represent a complicated sequence of more elementary steps:

1. Transport of gaseous species from the source to the substrate
2. Adsorption onto the substrate surface
3. Nucleation and growth of the film
4. Removal from the surface of unwanted species that might interfere with film growth

The nucleation and growth steps are described in Section W21.2. The thermal decomposition of the gaseous species can occur either in the vapor phase or on the

[†] W. Kern and D. A. Puotinen, *RCA Rev.*, **31**, 187 (1970).

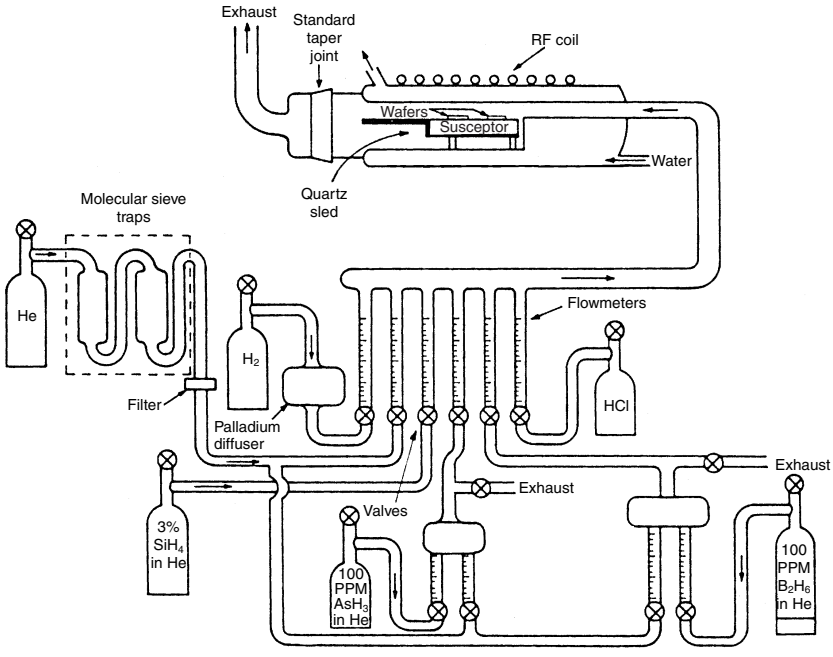

Figure W21.10. Typical cold-wall Si CVD system. (From D Richman et al., RCA Review, **31**, 613 (1970).)

heated substrate surface. The hydrodynamics of the flowing gases in the CVD system can have a significant influence on the growth process.

In the case of Si CVD, there are many possible choices for the molecular precursors, including SiH_4 and $SiHCl_3$. The important growth species present on the surface are then the highly reactive radicals silylene, SiH_2, and $SiCl_2$. These radicals are the products of the thermal decomposition of the feedstock gases and will undergo further reactions on the surface of the growing film. Carrier gases such as H_2 and He are often used to aid in the transport of vapor species to the substrate. The concentrations of atoms, radicals, and molecules adsorbed on the growing surface are controlled by their incident fluxes (i.e., by their partial pressures in the vapor phase) and by the substrate temperature T_s which controls their desorption rates.

Typical net chemical reactions resulting in the growth of the Si epilayer include the following:

$$\begin{aligned} SiH_4(g) &\longleftrightarrow Si(s) + 2H_2(g), \\ 2SiHCl_3 &\longleftrightarrow 2Si(s) + 3Cl_2(g) + H_2(g). \end{aligned} \tag{W21.17}$$

These reactions actually represent a series of elementary steps taking place in the vapor phase and on the substrate surface. Growth rates are ≈ 1 μm/ min at $T_s \approx 1100^\circ$C and decrease rapidly as T_s is lowered (see Fig. 21.3). Homogeneous vapor-phase reactions leading to the formation of disilane Si_2H_6 are

$$\begin{aligned} SiH_4(g) &\longrightarrow SiH_2(g) + H_2(g), \\ SiH_4(g) + SiH_2(g) &\longrightarrow Si_2H_6(g). \end{aligned} \tag{W21.18}$$

These reactions can ultimately lead to the formation of undesirable polymeric silicon hydride powder, $(SiH_2)_n$.

The partial pressures of the vapor species involved in growth must exceed their equilibrium vapor pressures with respect to the Si surface at T_s in order for the net deposition of a film to occur. The growth species must therefore be supersaturated in the vapor phase, with the *supersaturation ratio* SSR for the case of Si(g) atoms defined by

$$\mathrm{SSR}(\mathrm{Si}(g), T_s) = \frac{P(\mathrm{Si}(g))}{P_{\mathrm{eq}}(\mathrm{Si}(g), T_s)}, \tag{W21.19}$$

where $P(\mathrm{Si}(g))$ is the actual vapor pressure of Si(g) just above the substrate surface and $P_{\mathrm{eq}}(\mathrm{Si}(g), T_s)$ is the equilibrium vapor pressure of Si(g) with respect to pure Si(s).

A wide variety of investigations have allowed the following conclusions to be reached concerning the growth of Si epilayers via CVD:

1. The rate-controlling step for the growth of Si is either the removal from the surface of hydrogen in Si–H bonds via the desorption of H_2, or the dissociation of SiH_2 or $SiCl_2$ on the surface.
2. The rate-controlling step for obtaining high crystallinity in the Si epilayer is the diffusion of Si on the growing surface.
3. Lattice defects are generated when the Si adsorption rate exceeds the rate at which Si can diffuse on the surface and be incorporated into the growing film. Si atoms then enter nonideal, higher-energy bonding configurations.
4. Si atoms compete with other species on the surface, such as dopant atoms or molecules and hydrogen, oxygen, or carbon atoms, for the available bonding sites to Si substrate atoms, thereby limiting the Si atom diffusion rate.

The termination of the growing Si surface by hydrogen in Si–H bonds can play a critical role in the CVD of Si by inhibiting epitaxial growth through the blocking of surface sites for the adsorption of reactive species such as SiH_2 and SiH_3. This is particularly important at T_s less than about 400 to 500°C.

Recently, the CVD of Si and of Si–Ge alloys has been combined with UHV techniques to achieve a very high level of system and substrate cleanliness (e.g., the elimination of oxygen and carbon surface impurities). The use of this growth method, known as *UHV/CVD*, allows the deposition of epitaxial Si and Si–Ge layers at much lower pressures, $P \approx 10^{-3}$ torr, and lower T_s, $\approx$ 500 to 550°C, than are ordinarily used. Operation at lower pressures has several advantages: the undesirable homogeneous pyrolysis of precursors in the vapor phase is minimized, the very low partial pressures of O_2 and H_2O necessary for the maintenance of an active, SiO_2-free Si surface are more readily achieved,[†] and molecular flow conditions are obtained, with the result that recirculating flows, eddy currents, and turbulence are avoided. Due to the clean and hydrogen-stabilized surfaces of the Si wafers when they are placed into

[†] For experimental results and discussions of the interactions of O_2 and H_2O with Si at high temperatures, see F. W. Smith and G. Ghidini, *J. Electrochem. Soc.*, **129**, 1300 (1982); G. Ghidini and F. W. Smith, *J. Electrochem. Soc.*, **131**, 2924 (1984).

the UHV/CVD system, no further in situ treatment at high temperatures is required to prepare the Si surface for epitaxial growth.

The use of lower substrate temperatures reduces problems associated with dopant atom redistribution via diffusion and also is a very effective method of reducing defect concentrations in the films. Growth at lower T_s will reduce the equilibrium concentrations of defects such as vacancies and will also reduce the mobility of point defects and hence their tendency to interact with each other to form extended defects. In addition, thermal stresses which can also lead to the generation of defects in the film will be reduced at lower T_s. Better film thickness uniformity is also expected at lower T_s since the deposition process changes from one controlled by vapor-phase transport at higher T_s to one controlled by surface reactions at lower T_s, as discussed in Section 21.3. It is still necessary to maintain T_s well above the range in which the film will become noncrystalline or amorphous.

Nonequilibrium structures and alloys can also be prepared at low T_s. These include strained Si–Ge epilayers grown on Si with thicknesses well above the critical values for the generation of misfit dislocations and also alloys of Si with concentrations of dopant atoms such as B which are several orders of magnitude above equilibrium concentrations. Sharp transitions, particularly in dopant profiles, between the substrate and the epilayer are essential as device dimensions continue to shrink. Both the layer growth rate and dopant diffusion rates decrease exponentially as T_s decreases. Since the activation energy for diffusion, $E_a(\text{diff}) \approx 3.5$ eV, is much greater than that for growth, $E_a(\text{growth}) \approx 1.5$ eV, reasonable growth rates, ≈ 0.1 to 10 nm/min, can still be obtained at $T_s \approx 500^\circ$C, where dopant diffusion has been effectively frozen out.

A schematic of the hot-wall apparatus used in the UHV/CVD method is shown in Fig. W21.11. The carefully cleaned Si wafers have surfaces passivated by H termination (i.e., Si–H bonds), which can be thermally desorbed from the Si surface at $T_s > 400^\circ$C. In the UHV/CVD of Si the vapor phase consists entirely of SiH_4.

Films that are "defect-free" (i.e., with defect densities less than ≈ 100 cm^{-2}) are readily achieved via CVD. The most sensitive quantitative method of determining

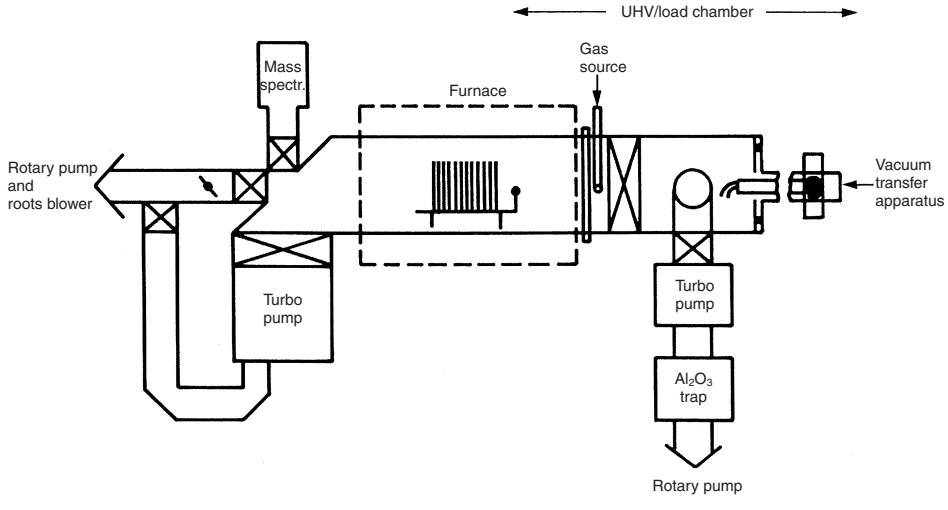

Figure W21.11. UHV/CVD system. (From B. S. Meyerson, *Appl. Phys. Lett.*, **48**, 797 (1987). Copyright 1987 by the American Institute of Physics.)

densities of structural defects such as dislocations in Si epitaxial layers is by means of chemical etching. Since the disordered regions of the lattice containing defects are in a state of higher energy, they are more rapidly attacked (i.e., etched) by appropriate acids. Optical microscopy can then be used to count the etch pits and also to identify the nature of the defects from the shape of the etch pit. Transmission electron microscopy (TEM) is the preferred method for probing the atomic perfection of the interface between the substrate and the epilayer. Electrically active defects such as impurity-related traps are not readily detected via etching or TEM. Their presence can be determined by the effects that they have on devices such as diodes, transistors, or metal–oxide–semiconductor (MOS) capacitors, which are fabricated from the Si epilayers.

Metallic elements such as Fe and other transition metals are undesirable impurities in Si due to the fact that they act as traps (i.e., as centers for the recombination of electrons and holes). Although they do not enter into CZ or FZ Si from the melt due to their very low distribution coefficients, they will diffuse rapidly into the bulk at elevated temperatures if they can reach the surface of the Si crystal through the vapor phase.

