28
Homogeneous Semiconductors

In Chapter 12 we observed that electrons in a completely filled band can carry no current. Within the independent electron model this result is the basis for the distinction between insulators and metals: In the ground state of an insulator all bands are either completely filled or completely empty; in the ground state of a metal at least one band is partially filled.

We can characterize insulators by the *energy gap*, E_g, between the top of the highest filled band(s) and the bottom of the lowest empty band(s) (see Figure 28.1). A solid with an energy gap will be nonconducting at $T = 0$ (unless the DC electric field is so strong and the energy gap so minute that electric breakdown can occur (Eq. (12.8)) or unless the AC field is of such high frequency that $\hbar\omega$ exceeds the energy gap).

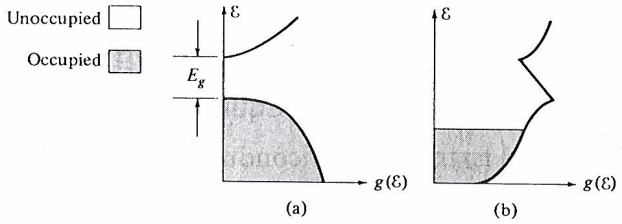

Figure 28.1
(a) In an insulator there is a region of forbidden energies separating the highest occupied and lowest unoccupied levels. (b) In a metal the boundary occurs in a region of allowed levels. This is indicated schematically by plotting the density of levels (horizontally) vs. energy (vertically).

However, when the temperature is not zero there is a nonvanishing probability that some electrons will be thermally excited across the energy gap into the lowest unoccupied bands, which are called, in this context, the *conduction bands*, leaving behind unoccupied levels in the highest occupied bands, called *valence bands*. The thermally excited electrons are capable of conducting, and hole-type conduction can occur in the band out of which they have been excited.

Whether such thermal excitation leads to appreciable conductivity depends critically on the size of the energy gap, for the fraction of electrons excited across the gap at temperature T is, as we shall see, roughly of order $e^{-E_g/2k_BT}$. With an energy gap of 4 eV at room temperature ($k_BT \approx 0.025$ eV) this factor is $e^{-80} \approx 10^{-35}$, and essentially no electrons are excited across the gap. If, however, E_g is 0.25 eV, then the factor at room temperature is $e^{-5} \approx 10^{-2}$, and observable conduction will occur.

Solids that are insulators at $T = 0$, but whose energy gaps are of such a size that thermal excitation can lead to observable conductivity at temperatures below the melting point, are known as *semiconductors*. Evidently the distinction between a semiconductor and an insulator is not a sharp one, but roughly speaking the energy gap in most important semiconductors is less than 2 eV and frequently as low as a few tenths of an electron volt. Typical room temperature resistivities of semiconductors are between 10^{-3} and 10^{9} ohm-cm (in contrast to metals, where $\rho \approx 10^{-6}$ ohm-cm, and good insulators, where ρ can be as large as 10^{22} ohm-cm).

Since the number of electrons excited thermally into the conduction band (and therefore the number of holes they leave behind in the valence band) varies exponentially with $1/T$, the electrical conductivity of a semiconductor should be a very rapidly

increasing function of temperature. This is in striking contrast to the case of metals. The conductivity of a metal (Eq. (1.6)),

$$\sigma = \frac{ne^2\tau}{m}, \tag{28.1}$$

declines with increasing temperature, for the density of carriers n is independent of temperature, and all temperature dependence comes from the relaxation time τ, which generally decreases with increasing temperature because of the increase in electron-phonon scattering. The relaxation time in a semiconductor will also decrease with increasing temperature, but this effect (typically described by a power law) is quite overwhelmed by the very much more rapid increase in the density of carriers with increasing temperature.[1]

Thus the most striking feature of semiconductors is that, unlike metals, their electrical resistance declines with rising temperature; i.e., they have a "negative coefficient of resistance." It was this property that first brought them to the attention of physicists in the early nineteenth century.[2] By the end of the nineteenth century a considerable body of semiconducting lore had been amassed; it was observed that the thermopowers of semiconductors were anomalously large compared with those of metals (by a factor of 100 or so), that semiconductors exhibited the phenomenon of photoconductivity, and that rectifying effects could be obtained at the junction of two unlike semiconductors. Early in the twentieth century, measurements of the Hall effect[3] were made confirming the fact that the temperature dependence of the conductivity was dominated by that of the number of carriers, and indicating that in many substances the sign of the dominant carrier was positive rather than negative.

Phenomena such as these were a source of considerable mystery until the full development of band theory many years later. Within the band theory they find simple explanations. For example, photoconductivity (the increase in conductivity produced by shining light on a material) is a consequence of the fact that if the band

---

[1] Thus the conductivity of a semiconductor is not a good measure of the collision rate, as it is in a metal. It is often advantageous to separate from the conductivity a term whose temperature dependence reflects only that of the collision rate. This is done by defining the *mobility*, μ, of a carrier, as being the ratio of the drift velocity it achieves in a field E, to the field strength: $v_d = \mu E$. If the carriers have density n and charge q, the current density will be $j = nqv_d$, and therefore the conductivity is related to the mobility by $\sigma = nq\mu$. The concept of mobility has little independent use in discussions of metals, since it is related to the conductivity by a temperature-independent constant. However, it plays an important role in descriptions of semiconductors (and any other conductors where the carrier density can vary, such as ionic solutions), enabling one to disentangle two distinct sources of temperature dependence in the conductivity. The usefulness of the mobility will be illustrated in our discussion of inhomogeneous semiconductors in Chapter 29.

[2] M. Faraday, *Experimental Researches on Electricity*, 1839, Facsimile Reprint by Taylor and Francis, London. R. A. Smith, *Semiconductors*, Cambridge University Press, 1964, provides one of the most pleasant introductions to the subject available. Most of the information in our brief historical survey is drawn from it.

[3] One might expect that the number of excited electrons would equal the number of holes left behind, so that the Hall effect would yield little direct information on the number of carriers. However, as we shall see, the number of electrons need not equal the number of holes in an impure semiconductor, and these were the only ones available at the time of the early experiments.

gap is small, then visible light can excite electrons across the gap into the conduction band, resulting in conduction by those electrons and by the holes left behind. The thermopower, to take another example, is roughly a hundred times larger in a semiconductor than in a metal. This is because the density of carriers is so low in a semiconductor that they are properly described by Maxwell-Boltzmann statistics (as we shall see below). Thus the factor of 100 is the same factor by which the early theories of metals (prior to Sommerfeld's introduction of Fermi-Dirac statistics) overestimated the thermopower (page 25).

The band theoretic explanations of these and other characteristic semiconducting properties will be the subject of this chapter and the next.

Compilation of reliable information on semiconductors in the early days was substantially impeded by the fact that data are enormously sensitive to the purity of the sample. An example of this is shown in Figure 28.2, where the resistivity of germanium is plotted vs. T for a variety of impurity concentrations. Note that concentrations as low as parts in 10^8 can lead to observable effects, and that the resistivity can vary at a given temperature by a factor of 10^{12}, as the impurity concentration changes by only a factor of 10^3. Note also that, for a given impurity concentration, the resistivity eventually falls onto a common curve as the temperature increases. This latter resistivity which is evidently the resistivity of an ideal perfectly pure sample, is known as the *intrinsic* resistivity, while the data for the various samples, except at temperatures so high that they agree with the intrinsic curve, are referred to as *extrinsic properties*. Quite generally, a semiconductor is intrinsic if its electronic properties are dominated by electrons thermally excited from the valence to the conduction band, and extrinsic if its electronic properties are dominated by electrons contributed to the conduction band by impurities (or captured from the valence band by impurities) in a manner to be described below. We shall return shortly to the question of why semiconducting properties are so very sensitive to the purity of the specimen.

EXAMPLES OF SEMICONDUCTORS

Semiconducting crystals come primarily from the covalent class of insulators.[4] The simple semiconducting elements are from column IV of the periodic table, silicon and germanium being the two most important elemental semiconductors. Carbon, in the form of diamond, is more properly classified as an insulator, since its energy gap is of order 5.5 eV. Tin, in the allotropic form of grey tin, is semiconducting, with a very small energy gap. (Lead, of course, is metallic.) The other semiconducting elements, red phosphorus, boron, selenium, and tellurium, tend to have highly complex crystal structures, characterized, however, by covalent bonding.

In addition to the semiconducting elements there is a variety of semiconducting compounds. One broad class, the III–V semiconductors, consists of crystals of the zincblende structure (page 81) composed of elements from columns III and V of the

[4] Among the various categories of insulating crystals, the covalent crystals have a spatial distribution of electronic charge most similar to metals. (See Chapter 19.)

Specimen	Donor concentration (cm^{-3})
1	5.3×10^{14}
2	9.3×10^{14}
5	1.6×10^{15}
7	2.3×10^{15}
8	3.0×10^{15}
10	5.2×10^{15}
12	8.5×10^{15}
15	1.3×10^{16}
17	2.4×10^{16}
18	3.5×10^{16}
20	4.5×10^{16}
21	5.5×10^{16}
22	6.4×10^{16}
23	7.4×10^{16}
24	8.4×10^{16}
25	1.2×10^{17}
26	1.3×10^{17}
27	2.7×10^{17}
29	9.5×10^{17}

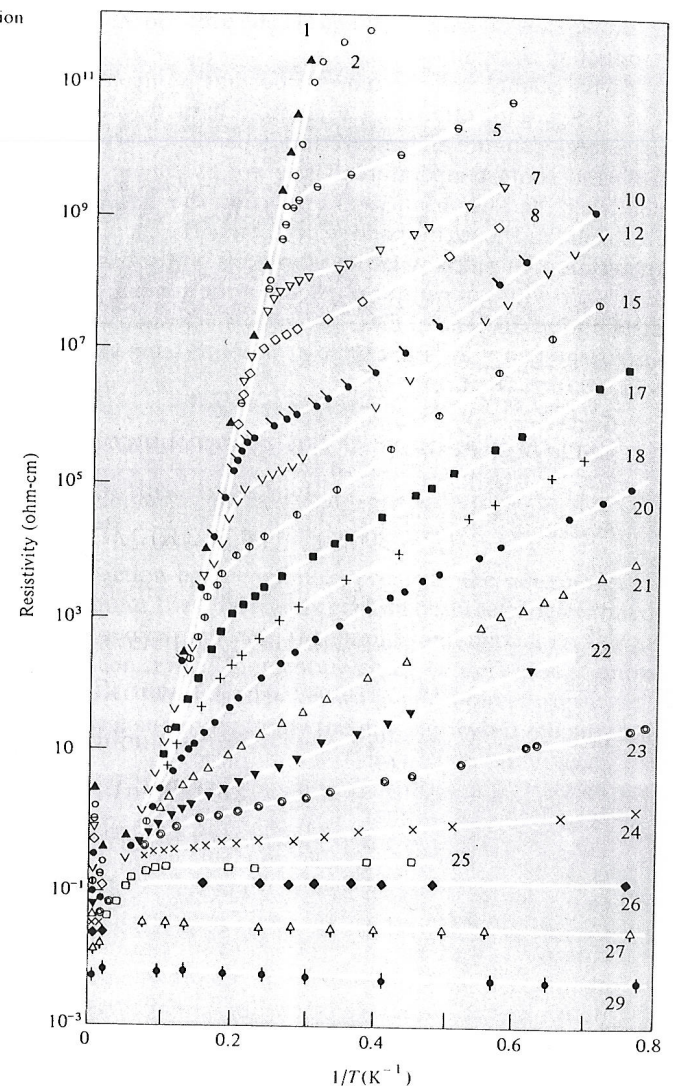

Figure 28.2
The resistivity of antimony-doped germanium as a function of $1/T$ for several impurity concentrations. (From H. J. Fritzsche, *J. Phys. Chem. Solids* **6**, 69 (1958).)

periodic table. As described in Chapter 19, the bonding in such compounds is also predominantly covalent. Semiconducting crystals made up of elements from columns II and VI begin to have a strong ionic as well as a covalent character. These are known as *polar semiconductors*, and can have either the zincblende structure or, as in the case of lead selenide, telluride, or sulfide, the sodium chloride structure more

characteristic of ionic bonding. There are also many far more complicated semiconducting compounds.

Some examples of the more important semiconductors are given in Table 28.1. The energy gaps quoted for each are reliable to within about 5 percent. Note that the energy gaps are all temperature-dependent, varying by about 10 percent between 0 K and room temperature. There are two main sources of this temperature dependence. Because of thermal expansion the periodic potential experienced by the electrons (and hence the band structure and the energy gap) can vary with temperature. In addition, the effect of lattice vibrations on the band structure and energy gap[5] will also vary with temperature, reflecting the temperature dependence of the phonon distribution. In general these two effects are of comparable importance, and lead to an energy gap that is linear in T at room temperature and quadratic at very low temperatures (Figure 28.3).

Table 28.1
ENERGY GAPS OF SELECTED SEMICONDUCTORS

MATERIAL	E_g (T = 300 K)	E_g (T = 0 K)	E_0 (LINEAR EXTRAPOLATION TO T = 0)	LINEAR DOWN TO
Si	1.12 eV	1.17	1.2	200 K
Ge	0.67	0.75	0.78	150
PbS	0.37	0.29	0.25	
PbSe	0.26	0.17	0.14	20
PbTe	0.29	0.19	0.17	
InSb	0.16	0.23	0.25	100
GaSb	0.69	0.79	0.80	75
AlSb	1.5	1.6	1.7	80
InAs	0.35	0.43	0.44	80
InP	1.3		1.4	80
GaAs	1.4		1.5	
GaP	2.2		2.4	
Grey Sn	0.1			
Grey Se	1.8			
Te	0.35			
B	1.5			
C (diamond)	5.5			

Sources: C. A. Hogarth, ed., *Materials Used in Semiconductor Devices*, Interscience, New York, 1965; O. Madelung, *Physics of III–V Compounds*, Wiley, New York, 1964; R. A. Smith, *Semiconductors*, Cambridge University Press, 1964.

The energy gap can be measured in several ways. The optical properties of the crystal are one of the most important sources of information. When the frequency of an incident photon becomes large enough for $\hbar\omega$ to exceed the energy gap, then, just as in metals (see pages 293, 294) there will be an abrupt increase in the absorption

[5] Via, for example, the kinds of effects described in Chapter 26.

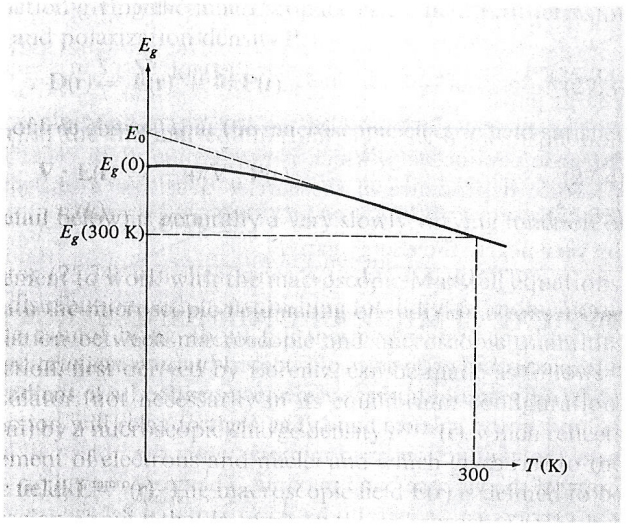

Figure 28.3
Typical temperature dependence of the energy gap of a semiconductor. Values of E_0, $E_g(0)$, and E_g(300 K) for several materials are listed in Table 28.1.

of incident radiation. If the conduction band minimum occurs at the same point in k-space as the valence band maximum, then the energy gap can be directly determined from the optical threshold. If, as is often the case, the minima and maxima occur at different points in k-space, then for crystal momentum to be conserved a phonon must also participate in the process,[6] which is then known as an "indirect transition" (Figure 28.4). Since the phonon will supply not only the missing crystal momentum

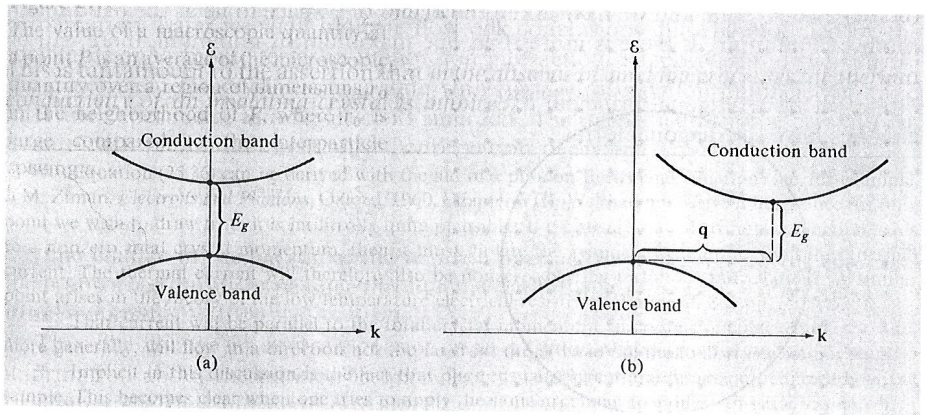

Figure 28.4
Photon absorption via (a) direct and (b) indirect transitions. In (a) the optical threshold is at $\omega = E_g/\hbar$; in (b) it occurs at $E_g/\hbar - \omega(\mathbf{q})$, since the phonon of wave vector $\mathbf{q}$ that must be absorbed to supply the missing crystal momentum also supplies an energy $\hbar\omega(\mathbf{q})$.

[6] At optical frequencies the crystal momentum supplied by the photon itself is negligibly small.

$\hbar\mathbf{k}$, but also an energy $\hbar\omega(\mathbf{k})$, the photon energy at the optical threshold will be less than E_g by an amount of order $\hbar\omega_D$. This is typically a few hundredths of an electron volt, and therefore of little consequence except in semiconductors with very small energy gaps.[7]

The energy gap may also be deduced from the temperature dependence of the intrinsic conductivity, which is predominantly a reflection of the very strong temperature dependence of the carrier densities. These vary (as we shall see below) essentially as $e^{-E_g/2k_BT}$, so that if $-\ln(\sigma)$ is plotted against $1/2k_BT$, the slope[8] should be very nearly the energy gap, E_g.