Other recent approaches to Si epitaxy via CVD include the use of intermediate layers such as cubic CaF_2, fluorite, whose lattice constant, $a = 0.546$ nm, matches that of Si, $a = 0.543$ nm, to within 0.6% at $T = 300$ K. The CaF_2 layer is deposited epitaxially onto the Si(100) surface first, followed by the deposition of the Si epilayer onto the CaF_2 layer. The top Si epilayer is then removed for further processing by dissolving the intermediate CaF_2 layer in an appropriate solvent. In this way the original Si(100) substrate can be reused.

A recent approach to understanding the growth of Si epilayers at low temperatures has involved the definition of a limiting *epitaxial thickness* h_{epi} above which the deposited films become amorphous. This is in contrast to the usual definition of a minimum *epitaxial temperature* T_{epi}, below which epitaxy is impossible, due to insufficient surface diffusion of atoms adsorbed on the surface. Epitaxial growth of Si can be observed in a very clean MBE system at all temperatures between $T = 50$ and 300°C, but only up to the thickness h_{epi}, which increases exponentially with increasing T and decreases with increasing growth rate. For Si films grown via MBE, h_{epi} was found to be 1 to 3 nm at room temperature. The transition from crystalline to amorphous growth at h_{epi} has been attributed to a surface-roughening effect, with the accumulation at the growing surface of impurity atoms such as hydrogen playing a major role in the roughening process.

W21.6 Molecular-Beam Epitaxial Growth of GaAs

The growth via *molecular-beam epitaxy* (MBE) of films of the group III–V semiconductor GaAs, as well as of other III–V and II–VI semiconductors, has many features in common with the CVD of epitaxial Si layers, including the steps of transport and adsorption of the appropriate precursor vapor species onto the substrate surface, nucleation and growth of the film, and removal of unwanted species from the substrate surface. In MBE molecular beams (i.e., beams of neutral molecules or atoms) are directed onto a heated substrate in a UHV system. Due to the low particle density of the beam and also to the very low background pressure in the growth chamber, the particles in the beam do not interact with each other and undergo essentially no collisions

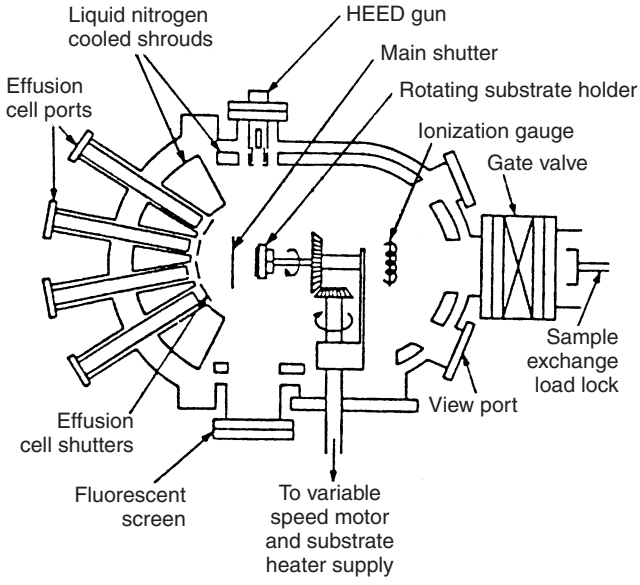

Figure W21.12. Typical MBE vacuum chamber. (Reprinted from A. Y. Cho, *Thin Solid Films*, **100**, 291 (1983), copyright 1983, with permission from Elsevier Science.)

with residual gas molecules on their path from the source to the substrate. A typical MBE growth chamber is shown schematically in Fig. W21.12. Along with the vacuum chamber and all the associated accessories, appropriate vacuum pumps and electronics for the control of the various components are required. The mass spectrometer is used for residual gas analysis. It can also be used to measure the fluxes of reactant species and can provide signals to be used for adjusting the effusion cell temperatures so that constant fluxes, and hence constant deposition rates, can be maintained.

Advances in UHV technology[†] have permitted the deposition via MBE of films at relatively low T_s with unparalleled control of composition, purity, and interface sharpness, involving literally atomic layer-by-layer growth. The low growth temperature has the advantage of reducing undesirable thermally activated processes such as diffusion, while the low growth rates ($\approx$ 10 nm/ min) offer the advantage of accurate control of film thickness. The UHV conditions employed in MBE also permit in situ monitoring of the film structure and thickness using high-energy electron beams reflected at very low angles from the surface of the growing film. This technique is known as *reflection high-energy electron diffraction* (RHEED). The chemical purity and composition of the substrate and of the film can also be monitored in situ using Auger electron spectroscopy (AES). Finally, the use of modulated-beam mass spectrometry (MBMS) employing separate beams of Ga and As_2 has allowed the detailed study of surface processes involved in the growth of GaAs via MBE.

The solids that are the source materials for the MBE of GaAs are contained in heated effusion cells within the vacuum chamber. Elemental Ga metal is used for the Ga flux, while solid GaAs is used for As_2 and solid elemental As for As_4. Additional elements

[†] See Weissler and Carlson (1979) for a useful description of UHV techniques.

used for doping, alloying, and for multilayer or junction depositions are contained in their own effusion cells. The nature and flux of the vapor species from each effusion cell are controlled by the temperature of the cell, with the flux directed through a small orifice in the wall of the cell toward the substrate. Shutters placed between each cell and the substrate are used to block individual beams when control of the composition or thickness of the growing film is desired. The substrates are mounted on heated holders whose temperature T_s can be controlled accurately by regulated internal heaters. The substrate holders can be rotated during growth in order to obtain extremely uniform epitaxial films.

Due to the very low background pressure in the MBE chamber during growth, $P \approx 10^{-9}$ torr ($\approx 10^{-7}$ Pa), very few unwanted residual gas molecules are incident on the substrate and incorporated into the films. Due to the cleanliness of the growth chamber, growth rates can be very low, 6 to 60 nm/min, which allows extremely thin layers with abrupt interfaces to be grown on surfaces that are essentially atomically smooth. Typical beam fluxes can be in the range 10^{11} to 10^{16} atoms (or molecules)/cm^2·s.

The substrates used for GaAs integrated-circuit fabrication are semi-insulating bulk GaAs crystals grown via the *liquid-encapsulated* Czochralski method. These undoped substrates typically contain 10^4 to 10^5 dislocations/cm^2. Before being placed in the growth chamber the substrates undergo a variety of polishing, etching, and rinsing procedures which are chosen carefully for each type of substrate. Further treatment of the substrate within the growth chamber is also possible and typically involves heating to about $T = 580^\circ$C to remove oxygen, followed by Ar ion bombardment to remove the less volatile carbon contamination. To obtain extremely clean growth surfaces, undoped epitaxial layers of GaAs are often grown in the MBE growth chamber on existing bulk substrates.

Stoichiometric GaAs films are typically grown in the range $T_s = 500$ to 600°C under an incident vapor flux that is enriched in As-containing species due to the instability of the heated GaAs surface with respect to the preferential loss of more volatile arsenic species. When As_2 is incident, stoichiometric GaAs films are obtained as long as the As_2 flux exceeds 50% of the Ga flux [i.e., as long as $R(As_2)/R(Ga) > 0.5$]. The sticking coefficient of Ga is equal to unity for T_s less than about 480°C and then decreases exponentially with an activation energy of $E_a \approx 2.5$ eV at higher temperatures. Under proper growth conditions any excess arsenic beyond that needed for stoichiometric growth is desorbed from the surface of the growing film. This is attributed to a high sticking coefficient for As_2 on a Ga-terminated surface and a low sticking coefficient for As_2 on an As-terminated surface, as observed experimentally. As a result, the growth rate of GaAs, which is controlled by the incident monoatomic Ga flux, can also be limited kinetically by the desorption of As-containing species that block sites for the incorporation of Ga atoms.

The GaAs growth process from Ga and As_2 has been shown by sensitive MBMS and RHEED studies to be limited by the first-order dissociative chemisorption of As_2 molecules when they encounter pairs of vacant As sites next to filled Ga sites. Growth of GaAs from Ga and As_4 has been shown to be more complicated, involving the dissociation of pairs of As_4 molecules on adjacent Ga atoms. Four of the resulting eight As atoms are incorporated into the growing film while the remaining four desorb as As_4. The doping of GaAs films for high-frequency and light-emitting device applications occurs during growth and is controlled by a variety of thermodynamic and kinetic

effects. For example, a dopant element such as Cd or Zn with a high vapor pressure can desorb from the growing surface and so may not be incorporated.

For a given substrate material there is a well-defined temperature range for the growth of high-quality epitaxial films. For example, MBE of GaAs is typically carried out for T_s between 500 and 600°C. The low-T_s limit is related to decreasing crystallinity, while the high-T_s limit is due to the high vapor pressure of As_2 and the resulting deviations from stoichiometry. The lower limit for T_s can be extended down to 200 to 300°C by using reduced arsenic fluxes, and the upper limit can be extended up to 700°C with the use of higher arsenic fluxes. Films deposited at $T_s = 700$°C are of higher quality (e.g., purer), due to reduced incorporation of impurities such as oxygen, which form volatile molecules that desorb from the growth surface at high T_s.

MBE systems are usually dedicated to the deposition of specific materials [e.g., either group III–V (GaAs, GaP, InP, etc.) or II–VI (ZnSe, CdTe, etc.) compound semiconductors]. For each group of materials the compositions and configurations of the films or superlattices deposited is essentially unlimited, with the only constraint being the imagination of the grower. MBE is a versatile deposition technique which, in addition to being used for group III–V and II–VI semiconductors, has also been used for the deposition of elemental semiconductors such as Si and Ge, for metals such as α-Fe, Co, and Al, and insulating layers such as CaF_2.

Other techniques used for the deposition of compound semiconductor thin films includes *metal–organic* CVD (MOCVD), metal–organic MBE (MOMBE), also known as *chemical beam epitaxy* (CBE), which make use of volatile organometallic compounds such as trimethyl gallium, $(CH_3)_3Ga$. When arsine, AsH_3, is used as the source of As, a typical reaction leading to the growth of GaAs is $(CH_3)_3Ga + AsH_3 \rightarrow GaAs + 3CH_4$.

W21.7 Plasma-Enhanced CVD of Amorphous Semiconductors

The use of energetic radio-frequency (RF) and microwave plasmas to produce highly-reactive chemical species (excited atoms, molecules, radicals, and ions) allows deposition of a wide variety of semiconducting and insulating thin films onto practically any substrate at low temperatures, typically in the range $T_s = 25$ to 500°C. Important advantages of this *plasma-enhanced* CVD (PECVD) method are that high-temperature materials such as oxides, nitrides, and carbides can be deposited without excessive heating of the substrate and also that large-area substrates can be coated. Low-temperature deposition is important because lower temperatures are required in integrated-circuit fabrication, due to the need to avoid diffusion of dopant atoms and due to the presence of the low-melting-point metal Al used for device interconnections. As a result of the lower T_s, the films deposited are usually *amorphous* and also often highly nonstoichiometric, with significant deviations from the nominal SiO_2, Si_3N_4, and SiC compositions in the case of Si-based films. Depending on the precursors employed and the substrate temperature, the films also can contain up to ≈ 40 at % hydrogen, which is chemically bonded in the random covalent network.