TYPICAL SEMICONDUCTOR BAND STRUCTURES

The electronic properties of semiconductors are completely determined by the comparatively small numbers of electrons excited into the conduction band and holes left behind in the valence band. The electrons will be found almost exclusively in levels near the conduction band minima, while the holes will be confined to the neighborhood of the valence band maxima. Therefore the energy vs. wave vector relations for the carriers can generally be approximated by the quadratic forms they assume in the neighborhood of such extrema:[9]

$$\mathcal{E}(\mathbf{k}) = \mathcal{E}_c + \frac{\hbar^2}{2}\sum_{\mu\nu} k_\mu(\mathbf{M}^{-1})_{\mu\nu}k_\nu \qquad \text{(electrons)},$$

$$\mathcal{E}(\mathbf{k}) = \mathcal{E}_v - \frac{\hbar^2}{2}\sum_{\mu\nu} k_\mu(\mathbf{M}^{-1})_{\mu\nu}k_\nu \qquad \text{(holes)}. \qquad \textbf{(28.2)}$$

Here $\mathcal{E}_c$ is the energy at the bottom of the conduction band, $\mathcal{E}_v$ is the energy at the top of the valence band, and we have taken the origin of k-space to lie at the band maximum or minimum. If there is more than one maximum or minimum, there will be one such term for each point. Since the tensor $\mathbf{M}^{-1}$ is real and symmetric, one can find a set of orthogonal principal axes for each such point, in terms of which the energies have the diagonal forms

$$\mathcal{E}(\mathbf{k}) = \mathcal{E}_c + \hbar^2\left(\frac{k_1^2}{2m_1} + \frac{k_2^2}{2m_2} + \frac{k_3^2}{2m_3}\right) \qquad \text{(electrons)},$$

$$\mathcal{E}(\mathbf{k}) = \mathcal{E}_v - \hbar^2\left(\frac{k_1^2}{2m_1} + \frac{k_2^2}{2m_2} + \frac{k_3^2}{2m_3}\right) \qquad \text{(holes)}. \qquad \textbf{(28.3)}$$

[7] To extract a really accurate band gap from the optical absorption data, however, it is necessary to determine the phonon spectrum and use it to analyze the indirect transitions.

[8] In deducing the energy gap in this way, however, one must remember that at room temperature the gaps of most semiconductors have a linear variation with temperature. If $E_g = E_0 - AT$, then the slope of the graph will be not E_g but E_0, the linear extrapolation of the room temperature gap to zero temperature (Figure 28.3). Values of E_0 extracted from this linear extrapolation procedure are also given in Table 28.1.

[9] The inverse of the matrix of coefficients in (28.2) is called $\mathbf{M}$ because it is a special case of the general effective mass tensor introduced on page 228. The electron mass tensor will not, of course, be the same as the hole mass tensor, but to avoid a multiplicity of subscripts we use the single generic symbol $\mathbf{M}$ for both.

Thus the constant energy surfaces about the extrema are ellipsoidal in shape, and are generally specified by giving the principal axes of the ellipsoids, the three "effective masses," and the location in k-space of the ellipsoids. Some important examples are:

Silicon The crystal has the diamond structure, so the first Brillouin zone is the truncated octahedron appropriate to a face-centered cubic Bravais lattice. The conduction band has six symmetry-related minima at points in the ⟨100⟩ directions, about 80 percent of the way to the zone boundary (Figure 28.5). By symmetry each

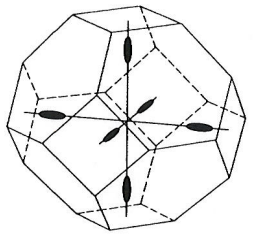

Figure 28.5
Constant-energy surfaces near the conduction band minima in silicon. There are six symmetry-related ellipsoidal pockets. The long axes are directed along ⟨100⟩ directions.

of the six ellipsoids must be an ellipsoid of revolution about a cube axis. They are quite cigar-shaped, being elongated along the cube axis. In terms of the free electron mass m, the effective mass along the axis (the longitudinal effective mass) is $m_L \approx 1.0m$ while the effective masses perpendicular to the axis (the transverse effective mass) are $m_T \approx 0.2m$. There are two degenerate valence band maxima, both located at $\mathbf{k} = \mathbf{0}$, which are spherically symmetric to the extent that the ellipsoidal expansion is valid, with masses of $0.49m$ and $0.16m$ (Figure 28.6).

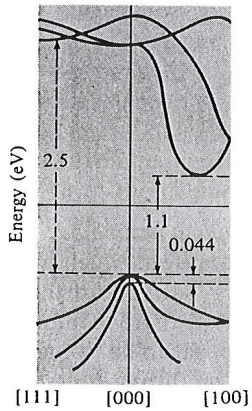

Figure 28.6
Energy bands in silicon. Note the conduction band minimum along [100] that gives rise to the ellipsoids of Figure 28.5. The valence band maximum occurs at $\mathbf{k} = 0$, where two degenerate bands with different curvatures meet, giving rise to "light holes" and "heavy holes." Note also, the third band, only 0.044 eV below the valence band maximum. This band is separated from the other two only by spin-orbit coupling. At temperatures on the order of room temperature ($k_BT = 0.025$ eV) it too may be a significant source of carriers. (From C. A. Hogarth, ed., *Materials Used in Semiconductor Devices*, Interscience, New York, 1965.)

Germanium The crystal structure and Brillouin zone are as in silicon. However, the conduction band minima now occur at the zone boundaries in the ⟨111⟩ directions. Minima on parallel hexagonal faces of the zone represent the same physical levels, so there are four symmetry-related conduction band minima. The ellipsoidal constant energy surfaces are ellipsoids of revolution elongated along the ⟨111⟩ directions, with effective masses $m_L \approx 1.6m$, and $m_T \approx 0.08m$ (Figure 28.7). There are again two

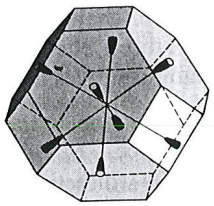

Figure 28.7
Constant-energy surfaces near the conduction band minima in germanium. There are eight symmetry-related half ellipsoids with long axes along ⟨111⟩ directions centered on the midpoints of the hexagonal zone faces. With a suitable choice of primitive cell in k-space these can be represented as four ellipsoids, the half ellipsoids on opposite faces being joined together by translations through suitable reciprocal lattice vectors.

degenerate valence bands, both with maxima at $\mathbf{k} = \mathbf{0}$, which are spherically symmetric in the quadratic approximation with effective masses of $0.28m$ and $0.044m$ (Figure 28.8).

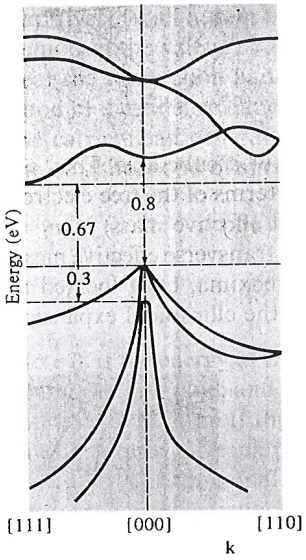

Figure 28.8
Energy bands in germanium. Note the conduction band minimum along [111] at the zone boundary that gives rise to the four ellipsoidal pockets of Figure 28.7. The valence band maximum, as in silicon, is at $\mathbf{k} = \mathbf{0}$, where two degenerate bands with different curvatures meet, giving rise to two pockets of holes with distinct effective masses. (From C. A. Hogarth, ed., *Materials Used in Semiconductor Devices*, Interscience, New York, 1965.)

Indium antimonide This compound, which has the zincblende structure, is interesting because all valence band maxima and conduction band minima are at $\mathbf{k} = 0$. The energy surfaces are therefore spherical. The conduction band effective mass is very small, $m^* \approx 0.015m$. Information on the valence band masses is less unambiguous, but there appear to be two spherical pockets about $\mathbf{k} = 0$, one with an effective mass of about $0.2m$ (heavy holes) and another with effective mass of about $0.015m$ (light holes).

CYCLOTRON RESONANCE

The effective masses discussed above are measured by the technique of cyclotron resonance. Consider an electron close enough to the bottom of the conduction band (or top of the valence band) for the quadratic expansion (28.2) to be valid. In the

presence of a magnetic field $\mathbf{H}$ the semiclassical equations of motion (12.32) and (12.33) imply that the velocity $\mathbf{v}(\mathbf{k})$ obeys the single set of equations

$$\mathbf{M}\frac{d\mathbf{v}}{dt} = \mp\frac{e}{c}\mathbf{v} \times \mathbf{H}. \tag{28.4}$$

In a constant uniform field (taken along the z-axis) it is not difficult to show (Problem 1) that (28.4) has an oscillatory solution

$$\mathbf{v} = \mathrm{Re}\,\mathbf{v}_0 e^{-i\omega t}, \tag{28.5}$$

provided that

$$\omega = \frac{eH}{m^* c}, \tag{28.6}$$

where m^*, the "cyclotron effective mass," is given by

$$m^* = \left(\frac{\det \mathbf{M}}{M_{zz}}\right)^{1/2}. \tag{28.7}$$

This result can also be written in terms of the eigenvalues and principal axes of the mass tensor as (Problem 1):

$$m^* = \sqrt{\frac{m_1 m_2 m_3}{\hat{H}_1^2 m_1 + \hat{H}_2^2 m_2 + \hat{H}_3^2 m_3}}, \tag{28.8}$$

where the $\hat{H}_i$ are the components along the three principal axes of a unit vector parallel to the field.

Note that the cyclotron frequency depends, for a given ellipsoid, on the orientation of the magnetic field with respect to that ellipsoid, but not on the initial wave vector or energy of the electron. Thus for a given orientation of the crystal with respect to the field, all electrons in a given ellipsoidal pocket of conduction band electrons (and, by the same token, all holes in a given ellipsoidal pocket of valence band holes) precess at a frequency entirely determined by the effective mass tensor describing that pocket. There will therefore be a small number of distinct cyclotron frequencies. By noting how these resonant frequencies shift as the orientation of the magnetic field is varied, one can extract from (28.8) the kind of information we quoted above.

To observe cyclotron resonance it is essential that the cyclotron frequency (28.6) be larger than or comparable to the collision frequency. As in the case of metals, this generally requires working with very pure samples at very low temperatures, to reduce both impurity scattering and phonon scattering to a minimum. Under such conditions the electrical conductivity of a semiconductor will be so small that (in contrast to the case of a metal (page 278)) the driving electromagnetic field can penetrate far enough into the sample to excite the resonance without any difficulties associated with a skin depth. On the other hand, under such conditions of low temperatures and purity the number of carriers available in thermal equilibrium to participate in the resonance may well be so small that carriers will have to be created by other means—such as photoexcitation. Some typical cyclotron resonance data are shown in Figure 28.9.

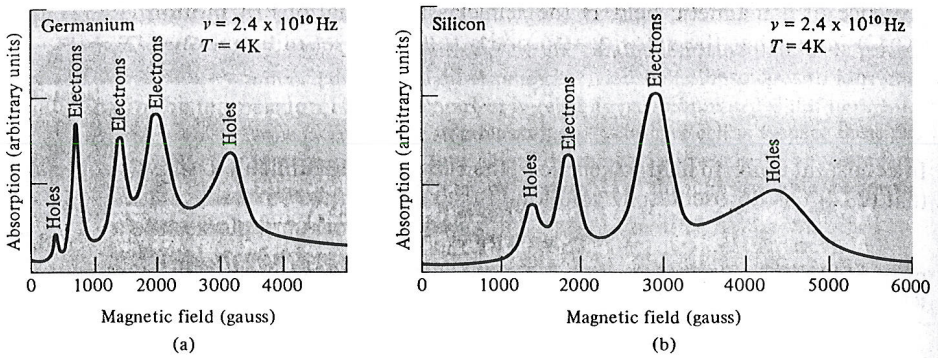

Figure 28.9
Typical cyclotron resonance signals in (a) germanium and (b) silicon. The field lies in a (110) plane and makes an angle with the [001] axis of 60° (Ge) and 30° (Si). (From G. Dresselhaus et al., *Phys. Rev.* **98**, 368 (1955).)

NUMBER OF CARRIERS IN THERMAL EQUILIBRIUM

The most important property of any semiconductor at temperature T is the number of electrons per unit volume in the conduction band, n_c, and the number of holes[10] per unit volume in the valence band, p_v. The determination of these as a function of temperature is a straightforward, though sometimes algebraically complicated, exercise in the application of Fermi-Dirac statistics to the appropriate set of one-electron levels.

The values of $n_c(T)$ and $p_v(T)$ depend critically, as we shall see, on the presence of impurities. However, there are certain general relations that hold regardless of the purity of the sample, and we consider these first. Suppose the density of levels (page 143) is $g_c(\mathcal{E})$ in the conduction band and $g_v(\mathcal{E})$ in the valence band. The effect of impurities, as we shall see below, is to introduce additional levels at energies between the top of the valence band, $\mathcal{E}_v$, and the bottom of the conduction band, $\mathcal{E}_c$, without, however, appreciably altering the form of $g_c(\mathcal{E})$ and $g_v(\mathcal{E})$. Since conduction is entirely due to electrons in conduction band levels or holes in valence band levels, regardless of the concentration of impurities the numbers of carriers present at temperature T will be given by

$$\begin{aligned} n_c(T) &= \int_{\mathcal{E}_c}^{\infty} d\mathcal{E}\, g_c(\mathcal{E}) \frac{1}{e^{(\mathcal{E}-\mu)/k_BT}+1}, \\ p_v(T) &= \int_{-\infty}^{\mathcal{E}_v} d\mathcal{E}\, g_v(\mathcal{E}) \left(1 - \frac{1}{e^{(\mathcal{E}-\mu)/k_BT}+1}\right) \\ &= \int_{-\infty}^{\mathcal{E}_v} d\mathcal{E}\, g_v(\mathcal{E}) \frac{1}{e^{(\mu-\mathcal{E})/k_BT}+1}. \end{aligned} \tag{28.9}$$

[10] Hole densities are conventionally denoted by the letter p (for positive). This widely used notation exploits the coincidence that the n denoting the *number* density of electrons can also be regarded as standing for "negative."

Impurities affect the determination of n_c and p_v only through the value of the chemical potential[11] μ to be used in Eq. (28.9). To determine μ one must know something about the impurity levels. However, one can extract some useful information from (28.9) which is independent of the precise value of the chemical potential, provided only that it satisfies the conditions:

$$\begin{aligned} \mathcal{E}_c - \mu &\gg k_BT, \\ \mu - \mathcal{E}_v &\gg k_BT. \end{aligned} \tag{28.10}$$

There will be a range of values of μ for which (28.10) holds even for energy gaps $E_g = \mathcal{E}_c - \mathcal{E}_v$ as small as a few tenths of an electron volt and temperatures as high as room temperature. Our procedure will be to assume the validity of (28.10), use it to simplify (28.9), and then, from the values of n_c and p_v so obtained and the appropriate information about possible impurity levels, compute the actual value of the chemical potential to check whether it does indeed lie in the range given by (28.10). If it does, the semiconductor is described as "nondegenerate," and the procedure is a valid one. If it does not, one is dealing with a "degenerate semiconductor" and must work directly with Eq. (28.9) without making the simplifications implied by (28.10).

Given Eq. (28.10), then since every conduction band level exceeds $\mathcal{E}_c$ and every valence band level is less than $\mathcal{E}_v$, we may simplify the statistical factors in (28.9):

$$\begin{aligned} \frac{1}{e^{(\mathcal{E}-\mu)/k_BT}+1} &\approx e^{-(\mathcal{E}-\mu)/k_BT}, \qquad \mathcal{E} > \mathcal{E}_c; \\ \frac{1}{e^{(\mu-\mathcal{E})/k_BT}+1} &\approx e^{-(\mu-\mathcal{E})/k_BT}, \qquad \mathcal{E} < \mathcal{E}_v. \end{aligned} \tag{28.11}$$

Equations (28.9) thereby reduce to

$$\boxed{\begin{aligned} n_c(T) &= N_c(T)e^{-(\mathcal{E}_c-\mu)/k_BT}, \\ p_v(T) &= P_v(T)e^{-(\mu-\mathcal{E}_v)/k_BT}, \end{aligned}} \tag{28.12}$$

where

$$\begin{aligned} N_c(T) &= \int_{\mathcal{E}_c}^{\infty} d\mathcal{E}\, g_c(\mathcal{E})e^{-(\mathcal{E}-\mathcal{E}_c)/k_BT}, \\ P_v(T) &= \int_{-\infty}^{\mathcal{E}_v} d\mathcal{E}\, g_v(\mathcal{E})e^{-(\mathcal{E}_v-\mathcal{E})/k_BT}. \end{aligned} \tag{28.13}$$

Because the ranges of integration in (28.13) include the points where the arguments of the exponentials vanish, $N_c(T)$ and $P_v(T)$ are relatively slowly varying functions

[11] It is the widespread practice to refer to the chemical potential of a semiconductor as "the Fermi level," a somewhat unfortunate terminology. Since the chemical potential almost always lies in the energy gap, there is no one-electron level whose energy is actually at "the Fermi level" (in contrast to the case of a metal). Thus the usual definition of the Fermi level (that energy below which the one-electron levels are occupied and above which they are unoccupied in the ground state of a metal) does not specify a unique energy in the case of a semiconductor: Any energy in the gap separates occupied from unoccupied levels at $T = 0$. The term "Fermi level" should be regarded as nothing more than a synonym for "chemical potential," in the context of semiconductors.

of temperature, compared with the exponential factors they multiply in (28.12). This is their most important feature. Usually, however, one can evaluate them explicitly. Because of the exponential factors in the integrands of (28.13) only energies within k_BT of the band edges contribute appreciably, and in this range the quadratic approximation, (28.2) or (28.3), is generally excellent. The level densities can then be taken to be (Problem 3):

$$g_{c,v}(\mathcal{E}) = \sqrt{2|\mathcal{E} - \mathcal{E}_{c,v}|}\,\frac{m_{c,v}^{3/2}}{\hbar^3\pi^2}, \tag{28.14}$$

and the integrals (28.13) then give

$$N_c(T) = \frac{1}{4}\left(\frac{2m_ck_BT}{\pi\hbar^2}\right)^{3/2},$$
$$P_v(T) = \frac{1}{4}\left(\frac{2m_vk_BT}{\pi\hbar^2}\right)^{3/2}. \tag{28.15}$$

Here m_c^3 is the product of the principal values of the conduction band effective mass tensor (i.e., its determinant),[12] and m_v^3 is similarly defined.