Despite the absence of long-range order, a considerable degree of short-range chemical order, corresponding to the strongest possible set of chemical bonds, is usually present in these films. This type of bonding results from the good atomic mixing taking place at the surface of the growing film as a result of energetic species (e.g., ions) incident from the plasma. This atomic mixing allows bonding configurations to be achieved

which correspond to a state of low enthalpy. The Gibbs free energy $G = H - TS$ for these amorphous films results from competition between achieving the lowest-possible enthalpy H, corresponding to the strongest set of chemical bonds in the network, and achieving the highest possible entropy S, corresponding to random bonding between the atoms in the network. A *free-energy model* for the bonding in amorphous covalent networks has been formulated which takes into account the effects of both enthalpy and entropy.[†]

Interesting and important examples of amorphous films deposited by PECVD include hydrogenated amorphous Si (i.e., a-Si:H), amorphous silicon oxide, nitride, and carbide (i.e. a-SiO_x:H, a-SiN_x:H, and a-SiC_x:H), and amorphous or diamond-like carbon (DLC) (i.e., a-C:H). One of the important advantages of the PECVD method is that films with a wide range of compositions can be deposited due to the wide variety of available gas-phase precursors and to the considerable range of deposition parameters such as T_s, discharge pressure and power, and substrate bias potential, which controls the bombardment of the film by ions. As a result, film properties such as the optical energy gap and the electrical conductivity at room temperature can be varied over wide ranges [e.g., between ≈ 0 and 5 eV and between 10^{-14} and $10^{-2}(\Omega\cdot\text{m})^{-1}$, respectively]. Available gaseous precursors include SiH_4, O_2, H_2O, NH_3, and hydrocarbons such as CH_4 and C_2H_2. Other precursors, such as borazine ($B_3N_3H_6$) and tetraethoxysilane [TEOS, $Si(OC_2H_5)_4$], can be generated from liquids. Gases such as diborane (B_2H_6) and phosphine (PH_3) can be added directly to the discharge when doping of the deposited layer (e.g., a-Si:H) is desired. Precursors that are typically used in the PECVD of thin films are listed in Table W21.4.

PECVD films have a wide range of semiconducting, dielectric, and protective-coating applications. Examples include n- and p-type a-Si:H in photovoltaic solar cells and thin-film transistors (TFTs), a-SiO_x:H as a dielectric layer and a-SiN_x:H as an encapsulating layer in semiconductor devices, p-type a-SiC_x:H as a window layer in a-Si:H solar cells, and a-C:H as a protective coating for magnetic-recording media, and so on.

As a specific example of the PECVD process, consider the deposition of hydrogenated amorphous silicon nitride, a-SiN_x:H, from SiH_4 and NH_3 mixtures using volume flow ratios $R = NH_3/SiH_4$. Under typical conditions [e.g., $T_s = 400^\circ$C and $P = 0.5$ torr ($= 66$ Pa) in RF discharges], the deposition rates of these a-SiN_x:H films are ≈ 0.1 to 0.5 nm/s and are controlled by the SiH_4 flow rate. This occurs because

TABLE W21.4 Typical Precursor Gases Used in PECVD

Film	Precursor Gases	Film	Precursor Gases
a-Si:H	SiH_4, SiH_4/H_2	a-Ge:H	GeH_4, GeH_4/H_2
a-C:H	C_2H_2, C_2H_4, C_6H_6	a-SiN_x:H	SiH_4/NH_3, SiH_4/N_2,
a-SiO_x:H	$Si(OC_2H_5)_4/O_2$,		SiH_2Cl_2/NH_3
	SiH_4/O_2,	a-SiC_x:H	SiH_4/C_2H_2
	$SiH_4/Ar/N_2O$	a-BN_x:H	$B_3N_3H_6$, B_2H_6/NH_3
a-C:F	CF_4, C_2F_4		

[†] For the application of the free-energy model to a-SiN_x:H, see Z. Yin and F. W. Smith, *Phys. Rev. B*, **43**, 4507 (1991); for a-C:H, see H. Efstathiadis, Z. L. Akkerman, and F. W. Smith, *J. Appl. Phys.*, **79**, 2954 (1996).

SiH_4 is dissociated much more rapidly than NH_3 in the plasma. For $R = 0$ a-Si:H films are deposited, and for $R \ll 1$ a fraction of the incorporated N atoms can act as substitutional donor impurities in a-Si:H. As R increases still further and more N is incorporated, the optical energy gap widens and the films become electrically more insulating. For very high ratios, $R \approx 60$, and for lower $T_s \approx 100^{\circ}C$, the films become N-rich, with N/Si ratios that can exceed the stoichiometric value of $\frac{4}{3}$ for Si_3N_4. These films do not correspond to a-Si_3N_4, even when N/Si $= \frac{4}{3}$ due to the incorporation of H in the range 10 to 30 at %.

The a-SiN_x:H films used in devices have N/Si ≈ 1 and typical compositions given by a-$Si_{0.4}N_{0.4}H_{0.2}$. Undesirable bonding configurations in these films include Si–Si bonds and Si–NH_2 bonding units. The former lead to an increase in the dielectric function and also cause optical absorption at low energies, while the latter lead to a lack of chemical and thermal stability. Films with higher H contents are in general not useful in devices. Films with compositions close to the compound silicon diimide [i.e., $Si(NH)_2$], the bonding analog of SiO_2, with NH units replacing O atoms, can be obtained at very high NH_3/SiH_4 flow ratios. Films of $Si(NH)_2$ are unstable in the presence of H_2O due to the chemical reaction $Si(NH)_2(s) + 2H_2O(g) \leftrightarrow SiO_2(s) + 2NH_3(g)$, particularly when Si–$NH_2$ bonding units are present. Films of a-SiN_x:H thus provide a typical example of how H incorporation can play a key role in controlling the properties of amorphous semiconducting and insulating films.

The plasmas used in PECVD processes include RF plasmas at 13.56 MHz (wavelength $\lambda = 22.1$ m) and microwave plasmas at 2.45 GHz ($\lambda = 12.2$ cm). The RF plasmas are often employed using a capacitively coupled parallel electrode configuration, as shown in Fig. W21.13, although inductive coupling is also used. The microwave plasmas typically consist of a plasma ball with dimensions of a few

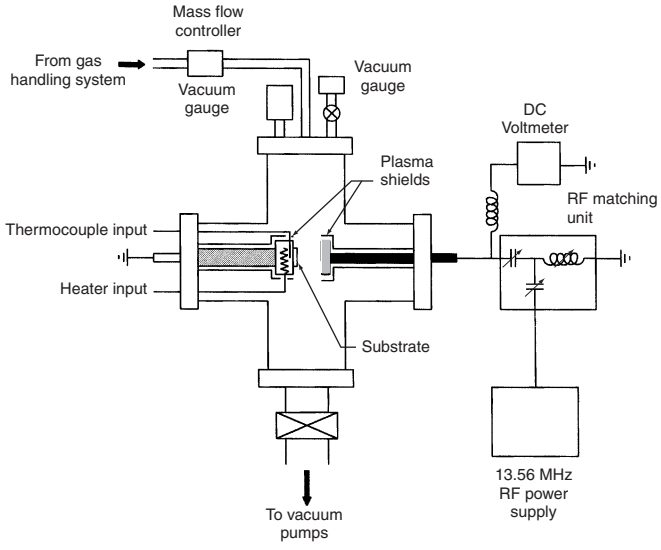

Figure W21.13. The RF plasmas used in plasma-enhanced CVD are typically employed in a capacitively coupled parallel electrode configuration, as shown here. (From K. Mui et al., *Phys. Rev. B*, **35**, 8089 (1987). Copyright 1987 by the American Physical Society.)

centimeters and are usually more confined in space than their RF counterparts. Electron cyclotron-resonance (ECR) plasmas which employ magnetic fields to aid in the coupling of energy into the plasma are also used in low-pressure discharges. Electron-impact dissociation of the feedstock gas in the plasma provides the excited neutral and charged species (i.e., free radicals and ions) needed for film deposition. Chemical reactions occurring in the gas phase and on the surface of the growing film can also produce species that are important for the deposition process.

A complete description and analysis of all the important processes occurring both in the plasma and on the surface of the growing film during PECVD is an extremely difficult task, due to the large number of possible species and processes and the often unknown rate constants and cross sections of these processes. A schematic model of the gas-phase and surface processes involved in the PECVD of a-Si:H from SiH_4 is shown in Fig. W21.14. The various ions, neutral radicals, and other molecular species present in the vapor phase are indicated, as are some of the surface reactions. The presence of the H-rich surface layer on the growing a-Si:H film is apparent. The net growth rate is the result of the competition between the deposition and etching rates. In most PECVD processes the substrate to be coated is mounted in a vacuum system on a heated substrate holder so that T_s can be varied from room temperature up to $\approx 400°C$. Typical discharge pressures are in the range 0.1 to 10 torr (13 to 1300 Pa) and typical plasma energy fluxes at the substrate are 10 to 100 mW/cm^2.

Hydrogen dilution (i.e., adding H_2 to the plasma) often has the advantage of actually reducing the hydrogen content of the deposited film by, for example, enhancing the removal from the growing surface of weakly bonded species such as SiH_2 or SiH_3.

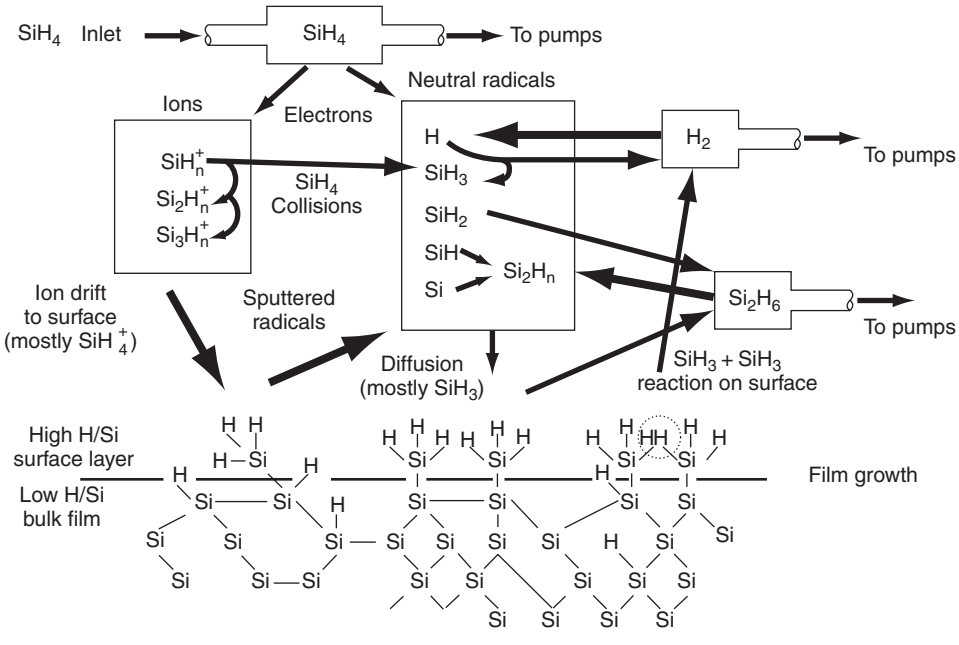

Figure 21.14. Gas-phase and surface processes involved in the plasma-enhanced CVD of a-Si:H from SiH_4. (From A. Gallagher, in *The Physics of Ionized Gases*, J. Puric and D. Belic, eds., World Scientific Press, 1987, p. 229.)

Another method used to reduce the hydrogen content is increasing T_s, which leads to increased mobility of the H atoms within the films, and their recombination into H_2 molecules, which can then diffuse to and desorb from the film surface. Higher deposition rates are also possible at higher T_s. The use of higher T_s allows greater atomic diffusion to occur in the films, which aids in the annealing (i.e., healing) of defects. Film stress and morphology are also strongly dependent on T_s as well as on ion bombardment.

Changes in the PECVD growth conditions, such as increasing the partial pressure of H_2 in SiH_4/H_2 mixtures, increasing the power density or the frequency of the plasma, or increasing the substrate temperature T_s, can lead to the deposition of *microcrystalline* (μc) films such as μc-Si:H. These μc-Si:H films have microstructures consisting of variable volume fractions of Si nanocrystals in an a-Si network. Preferential etching of the more weakly bonded amorphous component by H atoms is likely to play an important role in the deposition of μc-Si:H films.

In addition to deposition, reactive plasmas can also be used in a wide variety of etching processes, such as those used in the fabrication of Si devices. Some of these etching applications are discussed in Section W21.8. The plasma hardening of metal surfaces by the implantation of N or C ions, discussed in Section W21.13, and plasma doping by implantation of B ions into Si are also important materials processing procedures.

Another plasma-related mode of film deposition makes use of the *physical sputtering* of atoms from a target in, for example, an Ar plasma. The target material, as well as the deposited layer, can be a metal, semiconductor, or an insulator. The sputtered atoms are incident on the substrate, where they lead to the desired layer deposition. Physical sputtering is typically used for the deposition of metal films.

In another mode of operation, known as *reactive sputter deposition*, additional precursor gases are introduced into the plasma, where they are excited. These excited species contribute to the layer deposition since they can react with the target atoms both at the surface of the growing film and on the surface of the target. This method can readily be used to control the composition of the deposited layer. Reactive sputtering is typically used for the deposition of compound films such as oxides (including the high-T_c superconducting copper-based oxides), nitrides, carbides, and silicides. Typical precursor gases include O_2 and H_2O for oxygen, NH_3 and N_2 for nitrogen, CH_4 and C_2H_2 for carbon, SiH_4 for silicon, and H_2 when hydrogen is to be incorporated, as in a-Si:H.

W21.8 Fabrication of Si Devices

A brief overview of the important steps involved in the fabrication of Si-based electronic devices from Si wafers of sufficiently high resistivity is presented next. To illustrate the complexity of the process, consider the fabrication of a 256-Mbit dynamic random-access memory (DRAM). A wafer yields 16 chips, each 25 mm square and consisting of $\approx 3 \times 10^8$ devices with features as small as 0.25 μm. Due to the large number (≈ 300) of synthesis and processing steps involved in IC fabrication, it is not possible here to describe these procedures in detail. Wolf and Tauber (1990) and Maly (1987) provide useful descriptions of the steps involved in IC fabrication. Some of the important steps have already been described (e.g., the CVD of epitaxial Si films and the PECVD of silicon nitride dielectric films). The thermal oxidation of Si to form

passivating and protecting a-SiO_2 layers is discussed in Chapter 21. Other steps, such as diffusion (Chapter 6) and ion implantation (Section W21.3), are also discussed elsewhere. Therefore, only some additional details and current issues relevant to Si device fabrication are presented here.

Thermal Oxidation of Si. The *thermal oxidation* of Si to form layers of a-SiO_2 is repeated often during the fabrication of Si-based devices. In addition to protecting and passivating the surface of Si, oxide layers are also used as the surface for photoresist deposition, as masks for dopant diffusion, and as buried dielectric layers to isolate components of the device structure. Repeated oxidations of a given Si substrate can be carried out as often as necessary for the patterning of different circuit configurations via the photolithographic process, described later. For example, windows can be opened into an a-SiO_2 layer which can be used as diffusion masks, first for p-type doping into a n-type layer and then for n-type doping into the resulting p-type region in order to fabricate an *npn* transistor. This type of process is illustrated in Fig. W21.15.

The oxide dielectric layers include the thin *gate oxides* separating a metallic gate from, for example, the p-type region of a MOSFET, thicker *field oxides* which isolate transistors from metallic interconnecting wires, and dielectric caps which protect the device from the surrounding environment. Gate oxide thicknesses are typically ≈ 15 to 100 nm and are expected to decrease to the range 3.5 to 4.5 nm, and those of field oxides are ≈ 0.3 to 1 μm. These oxide layers are fabricated via the usual thermal oxidation process or via a plasma deposition process, discussed later. Thin gate oxides often include a region incorporating nitrogen (i.e., an oxynitride layer), which serves to suppress diffusion of boron from the polysilicon gate into the MOSFET channel.

The Si/a-SiO_2 interfaces can be prepared to be atomically or chemically abrupt, at least to within 0.5 nm, the dimensions of an $Si{-}O_4$ tetrahedron, and are flat on the scale of hundreds of nanometers. Nevertheless, the actual width of the interface (i.e., the region in which the properties of the Si and a-SiO_2 differ from their bulk

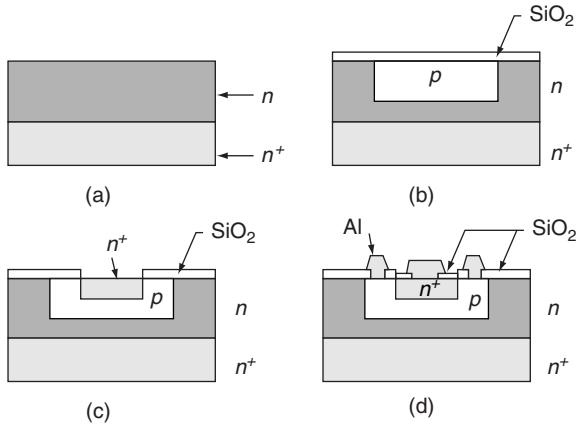

Figure W21.15. Fabrication of an *npn* transistor involving repeated oxidation, lithographic, and diffusion processing steps. In the case shown windows are created in an a-SiO_2 layer which can then be used as diffusion masks, first for p-type doping into a n-type layer and then for n-type doping into the resulting p-type region. (From B. Sapoval et al., *Physics of Semiconductors*, Springer-Verlag, New York, 1993.)

values) has been found to be ≈ 3 nm from sensitive core-level spectroscopies which can determine the strain in Si–O–Si bonding units. The properties of these interfaces are critically important for the operation of devices, and their physical and chemical structures and properties are discussed in Section 20.11.

Lithography. Optical lithography (i.e., *photolithography*) involves the patterning of two-dimensional circuits or designs onto Si wafers by means of the passage of light through a mask that corresponds to the outline of the desired circuit. This is illustrated in Fig. W21.16 and consists of the following sequence of steps:

1. A uniform a-SiO_2 layer is deposited onto the Si.
2. The a-SiO_2 layer is then covered by a layer of photosensitive polymeric material known as a *photoresist*. The photoresist is applied as a uniform liquid layer, using a spin-on procedure that is discussed in Section W21.24, and is then solidified via the application of heat.
3. The photoresist undergoes polymerization or cross-linking during exposure to light through a mask; this is the photoresist development step.
4. In the case illustrated involving the use of a negative photoresist, the unilluminated and hence unpolymerized areas of photoresist are removed via etching with an appropriate chemical solvent.
5. The exposed a-SiO_2 pattern is removed via etching using an acid that does not attack the polymerized photoresist.
6. The polymerized photoresist is finally removed via another suitable chemical solvent.

The patterned a-SiO_2 layer that remains on the surface can act as an insulating layer in the structure or can be used as a diffusion barrier in a subsequent processing step. The predominant method of photoresist removal is currently the use of oxygen plasmas which are described later in the discussion of etching processes.

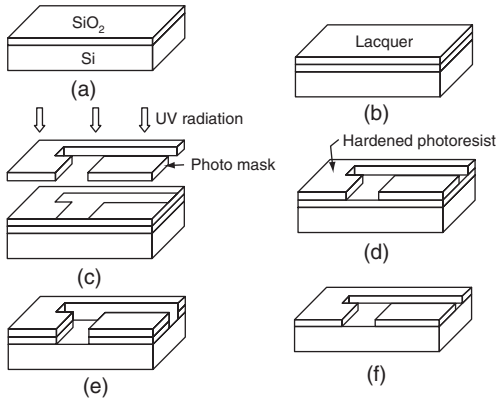

Figure W21.16. Optical lithography process involving the patterning of two-dimensional circuits or designs onto wafers through the use of light passing through a mask. (From B. Sapoval et al., *Physics of Semiconductors*, Springer-Verlag, New York, 1993.)

The interaction of light with photoresist materials such as the high-molecular-weight polymer polymethylmethacrylate (PMMA, also known as Plexiglas or Lucite) is discussed in Section 14.10. The light-induced breaking of bonds (i.e., photodissociation) in the long polymeric chains in the illuminated portions of the PMMA photoresist layer renders these regions susceptible to removal via etching. There are two types of photoresists in use: *negative photoresists*, which undergo light-induced cross-linking and so become insoluble and harder to remove after illumination, and *positive photoresists* like PMMA, which undergo light-induced chain breaking and so become more soluble and easier to remove after illumination. While negative photoresists are usually more photosensitive than positive photoresists and require less illumination, they have lower resolution and hence their use is not desirable in high-density ICs. PMMA is the photoresist with the highest-known resolution.

As the dimensions of features in ICs continue to decrease below 0.25 μm, optical lithography using UV light (e.g., the ArF laser line at $\lambda = 193$ nm) may no longer be possible since the minimum size of a feature is controlled by diffraction effects that limit the definition of the image to about one-half of the wavelength of the light used. The resolution limit D is given by

$$D = \frac{\lambda}{2 \sin \theta}, \tag{W21.20}$$

where θ is the angle subtended by the mask opening at a point on the surface and $\sin \theta$ is the numerical aperture (NA). For an opening of width w that is a height H above the substrate, $\tan \theta = w/2H$. The corresponding depth of focus, h, is given by

$$h = \frac{\lambda}{\sin^2 \theta}. \tag{W21.21}$$

Another important length scale governing the exposure depth is $1/\alpha$, the inverse of the absorption coefficient of the light in the photoresist.

Nanolithographic technologies (i.e., technologies with the higher resolution needed for producing geometrical circuit features with sizes below ≈ 0.1 μm) are based on shorter-wavelength beams of electrons or x-rays, or on the use of scanning probe microscopies such as scanning tunneling microscopy (STM) and atomic force microscopy (AFM). These advanced technologies are being explored as alternatives to optical lithography. Electron beams have the advantages of being able to be steered and focused rapidly using electric and magnetic fields. There are as yet no suitable photoresist materials for features smaller than 0.1 μm.

In the *LIGA process* (*lithographie galvanoformung abformung*), synchrotron radiation is employed to expose the photoresist polymer PMMA. Exceptionally sharp walls are produced, resembling steep cliffs. Metallization of the structure can even result in excellent molds from which replicas may be cast.

Diffusion. The thermal diffusion of dopants into a device in order to create junctions between n- and p-type regions, or just to change the electrical resistivity of a region, occurs repeatedly during device fabrication. Since solid-state diffusion is discussed in Chapter 6, only some details relevant to Si device fabrication are mentioned here.

Due to the need to limit the region of doping in the substrate, all diffusion processes are preceded by oxidation and mask-patterning lithographic steps. Layers

of a-SiO_2 serve as good mask materials for diffusion processes due to the low diffusion coefficients of typical dopants in the oxide. At typical diffusion temperatures of $T = 900$ to 1100°C, dopants present in a source at the Si surface will diffuse through the opening in the mask into the Si both vertically (i.e., normal to the surface), and laterally.

Two methods of dopant diffusion are typically used, constant-source diffusion or two-step diffusion. In the first method, used when shallow junctions are desired, a thick layer consisting of a mixture of B_2O_3 or P_2O_5 and SiO_2 is deposited onto the surface. This layer acts as a constant source of dopant atoms, so the dopant concentration at the surface remains essentially constant as diffusion occurs deeper and deeper into the substrate (see Fig. W6.2). The second method, used when deeper junctions are desired, starts with a predeposition step which is essentially the same as the constant-source method. After removal of the dopant source from the surface, a second, high-temperature step is used to drive the dopant atoms farther into the substrate (see Fig. W6.1).

Complicating the diffusion of acceptors such as B in Si are the effects known as *oxidation-enhanced diffusion* (OED) and *transient-enhanced diffusion* (TED). OED and TED both result from the injection of excess Si interstitials into the Si substrate and away from the Si/a-SiO_2 interface in the case of OED and out of a damaged ion-implanted layer in the case of TED. Dopants such as B must pair with defects such as vacancies or interstitials to move through the lattice, and as a result, their diffusion is affected by the motion of excess interstitials.