Equation (28.15) can be cast in the numerically convenient forms:

$$\boxed{\begin{aligned} N_c(T) &= 2.5\left(\frac{m_c}{m}\right)^{3/2}\left(\frac{T}{300\text{ K}}\right)^{3/2} \times 10^{19}/\text{cm}^3, \\ P_v(T) &= 2.5\left(\frac{m_v}{m}\right)^{3/2}\left(\frac{T}{300\text{ K}}\right)^{3/2} \times 10^{19}/\text{cm}^3, \end{aligned}} \tag{28.16}$$

where T is to be measured in degrees Kelvin. Since the exponential factors in (28.12) are less than unity by at least an order of magnitude, and since m_c/m and m_v/m are typically of the order of unity, Eq. (28.16) indicates that 10^{18} or 10^{19} carriers/cm^3 is an absolute upper limit to the carrier concentration in a nondegenerate semiconductor.

We still cannot infer $n_c(T)$ and $p_v(T)$ from (28.12) until we know the value of the chemical potential μ. However, the μ dependence disappears from the product of the two densities:

$$\boxed{\begin{aligned} n_cp_v &= N_cP_ve^{-(\mathcal{E}_c-\mathcal{E}_v)/k_BT} \\ &= N_cP_ve^{-E_g/k_BT}. \end{aligned}} \tag{28.17}$$

This result (sometimes called the "law of mass action"[13]) means that at a given temperature it suffices to know the density of one carrier type to determine that of the other. How this determination is made depends on how important the impurities are as a source of carriers.

[12] If there is more than one conduction band minimum one must add together terms of the form (28.14) and (28.15) for each minimum. These sums will continue to have the same forms as (28.14) and (28.15), provided that the definition of m_c is altered to $m_c^{3/2} \to \sum m_c^{3/2}$.

[13] The analogy with chemical reactions is quite precise: A carrier is provided by the dissociation of a combined electron and hole.

Intrinsic Case

If the crystal is so pure that impurities contribute negligibly to the carrier densities, one speaks of an "intrinsic semiconductor." In the intrinsic case, conduction band electrons can only have come from formerly occupied valence band levels, leaving holes behind them. The number of conduction band electrons is therefore equal to the number of valence band holes:

$$n_c(T) = p_v(T) \equiv n_i(T). \tag{28.18}$$

Since $n_c = p_v$, we may write their common value n_i as $(n_cp_v)^{1/2}$. Equation (28.17) then gives

$$n_i(T) = [N_c(T)P_v(T)]^{1/2}e^{-E_g/2k_BT}, \tag{28.19}$$

or, from (28.15) and (28.16):

$$\boxed{\begin{aligned} n_i(T) &= \frac{1}{4}\left(\frac{2k_BT}{\pi\hbar^2}\right)^{3/2}(m_cm_v)^{3/4}e^{-E_g/2k_BT} \\ &= 2.5\left(\frac{m_c}{m}\right)^{3/4}\left(\frac{m_v}{m}\right)^{3/4}\left(\frac{T}{300\text{ K}}\right)^{3/2}e^{-E_g/2k_BT} \times 10^{19}/\text{cm}^3. \end{aligned}} \tag{28.20}$$

We may now establish in the intrinsic case the condition for the validity of assumption (28.10) on which our analysis has been based. Defining μ_i to be the value of the chemical potential in the intrinsic case, we find that Eqs. (28.12) give values of n_c and p_v equal to n_i (Eq. (28.19)), provided that

$$\mu = \mu_i = \mathcal{E}_v + \tfrac{1}{2}E_g + \tfrac{1}{2}k_BT\ln\left(\frac{P_v}{N_c}\right), \tag{28.21}$$

or, from Eq. (28.15),

$$\boxed{\mu_i = \mathcal{E}_v + \tfrac{1}{2}E_g + \tfrac{3}{4}k_BT\ln\left(\frac{m_v}{m_c}\right).} \tag{28.22}$$

This asserts that as $T \to 0$, the chemical potential μ_i lies precisely in the middle of the energy gap. Furthermore, since $\ln(m_v/m_c)$ is a number of order unity, μ_i will not wander from the center of the energy gap by more than order k_BT. Consequently, at temperatures k_BT small compared with E_g, the chemical potential will be found far from the boundaries of the forbidden region, $\mathcal{E}_c$ and $\mathcal{E}_v$, compared with k_BT (Figure 28.10), and the condition for nondegeneracy (28.10) will be satisfied. Therefore (28.20) is a valid evaluation of the common value of n_c and p_v in the intrinsic case, provided only that E_g is large compared with k_BT, a condition that is satisfied in almost all semiconductors at room temperature and below.

Extrinsic Case: Some General Features

If impurities contribute a significant fraction of the conduction band electrons and/or valence band holes, one speaks of an "extrinsic semiconductor." Because of these

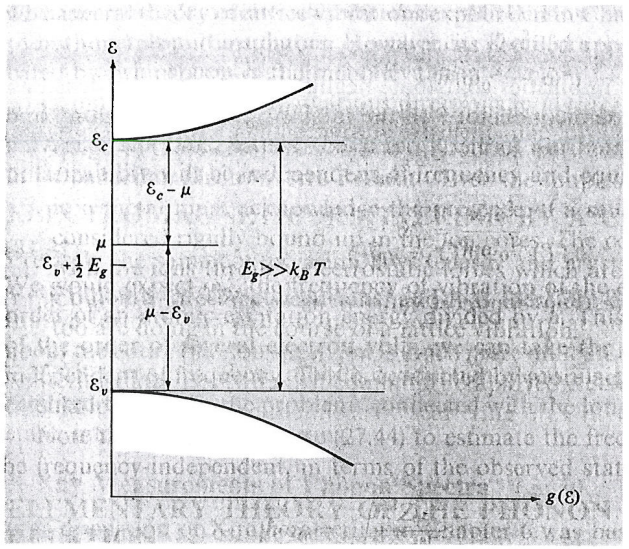

Figure 28.10
In an intrinsic semiconductor with an energy gap E_g large compared with k_BT, the chemical potential μ lies within order k_BT of the center of the energy gap, and is therefore far compared with k_BT from both boundaries of the gap at ε_c and ε_v.

added sources of carriers the density of conduction band electrons need no longer be equal to the density of valence band holes:

$$n_c - p_v = \Delta n \neq 0. \tag{28.23}$$

Since the law of mass action Eq. (28.17) holds regardless of the importance of impurities, we can use the definition (28.19) of $n_i(T)$ to write quite generally,

$$n_c p_v = n_i^2. \tag{28.24}$$

Equations (28.24) and (28.23) permit one to express the carrier densities in the extrinsic case in terms of their intrinsic values n_i and the deviation Δn from intrinsic behavior:

$$\begin{Bmatrix} n_c \\ p_v \end{Bmatrix} = \frac{1}{2}\left[(\Delta n)^2 + 4n_i^2\right]^{1/2} \pm \tfrac{1}{2}\Delta n. \tag{28.25}$$

The quantity $\Delta n/n_i$, which measures the importance of the impurities as a source of carriers, can be given a particularly simple expression as a function of chemical potential μ, if we note that Eqs. (28.12) have the form[14]

$$n_c = e^{\beta(\mu-\mu_i)}n_i; \quad p_v = e^{-\beta(\mu-\mu_i)}n_i. \tag{28.26}$$

Therefore

$$\frac{\Delta n}{n_i} = 2\sinh\beta(\mu - \mu_i). \tag{28.27}$$

[14] To verify these relations one need not substitute the explicit definitions of n_i and μ_i; it is enough to note that n_c and p_v are proportional to $\exp(\beta\mu)$ and $\exp(-\beta\mu)$, respectively, and that both reduce to n_i when $\mu = \mu_i$.

We have noted that if the energy gap E_g is large compared with k_BT, then the intrinsic chemical potential μ_i will satisfy the assumption (28.10) of nondegeneracy. But Eq. (28.27) requires that if μ_i is far from ε_c or ε_v on the scale of k_BT, then μ must be as well, unless Δn is many orders of magnitude larger than the intrinsic carrier density n_i. Thus the nondegeneracy assumption underlying the derivation of (28.27) is valid when $E_g \gg k_BT$, unless we are in a region of extreme extrinsic behavior.

Note also that when Δn is large compared with n_i, then Eq. (28.25) asserts that the density of one carrier type is essentially equal to Δn, while that of the other type is smaller by a factor of order $(n_i/\Delta n)^2$. Thus when impurities do provide the major source of carriers, one of the two carrier types will be dominant. An extrinsic semiconductor is called "n-type" or "p-type" according to whether the dominant carriers are electrons or holes.

To complete the specification of the carrier densities in extrinsic semiconductors one must determine Δn or μ. To do this we must examine the nature of the electronic levels introduced by the impurities and the statistical mechanics of the occupation of these levels in thermal equilibrium.

IMPURITY LEVELS

Impurities that contribute to the carrier density of a semiconductor are called *donors* if they supply additional electrons to the conduction band, and *acceptors* if they supply additional holes to (i.e., capture electrons from) the valence band. Donor impurities are atoms that have a higher chemical valence than the atoms making up the pure (host) material, while acceptors have a lower chemical valence.