Ion Implantation. Ion implantation is used as an alternative to the introduction of dopants by diffusion in IC fabrication when the high temperatures associated with diffusion cannot be tolerated. In addition, the lateral spreading of dopants associated with the diffusion process is minimized when ion implantation is used, a significant advantage in high-density devices. As with diffusion, implantation occurs through a mask and extends into the Si for a characteristic distance known as the *range*. The mask is an opening in an a-SiO_2 overlayer or any other overlayer (metal, photoresist, etc.). Some of the important aspects of ion implantation are discussed in Section W21.3. The *dose* and *energy* of the implanted ions determine the doping level and the position of the resulting junction within the implanted Si. When desirable, implantation through a thin overlayer is possible as long as the incident ions are sufficiently energetic. A schematic phase-space map of the typical ion energies (in electron volts) and ion beam currents (in particle-amperes) used in semiconductor processing is illustrated in Fig. W21.17.

The lattice disorder created in the Si by the incident energetic ions can lead to dopant deactivation when the dopant atoms do not enter the lattice substitutionally or when traps are generated. A subsequent annealing step must then be carried out to repair the damage and for dopant activation.

When plasmas are used to excite the species to be implanted, the process is known as *plasma-immersion ion implantation* (PIII). In this method the substrate is immersed directly in the plasma, and rather than using accelerated beams of energetic dopant ions, high fluxes of relatively low-energy dopant ions are instead extracted from the plasma by applying pulsed high negative voltages, ≈ 2 to 4 kV, to the substrate. When PIII is used to form shallow p^+-n junctions, the n-type Si substrate is first converted to amorphous Si by using SiF_4 in the plasma, followed by the introduction of BF_3 to

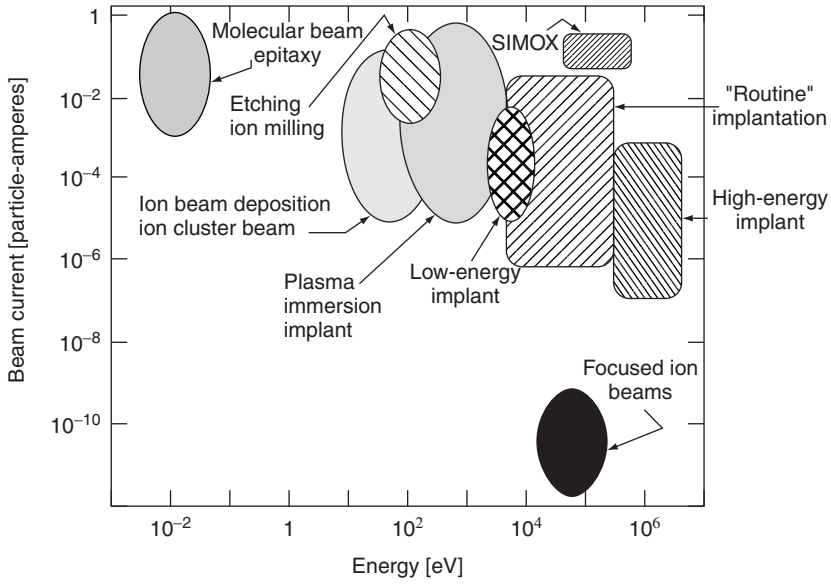

Figure W21.17. Schematic phase-space map of the typical ion energies (in electron volts) and ion beam currents (in particle-amperes) used in semiconductor processing. (From E. Chason et al., *J. Appl. Phys.*, **81**, 6513 (1997). Copyright 1997 by the American Institute of Physics.)

the plasma to implant B ions into the a-Si. An extremely shallow junction depth of 80 nm can be achieved following thermal activation of the dopant atoms using rapid thermal annealing of the implanted region at $T = 1060^{\circ}C$ for 1 s. The PIII process for dopant implantation is similar to the plasma carburizing and nitriding processes used to modify the surface properties of metals, as discussed in Section W21.13.

In the process known as *separation by implantation of oxygen* (i.e., SIMOX) a buried dielectric layer is created below the surface of a Si substrate via the implantation of oxygen ions. This process is a major candidate for the creation of Si-on-insulator (SOI) structures in which devices are isolated by being surrounded completely by an insulator rather than by using a reverse-biased *p-n* junction. The O^+ implantation consists of a high dose, $\approx 2 \times 10^{18}$ cm^{-2}, of ions, which leads to the formation of a continuous buried a-SiO_2 layer following an annealing step for 3 to 5 h at $T = 1100$ to $1175^{\circ}C$. The characteristic distance of the buried layer from the Si surface is 0.3 to 0.5 μm when O^+ ion energies of 150 to 180 keV are used.

Chemical and Physical Vapor Deposition. A variety of *chemical* and *physical* vapor deposition procedures are used to deposit the conducting, semiconducting, and insulating layers that are needed in device fabrication. Reactions between the incident vapor species and the substrate are not necessarily required to grow the desired films in these CVD and PVD procedures. As an example, a-SiO_2 layers must be deposited via PECVD when this dielectric layer is to be grown on a metallic layer instead of on Si. The CVD of epitaxial Si layers and the PECVD of the silicon oxide, nitride, and oxynitride layers used as dielectrics for interlevel isolation, for passivation, and as gate insulators have already been discussed. Si epilayers can be deposited on Si substrates with differing doping levels (e.g., an *n*-type Si epilayer deposited onto an

n^+ Si substrate). PVD in the form of electron-beam evaporation or sputtering is used for the deposition of Al layers.

A challenging problem is the deposition of conformal layers (i.e., layers of uniform thickness) on nonplanar substrates having steps, trenches, and holes. Examples of reliability problems in devices due to deposited layers with nonuniform thicknesses include inadequate electrical isolation in dielectric layers and nonuniform current densities in conducting layers, leading to enhanced electromigration in the conductors and hence open circuits. In the case of a-SiO_2 deposition, when mixtures such as $SiH_4/Ar/N_2O$ or $SiH_4/Ar/O_2$ are used, the sticking coefficients for SiH_n species are high, with the result that the a-SiO_2 layers tend not to be conformal. A method for obtaining conformal a-SiO_2 layers is plasma deposition using the liquid tetraethoxysilane (TEOS) as the source of the precursor in mixtures with O_2 or O_3 (ozone) and Ar. Oxide depositions using dilute TEOS/O_2 mixtures at $T = 200$ to 300°C result in lower deposition rates, < 50 nm/min, compared to SiH_4-based depositions, but the resulting layers have good conformality, due to the low sticking coefficients and higher surface mobility of the TEOS-based precursors.

Metallization. Aluminum and Al alloys have been the metals of choice for providing the electrical connections between circuit elements in ICs due to their desirable physical and chemical properties (e.g., excellent electrical conductivity, the ability to form both ohmic and Schottky barrier contacts to Si, good bonding and adherence to both Si and SiO_2 and also to diffusion barriers such as TiN and Ti, the ability to be patterned in Cl-based plasmas, and the ability to form a stable oxide, Al_2O_3, when exposed to air). Aluminum alloyed with 0.5 wt % Cu exhibits higher hardness and good electrical conductivity, along with improved resistance to electromigration, a process described in Section 12.9. The resistance to electromigration resulting from alloying Al with Cu is attributed to the precipitation of Cu at grain boundaries. This inhibits the harmful grain-boundary diffusion of Al, which leads to vacancy accumulation and void formation in the Al connecting lines. Even though Cu itself has low electrical resistivity and good resistance to electromigration, it has not been widely used so far as an interconnect metal because a successful dry-etching process has not been developed for patterning the Cu lines. In addition, diffusion barriers must be used between Cu lines and Si because Cu impurity atoms act as deep traps in Si.

Problems with Al layers deposited by PVD methods such as electron-beam evaporation and dc magnetron sputtering are associated with incomplete filling of vias and with poor step coverage for feature sizes below 0.5 μm. Other possible deposition procedures that may lead to improved via filling and step coverage include high-temperature Al-alloy sputtering processes, the use of Al reflow processes, and CVD at $T = 100$ to 200°C using Al-containing metal–organic molecules at deposition rates of 100 to 200 nm/min. Aluminum reflow processes involve the use of elevated deposition temperatures or postdeposition annealing to allow the deposited Al alloy to flow into and fill via/contact holes. The Al-alloy reflow temperatures lie below the alloy melting points by ≈ 150°C, with both temperatures decreasing with increased alloying of elements such as Cu or Ge.

The refractory metal W can be selectively deposited via CVD and allows much better step coverage and via and hole filling than Al. In addition, it exhibits excellent resistance to electromigration. Bilayers of Ti and TiN serve as *diffusion barriers* between W and Si and also as intermediate layers for the CVD of W. The initial Ti

layer is reacted with the underlying Si at $T \approx 700^{\circ}C$ to form a titanium silicide Ti_xSi_y phase with both good electrical conductivity and contact to the underlying Si. A TiN_x diffusion barrier layer is then deposited to prevent undesired reactions between the Ti_xSi_y layer and the fluorine involved in the CVD of W via the hydrogen reduction of the WF_6 precursor [i.e., $WF_6(g) + 3H_2(g) \rightarrow W(s) + 6HF(g)$]. When selective deposition of W and lower deposition temperatures are required, the silane reduction of WF_6 can be used [e.g., $2WF_6(g) + 3SiH_4(g) \rightarrow 2W(s) + 3SiF_4(g) + 6H_2(g)$].

Local interconnects formed from low-resistivity doped polycrystalline Si layers are useful because these layers can make good electrical contact to Si substrates and can also serve as diffusion barriers between Si and Al lines. Electrical contacts between pure Al and n^+ and p^+ Si are not stable at processing temperatures in the range $T = 350$ to $500^{\circ}C$, due to the solubility of Si in Al and also to the rapid diffusion of Si into the polycrystalline Al contacts. The reciprocal diffusion of Al into the Si layer can lead to the *spiking* (i.e., shorting) of shallow junctions. The use of polysilicon is restricted to buried contacts and to limited regions due to its relatively high sheet resistance of 20 to 30 Ω/square.

Etching Processes. Device fabrication involves a variety of processing steps employing the etching or controlled removal of material from the surface of the wafer. The etching or stripping process can employ either wet, liquid-phase or dry, gas-phase etchants. *Chemical etching*, in which the etchant reacts with the material to be removed, can occur in either the liquid or gas phases, is typically highly selective, and is isotropic (i.e., the etching occurs at the same rate in all directions). *Physical etching* is a gas-phase process in which material is removed by sputtering (i.e., via energy and momentum transfer from incident ions), is less selective than chemical etching, and is typically anisotropic (i.e., etching occurs preferentially in one direction). *Selectivity* refers to the ability of the etching process to remove some materials but not others. An example is positive-photoresist lithography, where liquid solvents etch away the illuminated portion of the photoresist while the unilluminated portion is unaffected, or as when an HF acid etch is used to remove a-SiO_2 but neither Si nor photoresist.

A plasma etching process with both chemical and physical components is *reactive-ion etching* (RIE), in which ions created in a plasma react with and also transfer kinetic energy to the material to be etched. An advantage of RIE is that it can be both selective and anisotropic. Plasma etching is used for the removal of Si, of a-SiO_2 and silicon nitride, of metals, and of photoresist. Appropriate etching species are chosen for each case: for example, F atoms and Ar^+ ions for etching Si or polysilicon (forming SiF_4) and O atoms for etching or stripping photoresist (forming CO, CO_2, and H_2O). The Ar^+ ions provide additional kinetic energy, which can greatly increase the yield of the etching process by enhancing chemical etching reaction rates on the surface. For example, a 1-keV Ar^+ ion can result in the removal of up to 25 Si atoms when a flux of F atoms is also incident on the surface. The use of Ar^+ ions can also increase the anisotropy of the etching but may decrease the etching selectivity.

Etch inhibitors are also used in RIE to prevent etching from occurring outside the area exposed to the ion beam. An example is the anisotropic etching of trenches and holes in Al using CCl_4/Cl_2 mixtures, where the CCl_4 molecules are the inhibitor precursors. A protective, etch-inhibiting amorphous chlorocarbon film is present on the areas of the Al surface not exposed directly to the ion beam, including on the sidewalls

of the features being etched. The presence of C in the etching mixture thus leads to an enhancement of the anisotropic etching of the desired trenches and holes.

Reactive-ion etching rates are very difficult to predict. This is due to difficulties associated with modeling the plasma processes giving rise to the incident fluxes of reactive atomic and molecular radicals and ions on the surface. There are also difficulties with modeling the many surface processes, including adsorption, diffusion, reaction, and desorption, involved in the generation of etching products. In addition, in the F etching of Si, a fluorinated SiF_x surface layer two to five monolayers thick is present and the diffusion of the etching species, F^- ions, through this layer plays an important role in the process. A rough estimate for the characteristic thickness of this layer is $d \approx D/R_e(\text{Si})$, where D is the diffusion coefficient for F^- ions in the surface layer and $R_e(\text{Si})$ is the etching rate in m/s.

The etching of Si by halogen atoms such as F and Cl is found to depend on the doping level and type of the Si substrate, with etching rates of n-type Si exceeding those of p-type Si by a factor of about 2 for F and by many orders of magnitude for Cl. These observations indicate that the position of the Fermi level and the concentrations of charge carriers near the Si surface can play important roles in the etching process. The current model is that electrons in n-type Si tunnel from the bulk through the SiF_x layer, leading to the formation of F^- or Cl^- ions that attack Si–Si bonds in either the surface layer or the bulk. Molecules such as CF_4 are typically used as etching precursors because the etching of Si by F_2 leads to roughening the surface through pitting. The overall etching reaction in this case can be written as

$$4CF_4 + Si \longrightarrow SiF_4 + 2C_2F_6. \tag{W21.22}$$

When wet chemical etching is used to remove an unprotected a-SiO_2 layer, the isotropic nature of the etching can cause unwanted undercutting of the oxide beneath the protective photoresist mask. As a result, the pattern obtained is not the one desired. Dry etching carried out at reduced pressures in the gas phase can combine the advantages of chemical etching in being selective and physical etching in being anisotropic, so that no undercutting of the oxide occurs.

The smallest feature size (e.g., the minimum trench width) that can be obtained via etching is

$$w \approx \frac{2d}{a_h}, \tag{W21.23}$$

where d is the depth of the trench and $a_h = R_{ev}/R_{eh}$ is the ratio of the vertical and horizontal etch rates of the material in which the trench is being etched. As an example, 0.2-μm-wide and 4-μm-deep trenches with the aspect ratio $d/w = a_h/2 = 20$ can be etched into single-crystal Si using F-based chemistry.

Remaining problems associated with the use of plasmas in device fabrication are related to ion-induced damage and plasma-induced contamination.

Annealing. Annealing at elevated temperatures is often required in IC fabrication for a variety of purposes:

1. To remove, or at least minimize, processing-induced defects (e.g., those created in the Si lattice during ion implantation).

2. To activate implanted dopants in Si or polysilicon following ion-implantation procedures.
3. To drive dopant atoms farther into the Si following their implantation in a shallow layer.
4. To promote the reactions between deposited metals such as Ti and the underlying Si in order to form desired silicides.
5. To deactivate deep trap-generating impurities such Cu and Fe via gettering, a process in which these impurities diffuse to and are immobilized in the strain fields of extended defects such as oxide precipitates or dislocations. In this way the traps are removed from the active area of the device.

The time and temperature of an anneal must be chosen so that unwanted dopant redistribution does not occur. Any exposure of the device to high temperatures must therefore be as brief as possible. A method for limiting the annealing time is the process of *rapid thermal annealing* (RTA), also known as *rapid thermal processing* (RTP). A typical RTA dopant drive-in procedure involves a rapid temperature increase to $T = 1050$ to 1150°C, a 10-s anneal, and a rapid decrease to temperatures at which diffusion is negligible.

W21.9 Processing of Microelectromechanical Systems

The fabrication of Si-based microstructures for use in *microelectromechanical systems* (MEMS) having typical dimensions ≈ 1 to 100 μm is an exciting new area of materials research.[†] In addition to its well-known and extremely versatile electronic properties, crystalline Si also possesses very useful mechanical and thermal properties, such as high durability, elasticity, and thermal conductivity, which can be exploited in very small electromechanical structures. With the development of MEMS, Si semiconductor device-fabrication technology can now also be exploited in sensors and actuators for measurement and control in the fields of thermodynamics, optics, magnetism, acoustics, and hydrodynamics. Besides Si, other materials used in MEMS include a-SiO_2, crystalline quartz, and other ceramics, such as SiC. Since MEMS technology is in a state of rapid development, only a brief survey is given here.

The fabrication of MEMS is involved primarily with the processing of Si wafers into the desired final forms using a variety of etching and micromachining procedures. These processing procedures currently include the following:

1. Anisotropic wet chemical etching, usually in KOH solutions
2. Dry etching (i.e., reactive-ion etching) with the etchant activated via plasma excitation
3. Surface micromachining involving the removal of a sacrificial layer of a-SiO_2 or porous Si via etching in HF
4. Porous Si technology, also involving surface micromachining but using much thicker sacrificial layers of porous Si, up to hundreds of micrometers thick

[†] A recent review article is W. Lang, *Mater. Sci. Eng.*, **R17**, 1 (1996).

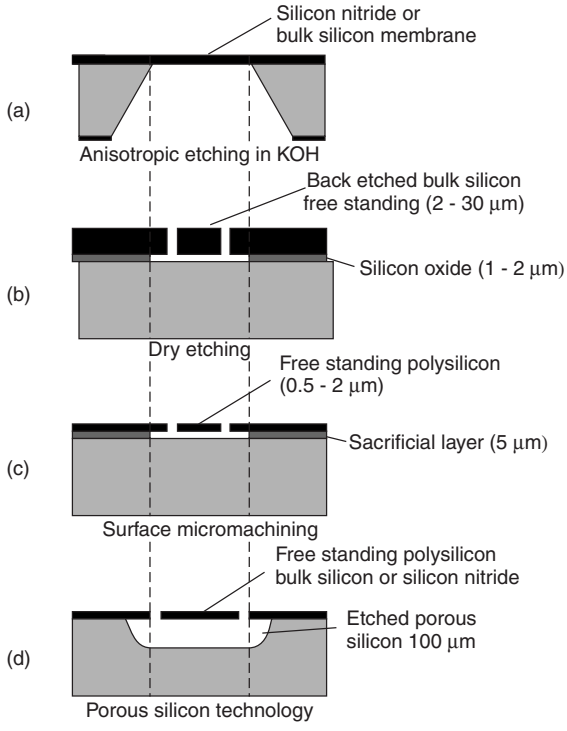

Figure W21.18. Micromachining processes currently used to fabricate microelectromechanical systems (MEMS) from Si wafers: (*a*) anisotropic wet chemical etching; (*b*) dry etching or reactive-ion etching; (*c*) surface micromachining involving a sacrificial layer of a-SiO_2; (*d*) porous Si technology, also involving surface micromachining but with much thicker sacrificial layers of porous Si. [Reprinted from W. Lang, *Mater. Sci. Eng.*, **R17**, 1 (1996). Copyright 1996, with permission from Elsevier Science.]

Examples of these processes are shown in Fig. W21.18. Free-standing features (e.g., Si cantilevers) are readily produced. The key to the rapid growth of MEMS technology is that most of these procedures involve deposition, lithography, and etching processes that have already reached an advanced level of development in Si electronic device fabrication. Porous Si, however, is a relatively new material consisting of variable volume fractions of crystalline Si filaments or wires and of empty pores, which is prepared by electrochemical anodic etching or anodization of crystalline Si in HF (see Fig. W11.9). The use of thick porous Si in MEMS is also compatible with Si device-fabrication techniques.

While Si electronic devices are essentially planar, containing circuit elements with typical thicknesses ≈ 1 μm, Si electromechanical devices or MEMS are truly three-dimensional and often contain free-standing structures such as cantilevers and bridges. The current trend in MEMS is to include several Si-based electronic devices and mechanical sensors and actuators in a single MEMS. The most widely used Si MEMS sensors at present are pressure transducers and thermopile radiation detectors. Other MEMS include micromotors, micromirrors in optical switches, accelerometers, microvalves, and flow sensors. In the future, MEMS actuators may be used to move STM tips in three dimensions as part of data storage systems at the near-atomic level.

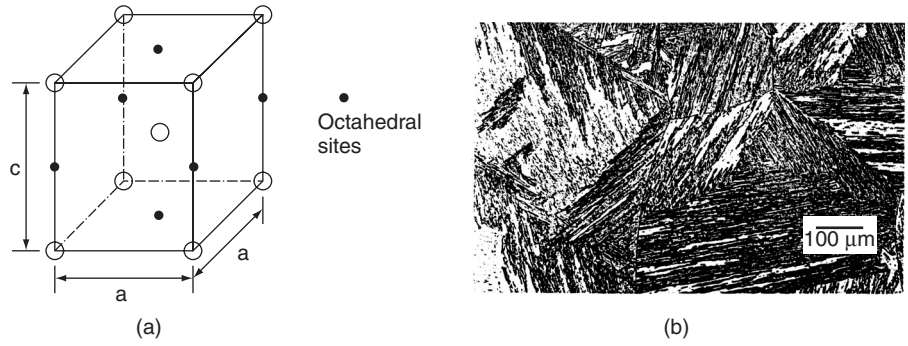

Figure W21.19. Martensite is a supersaturated solid solution of interstitial C in Fe. (*a*) Body-centered tetragonal (BCT) unit cell of martensite. The Fe atoms are actually displaced from their normal lattice sites to accommodate the C atoms in the octahedral sites. (*b*) Lath microstructure of martensite in a Fe–2Mn–0.03C wt % steel. (From ASM Handbook, 9th ed., Vol. 9, *Metallography and Microstructures*, ASM International, Materials Park, Ohio, 1985, p. 670.)

W21.10 Synthesis and Processing of Steels

While the simplest steels are just Fe–C alloys, steels in general can be very complex materials in both composition and microstructure. This complexity makes the design of a steel with a given set of properties quite challenging. It is useful first to review how the complex phases that may be present in steels are related to the simpler phases of pure Fe and Fe–C compounds and alloys.

Nonequilibrium Multicomponent Phases in Steels. The various nonequilibrium, multicomponent phases of Fe and Fe-based alloys and compounds which are the identifiable components of a wide variety of steels are described briefly next. These phases are all formed from the transformation or decomposition of *austenite* as the steel is cooled below the eutectoid temperature and include *pearlite*, *bainite*, *martensite*, and *acicular ferrite*. Table W21.5 summarizes the properties of these important phases and also of their multicomponent mixtures, which are found in the steels commonly used today.

Pearlite. *Pearlite* is a coarse, lamellar eutectoid mixture consisting of alternating layers of cementite and ferrite, shown in Fig. 21.11, which results from the decomposition of austenite as its temperature is lowered below $T_e \approx 727^\circ$C. Along with ferrite, it is a very common constituent of a broad range of steels in which it makes a substantial contribution to the strength of these materials. Pearlite also reduces the ductility and toughness of steels since cracks can nucleate at the ferrite–cementite interfaces.