Consider, for example, the case of substitutional impurities in a group IV semiconductor. Suppose that we take a crystal of pure germanium, and replace an occasional germanium atom by its neighbor to the right in the periodic table, arsenic (Figure 28.11). The germanium ion has charge $4e$ and contributes four valence electrons, while the arsenic ion has charge $5e$ and contributes five valence electrons. If, to a first approximation, we ignore the difference in structure between the arsenic and germanium ion cores, we can represent the substitution of an arsenic atom for

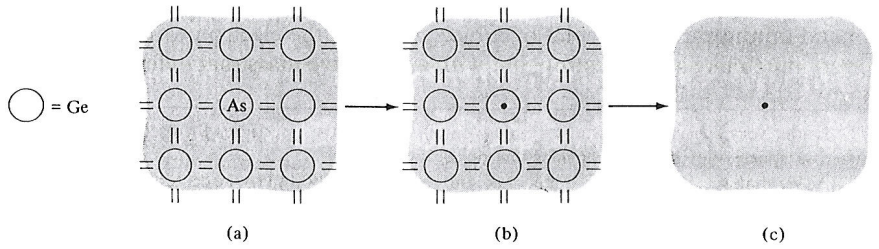

Figure 28.11
(a) Schematic representation of a substitutional arsenic (valence 5) donor impurity in a germanium (valence 4) crystal. (b) The arsenic (As) can be represented as a germanium atom *plus* an additional unit of positive charge fixed at the site of the atom (circled dot). (c) In the semiclassical approximation, in which the pure semiconductor is treated as a homogeneous medium, the arsenic impurity is represented as a fixed point charge $+e$ (dot).

a germanium atom by a slightly less drastic modification, in which the germanium atom is not removed, but an additional fixed positive charge of e is placed at its site, along with an additional electron.

This is the general model for a semiconductor doped with donor impurities. Distributed irregularly[15] throughout the perfect pure crystal are N_D fixed attractive centers of charge $+e$, per unit volume, along with the same number of additional electrons. As expected, each such center of charge $+e$ can bind[16] one of the additional electrons of charge $-e$. If the impurity were not embedded in the semiconductor, but in empty space, the binding energy of the electron would just be the first ionization potential of the impurity atom, 9.81 eV for arsenic. However (*and this is of crucial importance in the theory of semiconductors*), since the impurity is embedded in the medium of the pure semiconductor, this binding energy is enormously reduced (to 0.013 eV for the case of arsenic in germanium). This happens for two reasons:

1. The field of the charge representing the impurity must be reduced by the static dielectric constant ϵ of the semiconductor.[17] These are quite large ($\epsilon \approx 16$ in germanium), being typically between about 10 and 20 but ranging in some cases as high as 100 or more. The large dielectric constants are consequences of the small energy gaps. If there were no overall energy gap, the crystal would be a metal instead of a semiconductor, and the static dielectric constant would be infinite, reflecting the fact that a static electric field can induce a current in which electrons move arbitrarily far from their original positions. If the energy gap is not zero, but small, then the dielectric constant will not be infinite, but can be quite large, reflecting the relative ease with which the spatial distribution of electrons can be deformed.[18]
2. An electron moving in the medium of the semiconductor should be described not by the free space energy-momentum relation, but by the semiclassical relation (Chapter 12) $\mathcal{E}(\mathbf{k}) = \mathcal{E}_c(\mathbf{k})$, where $\hbar\mathbf{k}$ is the electronic crystal momentum, and $\mathcal{E}_c(\mathbf{k})$ is the conduction band energy-momentum relation; i.e., the additional electron introduced by the impurity should be thought of as being in a superposition of conduction band levels of the pure host material, which is appropriately altered by the additional localized charge $+e$ representing the impurity. The electron can minimize its energy by using only levels near the bottom of the conduction band, for which the quadratic approximation (28.2) is valid. Should the conduction band minimum be at a point of cubic symmetry, the electron would then behave very much like a free electron, but with an effective mass that differs from the free electron mass m. More generally, the energy wave vector relation will be some anisotropic quadratic function of k. In either case, however, to a first approximation, we may represent the electron as moving in free space but with a mass given by some appropriately defined effective mass m^*, rather than the free electron mass. In general, this mass will be smaller than the free electron mass, often by a factor of 0.1 or even less.

[15] Under very special circumstances it may be possible for the impurities themselves to be regularly arranged in space. We shall not consider this possibility here.

[16] As we shall see, the binding is quite weak, and the electrons bound to the center are readily liberated by thermal excitation.

[17] This use of macroscopic electrostatics in describing the binding of a single electron is justified by the fact (established below) that the wave function of the bound electron extends over many hundreds of angstroms.

[18] The connection between small energy gaps and large dielectric constants can also be understood from the point of view of perturbation theory: The size of the dielectric constant is a measure of the extent to which a weak electric field distorts the electronic wave function. But a small energy gap means there will be small energy denominators, and hence large changes, in the first-order wave functions.

These two observations suggest that we may represent an electron in the presence of a donor impurity of charge e within the medium of the semiconductor, as a particle of charge $-e$ and mass m^*, moving in free space in the presence of an attractive center of charge e/ϵ. This is precisely the problem of a hydrogen atom, except that the product $-e^2$ of the nuclear and electronic charges must be replaced by $-e^2/\epsilon$, and the free electron mass m, by m^*. Thus the radius of the first Bohr orbit, $a_0 = \hbar^2/me^2$, becomes

$$r_0 = \frac{m}{m^*}\epsilon a_0, \tag{28.28}$$

and the ground-state binding energy, $me^4/2\hbar^2 = 13.6$ eV becomes

$$\mathcal{E} = \frac{m^*}{m}\frac{1}{\epsilon^2} \times 13.6 \text{ eV}. \tag{28.29}$$

For reasonable values of m^*/m and ϵ, the radius r_0 can be 100 Å or more. This is very important for the consistency of the entire argument, for both the use of the semiclassical model and the use of the macroscopic dielectric constant are predicated on the assumption that the fields being described vary slowly on the scale of a lattice constant.

Furthermore, typical values of m^*/m and ϵ can lead to a binding energy $\mathcal{E}$ smaller than 13.6 eV by a factor of a thousand or more. Indeed, since small energy gaps are generally associated with large dielectric constants, it is almost always the case that *the binding energy of an electron to a donor impurity is small compared with the energy gap of the semiconductor.* Since this binding energy is measured relative to the energy of the conduction band levels from which the bound impurity level is formed, we conclude that donor impurities introduce additional electronic levels at energies $\mathcal{E}_d$ which are lower than the energy $\mathcal{E}_c$ at the bottom of the conduction band by an amount that is small compared with the energy gap E_g (Figure 28.12).

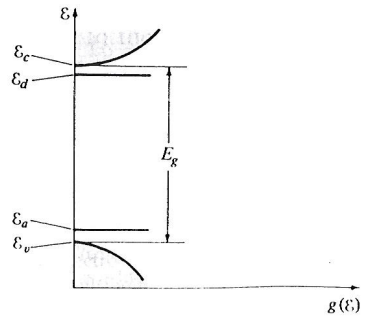

Figure 28.12
Level density for a semiconductor containing both donor and acceptor impurities. The donor levels $\mathcal{E}_d$ are generally close to the bottom of the conduction band, $\mathcal{E}_c$ compared with E_g, and the acceptor levels, $\mathcal{E}_a$, are generally close to the top of the valence band, $\mathcal{E}_v$.

A similar argument can be applied to acceptor impurities, whose valence is one less than that of the host atoms (e.g., gallium in germanium). Such an impurity can be represented by the superimposition of a fixed charge $-e$ on top of a host atom, along with the presence of one less electron in the crystal. The missing electron can be represented as a bound hole, attracted by the excess negative charge representing the impurity, with a binding energy that is again small[19] on the scale of the energy gap, E_g. In terms of the electron picture this bound hole will be manifested as an additional electronic level at an energy $\mathcal{E}_a$ lying slightly above the top of the valence band (Figure 28.12). The hole is bound when the level is empty. The binding energy of the hole is just the energy $\mathcal{E}_a - \mathcal{E}_v$ necessary to excite an electron from the top of the valence band into the acceptor level, thereby filling the hole in the vicinity of the acceptor and creating a free hole in the valence band.

Table 28.2
LEVELS OF GROUP V (DONORS) AND GROUP III (ACCEPTORS) IMPURITIES IN SILICON AND GERMANIUM

GROUP III ACCEPTORS (TABLE ENTRY IS $\mathcal{E}_a - \mathcal{E}_v$)	B	Al	Ga	In	Tl
Si	0.046 eV	0.057	0.065	0.16	0.26
Ge	0.0104	0.0102	0.0108	0.0112	0.01

GROUP V DONORS (TABLE ENTRY IS $\mathcal{E}_c - \mathcal{E}_d$)	P	As	Sb	Bi
Si	0.044 eV	0.049	0.039	0.069
Ge	0.0120	0.0127	0.0096	—

ROOM TEMPERATURE ENERGY GAPS ($E_g = \mathcal{E}_c - \mathcal{E}_v$)	
Si	1.12 eV
Ge	0.67 eV

Source: P. Aigrain and M. Balkanski, *Selected Constants Relative to Semiconductors*, Pergamon, New York, 1961.

The single most important fact about these donor and acceptor levels is that they lie very close to the boundaries of the forbidden energy region.[20] It is far easier thermally to excite an electron into the conduction band from a donor level, or a hole into the valence band from an acceptor level, than it is to excite an electron across the entire energy gap from valence to conduction band. Unless the concentration of donor and acceptor impurities is very small, they will therefore be a far more important source of carriers than the intrinsic mechanism of exciting carriers across the full gap.

[19] For the same reasons as in the case of donor impurities, the binding energy of the hole is quite weak; i.e., valence band electrons are readily lifted into the acceptor level by thermal excitation.

[20] Some measured donor and acceptor levels are given in Table 28.2.