The diffusion of C atoms is usually assumed to be the rate-controlling step for the nucleation and growth of pearlite in austenite. This is essentially a high-temperature reaction that occurs between T_e and $T \approx 550^\circ$C. Nucleation can take place at a variety of sites, including at austenite grain boundaries as well as on ferrite and cementite phases when they are already present in the austenite. At low transformation temperatures where the diffusion of C is slower, the lamellar spacing is much smaller and the resulting material is known as *fine pearlite*. The spacing of the lamellae in pearlite

TABLE W21.5 Important Phases of Fe, Fe–C Compounds and Alloys, and Their Multicomponent Mixtures Found in Steels

Phase	Structure and Description[a]	How Phase Is Obtained
	Equilibrium Phases of Pure Fe	
α-Fe (ferrite)	BCC, $a = 0.286$ nm at $T = 20^\circ$C; stable up to $T = 912^\circ$C; $T_C = 769^\circ$C	Stable phase at STP
γ-Fe (austenite)	FCC, $a = 0.364$ nm at $T = 912^\circ$C	Stable phase for $912 < T < 1394^\circ$C
δ-Fe (δ-ferrite)	BCC, $a = 0.293$ nm at $T = 1394^\circ$C; $T_m = 1538^\circ$C	Stable phase for $T > 1394^\circ$C
	Equilibrium Fe–C Compound	
Fe_3C (cementite)	Orthorhombic, $a = 0.509$, $b = 0.674$, $c = 0.452$ nm; a complex interstitial compound	Present in Fe–C alloys under conditions of metastable equilibrium (see Fig. 21.9)
	Equilibrium $Fe_{1-x}C_x$ Alloys	
α-Fe–C (ferrite)	Solubility limit of C in α-Fe at $T = 27^\circ$C: $x = 1.2 \times 10^{-6}$ (0.00012 at % or 1.2 ppm)	Present in Fe–C alloys under equilibrium conditions (see Fig. 21.9)
γ-Fe–C (austenite)	Solubility of C in γ-Fe at $T = 1150^\circ$C: $x \approx 0.09$ (9 at %)	Present in Fe–C alloys under equilibrium conditions (see Fig. 21.9)
	Nonequilibrium Multicomponent Phases	
Pearlite	A coarse, lamellar form of cementite in ferrite; a eutectoid structure	Formed between $T = 720$ and 550°C during cooling of austenite
Bainite	An intermediate structure composed of fine aggregates of ferrite plates (laths) and cementite particles	Formed between $T = 550$ and $\approx 250^\circ$C during cooling of austenite
Martensite	BC tetragonal, $c/a = 1 + 0.045$ wt % C; a supersaturated solid solution of interstitial C in ferrite, having a lath or lenticular microstructure	Rapid quenching of austenite to keep C in solution; formed between $T \approx 250^\circ$C and room temperature or below
Acicular ferrite	A disorganized structure of randomly oriented ferritic plates in a matrix such as martensite	Nucleation of ferrite at small, nonmetallic inclusions during cooling of austenite

[a]The range of thermal stability is given at $P = 1$ atm.

is larger at higher transformation temperatures due to the enhanced diffusion of C, with the resulting material known as *coarse pearlite*. The spacing is also controlled in part by the competition between the decrease in free energy associated with the more stable phase and the increases in surface energy associated with the interfaces between the ferrite and cementite lamellae and of any strain energy associated with the transformation.

Bainite. The term *bainite* refers to the intermediate structures found in steels, which are composed typically of fine aggregates of ferrite plates or laths and cementite particles. Bainite is formed at intermediate temperatures ($T \approx 250$ to 400°C for lower bainite and $T = 400$ to 550°C for upper bainite), below those at which pearlite ($T = 550$ to 720°C) is formed and above those at which martensite is formed (typically from room temperature up to $T \approx 250$°C). Bainite can also be formed when austenite is cooled too rapidly for the diffusion of C required for the formation of pearlite to occur and too slowly for martensite to be formed. Depending on the contents of C and of other alloying elements, the bainitic microstructure can be quite complicated, with austenite and martensite replacing cementite. There is a start temperature T_{Bs} for the austenite-to-bainite transition, with the amount of bainite that can be formed, increasing as T is lowered below T_{Bs}. The TTT diagram shown in Fig. 21.12 illustrates the formation of bainite at intermediate temperatures. Upper bainite is favored in low-carbon steels, while lower bainite is favored in high-carbon steels.

Martensite. *Martensite* is a supersaturated solid solution of interstitial C in Fe formed via the rapid quenching of austenite, which prevents the diffusion of C that would result in the formation of cementite. The body-centered tetragonal (BCT) crystal structure of martensite is shown in Fig. W21.19*a*. Carbon atoms are randomly distributed in the six equivalent octahedral interstitial sites at the midpoints of the edges along the c axis and in the centers of the basal faces. The lattice parameters of the BCT martensite unit cell depend on the C composition according to $a_{\text{mar}} = (0.286 \text{ nm})(1 - 0.0035 \text{ wt \% C})$ and $c_{\text{mar}} = (0.286 \text{ nm})(1 + 0.041 \text{ wt \% C})$, resulting in $c_{\text{mar}}/a_{\text{mar}} = (1 + 0.045 \text{ wt \% C})$. The lattice constant $a = 0.286$ nm of α-Fe has been used here for the zero-carbon limit.

The corresponding lath microstructure of martensite (Fig. W21.19*b*) can appear in a matrix of ferrite or pearlite. The martensitic transformation, known as a *diffusionless transformation*, involves the rapid appearance of shear strain in the FCC austenite lattice. The result is a change in shape of the unit cell from cubic to tetragonal. The preferential occupation of the octahedral sites by the C atoms distorts the structure, thus determining the c axis of the resulting BCT crystal structure. High densities of dislocations and also slip and twinning can occur in the martensite during its formation. Similar martensitic transformations or reactions occur in other alloys, such as Fe–Ni, In–Tl, and the shape-memory alloys discussed in Chapter W12.

The decomposition of metastable austenite to form martensite usually occurs over a well-defined range of temperatures, beginning at the *martensitic start temperature* T_{Ms} (often written as M_s), which ordinarily lies in the range from $T \approx 250$°C to below room temperature. Additional martensite is formed as the temperature is lowered further below T_{Ms}, until most of the austenite has been converted to martensite at the *finish temperature* T_{Mf} (or M_f). The transformation is an *athermal* one (i.e., it is not thermally activated and occurs essentially instantly once a nucleus of martensite is formed). Thus there is no time delay for the formation of martensite on the TTT diagram as observed for the formation of pearlite or bainite. The amount of austenite converted to martensite depends only on temperature and not on the time allowed for the transformation. Both T_{Ms} and T_{Mf} are lower when the austenite phase in the steel has been stabilized by carbon or other alloying elements. The cooling must occur rapidly enough so that the metastable austenite does not transform instead to ferrite, pearlite, or bainite at temperatures between T_e and T_{Ms}.

The actual microstructure present in a quenched steel will often exhibit spatial variations from the surface into the bulk, due to the fact that the cooling rate and temperature will be different at different depths within the sample. This is certainly the case in rapidly solidified steels, as discussed later.

Rapidly quenched steels that have both enhanced hardness and brittleness due to the formation of martensite from austenite are said to have good *hardenability*. The strength of the steel due to the martensite is enhanced as the C content is increased and can result from a variety of strengthening mechanisms, several of which are described later. When a martensitic steel is reheated so that the C can diffuse, the martensite will be transformed into other phases, such as pearlite and bainite. This process, known as *tempering*, is also described.

The cooling rates needed to transform a given steel completely to martensite can be determined from another type of temperature–time diagram, the continuous-cooling transformation (CCT) diagram shown in Fig. W21.20. This diagram provides information concerning the kinetics of the transformation which is not obtainable from the

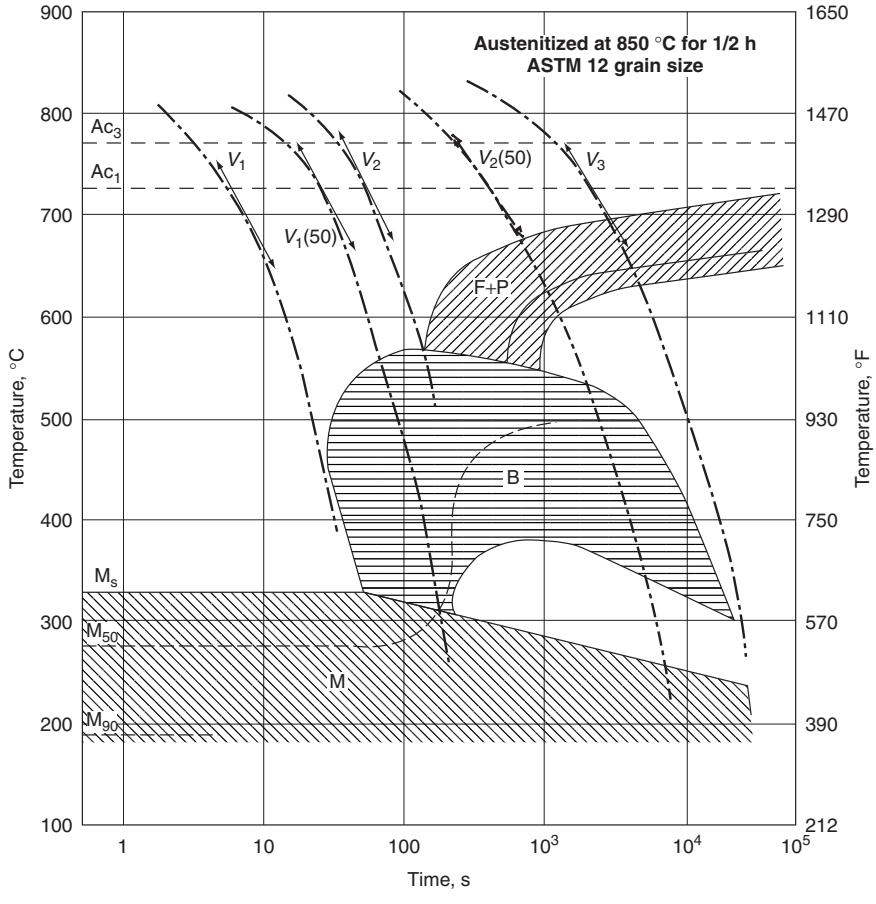

Figure W21.20. The cooling rates needed to transform a given steel completely to martensite (M) can be determined from the continuous-cooling or CCT diagram, shown here for 30 NC11 steel. The ferrite (F), pearlite (P), and bainite (B) phase regions are also shown. (From ASM Handbook, 9th ed., Vol. 4, *Heat Treatment*, ASM International, Materials Park, Ohio, 1991, p. 26.)

Figure W21.21. Coarse acicular ferrite, a disorganized structure of randomly oriented ferritic plates, is shown in a weld zone along with polygonal ferrite. The horizontal bar corresponds to 20 μm. (From ASM Handbook, 9th ed., Vol. 9, *Metallography and Microstructures*, ASM International, Materials Park, Ohio, 1985, p. 585.)

isothermal TTT diagram shown in Fig. 21.12. In the CCT diagram the ferrite, pearlite, and bainite phases are shown in addition to martensite.

Acicular Ferrite. *Acicular ferrite* is a nonequilibrium phase that has superior mechanical properties, including toughness, and consists of a disorganized structure of randomly oriented, interlocking ferritic plates in a matrix such as martensite. This phase can be obtained via the incorporation of small, nonmetallic inclusions that serve as nucleation sites for the plates. It can also appear in weld zones (Fig. W21.21). The morphology of this phase is three-dimensional since the ferritic plates can nucleate and grow in several different directions around an inclusion. Whether bainite or acicular ferrite is formed in a given steel as austenite is cooled depends on the ratio of nucleation sites at austenitic grain boundaries to those at the surfaces of inclusions, with grain-boundary nucleation leading preferentially to bainite. Ti_2O_3 and other oxide particles have been found to be especially effective in nucleating acicular ferrite, with the exact mechanism remaining unknown.

Processing Treatments for the Strengthening of Steels. A variety of processing treatments are used to strengthen steels and also other metals and alloys (e.g., Al alloys and Ni alloys). Important examples of these processes are given now, and a brief description of the strengthening mechanism is presented for each case. The strength of a given steel often results from contributions from more than one of these mechanisms. In practically every case the strengthening occurs via the pinning of dislocations, as discussed in Chapter 10. The specific application for which a given steel is designed will determine the conditions under which strength is needed (e.g., at high temperatures, under repeated loading, along with good ductility, etc.). Due to the large number of available processing variables, it is not possible to discuss here all of the important processing treatments that can be used to strengthen steels.