POPULATION OF IMPURITY LEVELS IN THERMAL EQUILIBRIUM

To assess the extent to which carriers can be thermally excited from impurity levels, we must compute the mean number of electrons in the levels at a given temperature and chemical potential. We assume that the density of impurities is low enough that the interaction of electrons (or holes) bound at different impurity sites is negligible. We may then calculate the number density of electrons n_d (or holes p_a) bound to donor (or acceptor) sites by simply multiplying by the density of donors N_d (or acceptors N_a) the mean number of electrons (or holes) there would be if there were only a single impurity. For simplicity we assume that the impurity introduces only a single one-electron orbital level.[21] We calculate its mean occupancy as follows:

Donor Level If we ignored electron-electron interactions the level could either be empty, could contain one electron of either spin, or two electrons of opposite spins. However, the Coulomb repulsion of two localized electrons raises the energy of the doubly occupied level so high that double occupation is essentially prohibited. Quite generally, the mean number of electrons in a system in thermal equilibrium is given by:

$$\langle n \rangle = \frac{\sum N_j e^{-\beta(E_j - \mu N_j)}}{\sum e^{-\beta(E_j - \mu N_j)}}, \tag{28.30}$$

where the sum is over all states of the system, E_j and N_j, are the energy and number of electrons in state j, and μ is the chemical potential. In the present case the system is a single impurity with just three states: one with no electrons present which makes no contribution to the energy, and two with a single electron present of energy $\mathcal{E}_d$. Therefore (28.30) gives

$$\langle n \rangle = \frac{2e^{-\beta(\mathcal{E}_d - \mu)}}{1 + 2e^{-\beta(\mathcal{E}_d - \mu)}} = \frac{1}{\frac{1}{2}e^{\beta(\mathcal{E}_d - \mu)} + 1}, \tag{28.31}$$

so that[22]

$$\boxed{n_d = \frac{N_d}{\frac{1}{2}e^{\beta(\mathcal{E}_d - \mu)} + 1}}. \tag{28.32}$$

Acceptor Level In contrast to a donor level, an acceptor level, when viewed as an electronic level, can be singly or doubly occupied, but not empty. This is easily seen from the hole point of view. An acceptor impurity can be regarded as a fixed, negatively charged attractive center superimposed on an unaltered host atom. This additional charge $-e$ can weakly bind one hole (corresponding to one electron being in the

[21] There is no general reason why a donor site cannot have more than one bound level, and we assume a single one only to simplify our discussion. Our qualitative conclusions, however, are quite general (see Problem 4c).

[22] Some insight into the curious factor of $\frac{1}{2}$ that emerges in (28.32) in contrast to the more familiar distribution function of Fermi-Dirac statistics can be gained by examining what happens as the energy of the doubly occupied level drops from $+\infty$ down to $2\mathcal{E}_d$. See Problem 4.

acceptor level). The binding energy of the hole is $\mathcal{E}_a - \mathcal{E}_v$, and when the hole is "ionized" an additional electron moves into the acceptor level. However, the configuration in which no electrons are in the acceptor level corresponds to two holes being localized in the presence of the acceptor impurity, which has a very high energy due to the mutual Coulomb repulsion of the holes.[23]

Bearing this in mind, we can calculate the mean number of electrons at an acceptor level from (28.30) by noting that the state with no electrons is now prohibited, while the two-electron state has an energy that is $\mathcal{E}_a$ higher than the two one-electron states. Therefore

$$\langle n \rangle = \frac{2e^{\beta\mu} + 2e^{-\beta(\mathcal{E}_a - 2\mu)}}{2e^{\beta\mu} + e^{-\beta(\mathcal{E}_a - 2\mu)}} = \frac{e^{\beta(\mu - \mathcal{E}_a)} + 1}{\frac{1}{2}e^{\beta(\mu - \mathcal{E}_a)} + 1}. \tag{28.33}$$

The mean number of holes in the acceptor level is the difference between the maximum number of electrons the level can hold (two) and the actual mean number of electrons in the level ($\langle n \rangle$): $\langle p \rangle = 2 - \langle n \rangle$, and therefore $p_a = N_a \langle p \rangle$ is given by

$$\boxed{p_a = \frac{N_a}{\frac{1}{2}e^{\beta(\mu - \mathcal{E}_a)} + 1}.} \tag{28.34}$$

THERMAL EQUILIBRIUM CARRIER DENSITIES OF IMPURE SEMICONDUCTORS

Consider a semiconductor doped with N_d donor impurities and N_a acceptor impurities per unit volume. To determine the carrier densities we must generalize the constraint $n_c = p_v$ (Eq. (28.18)) that enabled us to find these densities in the intrinsic (pure) case. We can do this by first considering the electronic configuration at $T = 0$. Suppose $N_d \geq N_a$. (The case $N_d < N_a$ is equally straightforward and leads to the same result (28.35).) Then in a unit volume of semiconductor N_a of the N_d electrons supplied by the donor impurities can drop from the donor levels into the acceptor levels.[24] This gives a ground-state electronic configuration in which the valence band and acceptor levels are filled, $N_d - N_a$ of the donor levels are filled, and the conduction band levels are empty. In thermal equilibrium at temperature T the electrons will be redistributed among these levels, but since their total number remains the same, the number of electrons in conduction band or donor levels, $n_c + n_d$, must exceed its value at $T = 0$, $N_d - N_a$, by precisely the number of empty levels (i.e., holes), $p_v + p_a$, in the valence band and acceptor levels:

$$\boxed{n_c + n_d = N_d - N_a + p_v + p_a.} \tag{28.35}$$

[23] When describing acceptor levels as electronic levels one usually ignores the electron that *must* be in the level, considering only the presence or absence of the second electron. One describes the level as empty or filled according to whether the second electron is absent or present.

[24] Since $\mathcal{E}_d$ is just below the conduction band minimum $\mathcal{E}_c$, and $\mathcal{E}_a$ is just above the valence band maximum, $\mathcal{E}_v$, we have $\mathcal{E}_d > \mathcal{E}_a$ (see Figure 28.12).

This equation, together with the explicit forms we have found for n_c, p_v, n_d, and n_a as functions of μ and T, permits one to find μ as a function of T, and therefore to find the thermal equilibrium carrier densities at any temperature. A general analysis is rather complicated, and we consider here only a particularly simple and important case:

Suppose that

$$\begin{aligned} \mathcal{E}_d - \mu &\gg k_B T, \\ \mu - \mathcal{E}_a &\gg k_B T. \end{aligned} \tag{28.36}$$

Since $\mathcal{E}_d$ and $\mathcal{E}_a$ are close to the edges of the gap, this is only slightly more restrictive than the nondegeneracy assumption (28.10). Condition (28.36) and the expressions (28.32) and (28.34) for n_d and p_a insure that thermal excitation fully "ionizes" the impurities, leaving only a negligible fraction with bound electrons or holes: $n_d \ll N_d$, $p_a \ll N_a$. Equation (28.35) therefore becomes

$$\Delta n = n_c - p_v = N_d - N_a, \tag{28.37}$$

so Eqs. (28.25) and (28.27) now give the carrier densities and chemical potential as explicit functions of the temperature alone:

$$\begin{Bmatrix} n_c \\ p_v \end{Bmatrix} = \frac{1}{2}[(N_d - N_a)^2 + 4n_i^2]^{1/2} \pm \frac{1}{2}[N_d - N_a] \tag{28.38}$$

$$\frac{N_d - N_a}{n_i} = 2 \sinh \beta(\mu - \mu_i). \tag{28.39}$$

If the gap is large compared with $k_B T$, the assumption (28.36) we began with should remain valid unless μ is quite far from μ_i on the scale of $k_B T$. According to Eq. (28.39), this will only happen when $|N_d - N_a|$ is several orders of magnitude greater than the intrinsic carrier density n_i. Therefore Eq. (28.38) correctly describes the transition from predominantly intrinsic behavior ($n_i \gg |N_d - N_a|$) well into the region of predominantly extrinsic behavior ($n_i \ll |N_d - N_a|$). Expanding (28.38), we find that at low impurity concentrations the corrections to the purely intrinsic carrier densities are

$$\begin{Bmatrix} n_c \\ p_v \end{Bmatrix} \approx n_i \pm \tfrac{1}{2}(N_d - N_a), \tag{28.40}$$

while for a considerable range of carrier concentrations in the extrinsic regime,

$$\begin{aligned} n_c &\approx N_d - N_a \\ p_v &\approx \frac{n_i^2}{N_d - N_a} \end{aligned} \Bigg\} \quad N_d > N_a;$$

$$\begin{aligned} n_c &\approx \frac{n_i^2}{N_a - N_d} \\ p_v &\approx N_a - N_d \end{aligned} \Bigg\} \quad N_a > N_d. \tag{28.41}$$

Equation (28.41) is quite important in the theory of semiconducting devices (Chapter 29). It asserts that the net excess of electrons (or holes) $N_d - N_a$ introduced by the impurities is almost entirely donated to the conduction (or valence) band; the other band has the very much smaller carrier density $n_i^2/(N_d - N_a)$, as required by the law of mass action, (28.24).