Mechanical Work Hardening. The tensile strength of a plain carbon steel that contains no other alloying elements can be increased up to 1500 MPa when it is drawn down

(e.g., to a wire) in a *work-hardening* or *cold-working* process in which its cross-sectional area is reduced by up to 95%. This large increase in strength produced by plastic deformation results from the generation of defects such as dislocations and dislocation arrays which reduce the mobility of other dislocations. The measured shear stress typically arises from two dislocation-pinning mechanisms, one arising from "small" defects, such as isolated dislocations, and the other from "larger" defects, such as dislocation arrays. The former mechanism decreases with increasing T, due to the thermally activated motion of dislocations around small defects while the latter is temperature independent. Work hardening is discussed in more detail in Section 10.13, where the dependence of the shear yield stress τ_y on dislocation density and strain is discussed in detail.

Solid-Solution Strengthening. Steels can also be strengthened or hardened by the presence of *interstitial* or *substitutional* impurities. The strong, attractive interactions between dislocations and the interstitial impurities C and N play an important role in this strengthening mechanism. Since interstitial C and N atoms as well as dislocations produce their own strain fields in the material, the attractive interaction arises from an overall reduction in strain energy when the C and N atoms reside in the strain field of a dislocation. The binding energy of a C atom to a dislocation in Fe is ≈ 0.5 eV. At high interstitial concentrations the resulting distribution of interstitial atoms surrounding the dislocation, known as the *Cottrell atmosphere*, can condense at the dislocation core. The movement of dislocations under the influence of an external stress will clearly be impeded by this interaction since the Cottrell atmosphere of interstitials has the effect of increasing the effective mass or inertia of the dislocation.

The condensation of interstitial atoms near dislocations can occur in steels at temperatures even as low as room temperature, due to the high diffusivity of C and N through defect-free regions of the material. Under applied stress and at higher temperatures, thermal activation of dislocations away from the atmosphere of interstitials can lead to a reduction of the yield strength. The strengthening process known as *strain aging* occurs under an applied stress after the yield point has been reached when interstitial atoms condense on newly generated dislocations.

The martensite structure, formed by rapid quenching, is usually very hard, due primarily to interstitial C and the resulting solid-solution strengthening but also due to the high densities of dislocations caused by the transformation of austenite to martensite. Martensite can, however, be brittle and not very ductile. The process known as *tempering*, (discussed later), is often used to increase its ductility and toughness.

The strengthening resulting from solid solutions of substitutional impurities such as Si, Mn, Cr, and Mo in steels results from the strain introduced into the structure by these impurities and thus is greater for impurity atoms, whose sizes are quite different from that of the host Fe atom. The increase of yield stress $\Delta\sigma_y$ of steel for various interstitial and substitutional impurities is illustrated in Fig. W21.22. The interstitial impurities C and N can be seen to have a much larger effect on σ_y than the substitutional impurities Si, Mn, Mo, and Ni due to the tetragonal distortions introduced into the lattice by C and N. These tetragonal distortions allow the stress fields of C and N impurities to interact with both edge and screw dislocations, while substitutional impurities have spherically symmetric stress fields and so can interact only with edge dislocations. Since substitutional alloying elements are usually added to the steel for other reasons

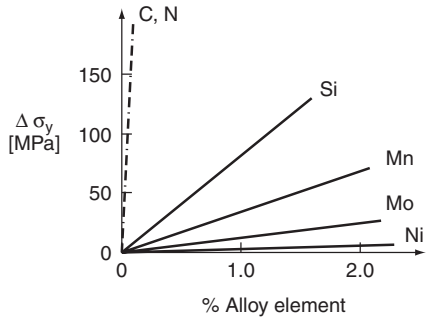

Figure W21.22. Increase $\Delta\sigma_y$ of the yield stress of steel for various interstitial and substitutional impurities. (From ASM Handbook, Vol. 1, *Properties and Selection: Iron, Steels, and High-Performance Alloys*, ASM International, Materials Park, Ohio, 1990, p. 400.)

(e.g., to improve corrosion resistance or to combine with oxygen or sulfur), the increase in strength associated with their presence can be considered a bonus.

Strengthening via Grain-Size Reduction. The reduction of grain size and the resulting increase in the number of grain boundaries are some of the most effective ways of increasing the strengths of steels. The *Hall–Petch relation* between the yield stress σ_y and the average grain size d of a material,

$$\sigma_y(d) = \sigma_0 + \frac{k_y}{\sqrt{d}}, \tag{W21.24}$$

is described in Section 10.14. Here σ_0, the yield stress for a single crystal with no grain boundaries, and k_y are constants that are independent of d for a given steel. The strengthening effect of grain boundaries results from their ability to pin dislocations. Reduction of the grain size in steels into the range 2 to 10 μm can produce yield stresses of over 500 MPa. This reduction is typically achieved via hot rolling and the addition of small amounts of certain alloying elements. The grain size can also be controlled by varying the cooling rate (i.e., the time available for the grains to grow). The kinetics of grain growth in metals are discussed in Section 21.5.

The growth of larger grains can be inhibited by the addition of small amounts, < 0.1 wt %, of grain-refining elements such as V, Al, Nb, and Ti, which form carbides, nitrides (e.g., VC and AlN), or carbonitrides. The 3 to 10-nm carbide and nitride particles that are formed tend to pin grain boundaries, thus helping to prevent grain growth. The resulting steels, which also contain 0.008 to 0.03 wt % C and up to 1.5 wt % Mn, have yield strengths in the range 450 to 550 MPa and are known as high-strength low-alloy (HSLA) steels or micro-alloyed steels.

Dispersion Strengthening. The strengthening of steels through the introduction of more than one structural phase in the ferrite matrix is known as *dispersion strengthening*. The typical phases present in plain carbon steels include carbides such as cementite, nonequilibrium phases such as pearlite, bainite, and martensite, and the precipitates formed by tempering. In alloy steels the thermodynamically more stable carbides of Si, Mn, and V often replace iron carbides. Other possible phases in steels include nitrides, other intermetallic compounds, and graphite.

A simple relation has been developed by Orowan for the yield stress σ_y of an alloy containing a random distribution of spherical particles of a different phase which are impenetrable by dislocations. With an average interparticle spacing Λ, the result is

$$\sigma_y(\Lambda) = \sigma_0 + \frac{2T_L}{b\Lambda}, \tag{W21.25}$$

where σ_0 is the yield stress of the particle-free matrix and T_L and b are the line tension (i.e., energy per unit length) and Burgers vector of a typical dislocation, respectively. An order-of-magnitude estimate for the line tension is $T_L \approx Gb^2 \approx 1.7 \times 10^{-9}$ J/m $\approx$ 10 eV/nm, using $G \approx 82$ GPa as the shear modulus and $b = a/2 = 0.144$ nm for Fe. The term $2T_L/b\Lambda$ is the stress required to move a dislocation past a second-phase particle via bowing. This process leaves a dislocation loop around each such particle. Equation (W21.25) is only approximately valid for steels in which the precipitates are plates or rods. In pearlite where the microstructure consists of a lamellar mixture of cementite and ferrite, the parameter controlling the strength is usually the average size of the uninterrupted ferritic regions, known as the *mean free ferrite path* (MFFP). In this case the flow stress is proportional to $(\text{MFFP})^{-1/2}$, a relationship of the Hall–Petch type [see Eq. (W21.24)]. Thus the fine pearlite formed at lower T will be stronger than the coarse pearlite formed at higher T.

The extent of the dispersion strengthening in a given steel is controlled by the C content, by alloying, and by the processes that determine which phases are present (e.g., heat treatment, tempering, etc.). When steels are quenched in order to form martensite, they are typically very strong but also tend to be quite brittle. Subsequent reheating or tempering of martensitic steels at an intermediate temperature between $T \approx 150$ and 700°C (i.e., below the eutectoid temperature T_e) is used to improve their ductility and toughness without at the same time causing too large a decrease in strength. The tempering process is controlled by the diffusion of carbon, which comes out of the supersaturated solid solution found in martensite and forms finely divided carbide phases. The martensite is thus converted to ferrite and the resulting material is then a dispersion of fine particles of cementite or transition metal (TM) carbides in a ferrite matrix. The formation of TM carbides such as MoC, Mo_2C, WC, W_2C, and VC_x ($x \approx 0.75$) occurs via precipitation and at much higher temperatures, $T \approx 500$ to 600°C, than that of cementite due to the much lower diffusivities in ferrite of these substitutional impurities as compared to that of C. This process, which can involve the conversion of cementite to TM carbides, is known as *secondary hardening* and is a type of age hardening.

Alloying elements such as Ni, Mn, and Si are often added to steels to make them heat treatable (i.e., to facilitate the heat treatment of austenite to produce martensite). This occurs because the formation of pearlite is retarded and so the desired martensite is more easily formed.

When the steel includes a high TM content (e.g., 18 to 25 wt % Ni along with Mo and Ti), particles of intermetallic compounds such as Ni_3Mo and Ni_3Ti can be formed via precipitation. Such materials are known as *maraging steels* and can have very high yield stresses, $\sigma_y \approx 2000$ MPa, along with good ductility and toughness.

The nucleation and growth of particles, often of a second phase, in a matrix is a recurrent theme in steels, especially in the discussion of dispersion-strengthening. This topic is also discussed in Section 21.5, where the Johnson–Mehl equation for the annealing and recrystallization (i.e., grain growth) of metals is discussed.

In addition to their uses in the strengthening processes just described, *heat treatments* of steels are used for a variety of other purposes. Various heat treatments are given to plain carbon steels containing pearlite in order to achieve the desired pearlite microstructures. As an example, *spheroidizing annealing* at just below T_e is used to transform the lamellar pearlite structure into one in which the pearlite takes on a spheroidal microstructure (i.e., the cementite lamellae have been spheroidized). This process leads to improved ductility and machinability of the steel. The driving force for this process is the reduction of the surface energy between the cementite and ferrite phases. This process is similar to the tempering of martensite discussed earlier, which, however, results in much smaller cementite particles, due to the lower temperatures used for tempering.

As just described, *tempering* is the term often used for the heat treatment or annealing of steels to achieve desired changes in microstructure and mechanical properties such as improved ductility. For example, the strength of martensite falls quickly and its ductility improves during tempering, due to the precipitation of C in carbides or carbon-containing intermetallic compounds. In contrast, tempering has little effect on bainite because there is not much C in solid solution. The effects of tempering on the mechanical properties of a steel are illustrated in Fig. W21.23. Similar behavior is observed for the tempering or annealing of nonferrous metals and alloys.

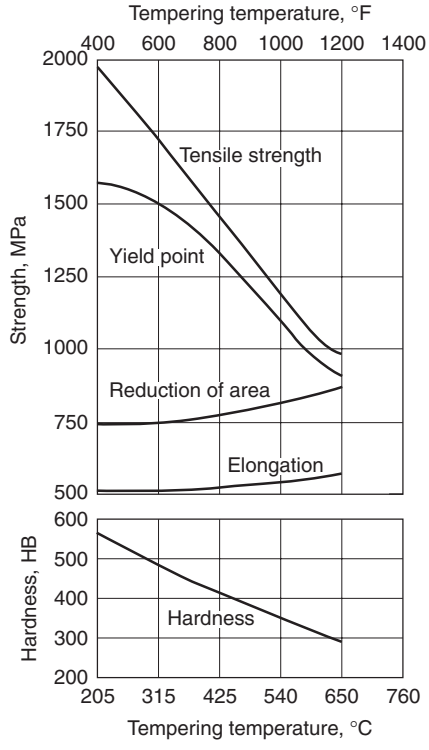

Figure W21.23. Effects of tempering at various temperatures on the mechanical properties (Brinell hardness, tensile and yield strengths, reduction of area, and elongation) for a 4340 steel bar. (From ASM Handbook, 9th ed., Vol. 4, *Heat Treating*, ASM International, Materials Park, Ohio, 1991, p. 123.)