If the temperature is too low (or the impurity concentration too high), condition (28.36) eventually fails to hold, and either n_d/N_d or p_a/N_a (but not both) ceases to be negligible, i.e., one of the impurity types is no longer fully ionized by thermal excitation. As a result, the dominant carrier density declines with decreasing temperature (Figure 28.13).[25]

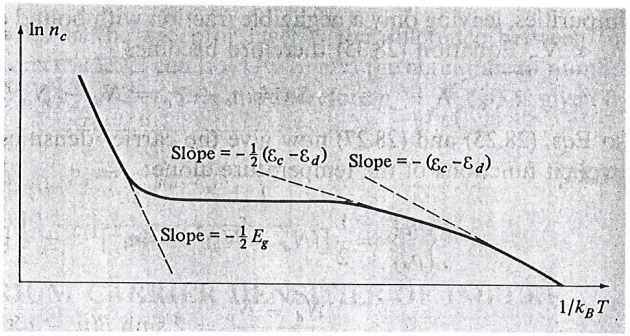

Figure 28.13
Temperature dependence of the majority carrier density (for the case $N_d > N_a$). The two high-temperature regimes are discussed in the text; the very low-temperature behavior is described in Problem 6.

IMPURITY BAND CONDUCTION

As the temperature approaches zero, so does the fraction of ionized impurities, and therefore also the density of carriers in the conduction or valence bands. Nevertheless, some small residual conductivity is observed even at the lowest temperatures. This is because the wave function of an electron (or hole) bound to an impurity site has considerable spatial extent, and therefore the overlap of wave functions at different impurity sites is possible even at fairly low concentrations. When this overlap is not negligible, it is possible for an electron to tunnel from one site to another. The resulting transport of charge is known as "impurity band conduction."

The use of the term "band" in this context is based on an analogy with the tight-binding method (Chapter 10), which shows that a set of atomic levels with a single energy can broaden into a band of energies, when wave function overlap is taken into account. The impurities, however, are usually not situated at the sites of a Bravais lattice, and one must therefore be cautious in attributing to the impurity "bands" features associated with electronic bands in *periodic* potentials.[26]

[25] This behavior is described more quantitatively in Problem 6.

[26] The problem of electronic behavior in aperiodic potentials (which arises not only in connection with impurity bands, but also, for example, in the case of disordered alloys) is still in its infancy, and is one of the very lively areas of current research in solid state physics.

THE THEORY OF TRANSPORT IN NONDEGENERATE SEMICONDUCTORS

It is a straightforward consequence (Problem 7) of Fermi-Dirac statistics and the nondegeneracy assumption (28.10) that the thermal equilibrium velocity distribution for electrons near a particular conduction band minimum (or holes near a particular valence band maximum) has the form:

$$f(\mathbf{v}) = n \frac{|\det \mathbf{M}|^{1/2}}{(2\pi k_B T)^{3/2}} \exp\left\{-\frac{\beta}{2}\sum_{\mu\nu} v_\mu \mathbf{M}_{\mu\nu} v_\nu\right\}, \tag{28.42}$$

where n is their contribution to the total carrier density.

This is just the form assumed by the thermal equilibrium molecular velocity distribution in a classical gas, with two exceptions:

1. In a classical gas, the density of molecules n is specified; in a semiconductor, n is an extremely sensitive function of temperature.
2. In a classical gas the mass tensor $\mathbf{M}$ is diagonal.

As a result, the theory of transport in a nondegenerate semiconductor is similar to the theory of transport in a classical gas of several charged components,[27] and many results of the classical theory can be applied directly to semiconductors, when allowance is made for the temperature dependence of the carrier densities and tensor character of the mass. For example, the anomalously high thermopower of a semiconductor (page 563) is only anomalous in comparison with the thermopower of metals; it is quite in accord with the properties of a classical charged gas. Indeed, the thermopower of metals was considered anomalously low in the early days of electron theory, before it was realized that metallic electrons must be described by Fermi-Dirac, rather than classical, statistics.

PROBLEMS

1. ***Cyclotron Resonance in Semiconductors***

(a) Show that the formulas (28.6) and (28.7) for the cyclotron resonance frequency follow from substituting the oscillatory velocity (28.5) into the semiclassical equation of motion (28.4), and requiring that the resulting homogeneous equation have a nonzero solution.

(b) Show that (28.7) and (28.8) are equivalent representations of the cyclotron effective mass by evaluating (28.7) in the coordinate system in which the mass tensor $\mathbf{M}$ is diagonal.

2. ***Interpretation of Cyclotron Resonance Data***

(a) Compare the cyclotron resonance signal from silicon in Figure 28.9b with the geometry of the conduction band ellipsoids shown in Figure 28.5, and explain why there are only two electron peaks although there are six pockets of electrons.

[27] Such a theory was extensively developed by Lorentz, as an attempt at refining the Drude model of metals. Although Lorentz's theory requires substantial modification to be applicable to metals (i.e., the introduction of degenerate Fermi-Dirac statistics and band structure), many of his results can be applied to the description of nondegenerate semiconductors with very little alteration.

(b) Verify that the positions of the electron resonances in Figure 28.9b are consistent with the electron effective masses given for silicon on page 569 and the formulas, (28.6) and (28.8), for the resonance frequency.

(c) Repeat (a) for the resonance in germanium (Figure 28.9a), noting that Figure 28.7 shows four electron pockets.

(d) Verify that the positions of the electron resonances in Figure 28.9a are consistent with the electron effective masses given for germanium on page 569.

3. *Level Density for Ellipsoidal Pockets*

(a) Show that the contribution of an ellipsoidal pocket of electrons to the conduction band density of levels $g_c(\mathcal{E})$, is given by $(d/d\mathcal{E})h(\mathcal{E})$, where $h(\mathcal{E})$ is the number of levels per unit volume in the pocket with energies less than $\mathcal{E}$.

(b) Show, similarly, that the contribution of an ellipsoidal pocket of holes to the valence band density of levels $g_v(\mathcal{E})$ is given by $(d/d\mathcal{E})h(\mathcal{E})$, where $h(\mathcal{E})$ is the number of electronic levels per unit volume in the pocket with energies greater than $\mathcal{E}$.

(c) Using the fact that a volume Ω of k-space contains $\Omega/4\pi^3$ electronic levels per cubic centimeter and the formula $V = (4\pi/3)abc$ for the volume of the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$, show that formulas (28.14) follow directly from (a) and (b), when the conduction (or valence) band has a single ellipsoidal pocket.

4. *Statistics of Donor Levels*

(a) Show that if the energy of a doubly occupied donor level is taken to be $2\mathcal{E}_d + \Delta$, then Eq. (28.32) must be replaced by

$$n_d = N_d \frac{1 + e^{-\beta(\mathcal{E}_d - \mu + \Delta)}}{\frac{1}{2}e^{\beta(\mathcal{E}_d - \mu)} + 1 + \frac{1}{2}e^{-\beta(\mathcal{E}_d - \mu + \Delta)}}. \tag{28.43}$$

(b) Verify that Eq. (28.43) reduces to (28.32) as $\Delta \to \infty$, and that it reduces to the expected result for independent electrons as $\Delta \to 0$.

(c) Consider a donor impurity with many bound electronic orbital levels, with energies $\mathcal{E}_i$. Assuming that the electron-electron Coulomb repulsion prohibits more than a single electron from being bound to the impurity, show that the appropriate generalization of (28.32) is

$$\frac{N_d}{1 + \frac{1}{2}(\sum e^{-\beta(\mathcal{E}_i - \mu)})^{-1}} \tag{28.44}$$

Indicate how (if at all) this alters the results described on pages 582–584.

5. *Constraint on Carrier Densities in p-Type Semiconductors*

Describe the electronic configuration of a doped semiconductor as $T \to 0$, when $N_a > N_d$. Explain why (28.35) (derived in the text when $N_d \geq N_a$) continues to give a correct constraint on the electron and hole densities at nonzero temperatures, when $N_a > N_d$.

6. *Carrier Statistics in Doped Semiconductors at Low Temperatures*

Consider a doped semiconductor with $N_d > N_a$. Assume that the nondegeneracy condition (28.10) holds, but that $(N_d - N_a)/n_i$ is so large that (28.39) does not necessarily yield a value of μ compatible with (28.36).

(a) Show under these conditions that p_v is negligible compared with n_c, and p_a is negligible compared with N_a, so that the chemical potential is given by the quadratic equation

$$N_c e^{-\beta(\mathcal{E}_c - \mu)} = N_d - N_a - \frac{N_d}{\frac{1}{2}e^{\beta(\mathcal{E}_d - \mu)} + 1}. \tag{28.45}$$

(b) Deduce from this that if the temperature drops so low that n_c ceases to be given by $N_d - N_a$ (Eq. (28.41)), then there is a transition to a regime in which

$$n_c = \sqrt{\frac{N_c(N_d - N_a)}{2}}\, e^{-\beta(\mathcal{E}_c - \mathcal{E}_d)/2}. \tag{28.46}$$

(c) Show that as the temperature drops still lower, there is another transition to a regime in which

$$n_c = \frac{N_c(N_d - N_a)}{N_a} e^{-\beta(\mathcal{E}_c - \mathcal{E}_d)}. \tag{28.47}$$

(d) Derive the results analogous to (28.45)–(28.47) when $N_a > N_d$.

7. *Velocity Distribution for Carriers in an Ellipsoidal Pocket*

Derive the velocity distribution (28.42) from the **k**-space distribution function

$$f(\mathbf{k}) \propto \frac{1}{e^{\beta(\mathcal{E}(\mathbf{k}) - \mu)} + 1}, \tag{28.48}$$

by assuming the nondegeneracy condition (28.10), changing from the variable **k** to the variable **v**, and noting that the contribution of the pocket to the carrier density is just $n = \int d\mathbf{v}\, f(\mathbf{v})$.