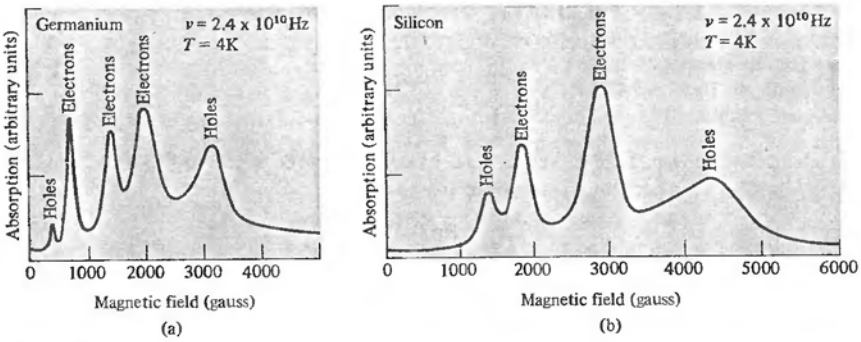

Figure 28.9
Typical cyclotron resonance signals in (a) germanium and (b) silicon. The field lies in a (110) plane and makes an angle with the [001] axis of 60° (Ge) and 30° (Si). (From G. Dresselhaus et al., *Phys. Rev.* 98, 368 (1955).)

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 $g_c(\varepsilon)$ in the conduction band and $g_v(\varepsilon)$ 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, ε_v, and the bottom of the conduction band, ε_c, without, however, appreciably altering the form of $g_c(\varepsilon)$ and $g_v(\varepsilon)$. 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

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

[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:

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

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):

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

Equations (28.9) thereby reduce to

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

where

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

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 $k_B T$ 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

$$\begin{aligned} N_c(T) &= \frac{1}{4}\left(\frac{2m_c k_B T}{\pi\hbar^2}\right)^{3/2}, \\ P_v(T) &= \frac{1}{4}\left(\frac{2m_v k_B T}{\pi\hbar^2}\right)^{3/2}. \end{aligned} \tag{28.15}$$

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:

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

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:

$$\boxed{\begin{aligned} n_c p_v &= N_c P_v e^{-(\mathcal{E}_c - \mathcal{E}_v)/k_B T} \\ &= N_c P_v e^{-E_g/k_B T}. \end{aligned}} \tag{28.17}$$

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 $(n_c p_v)^{1/2}$. 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):

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

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

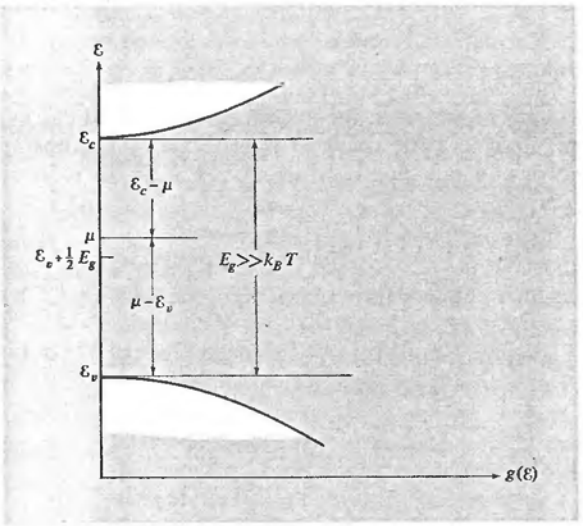

Figure 28.10
In an intrinsic semiconductor with an energy gap E_g large compared with k_BT, the chemical potential μ lies within order k_BT of the center of the energy gap, and is therefore far compared with k_BT from both boundaries of the gap at ε_c and ε_v.

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:

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

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

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

[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

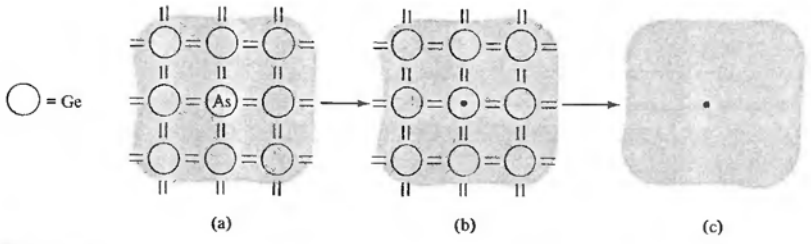

Figure 28.11
(a) Schematic representation of a substitutional arsenic (valence 5) donor impurity in a germanium (valence 4) crystal. (b) The arsenic (As) can be represented as a germanium atom *plus* an additional unit of positive charge fixed at the site of the atom (circled dot). (c) In the semiclassical 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

---

[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.

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.

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).

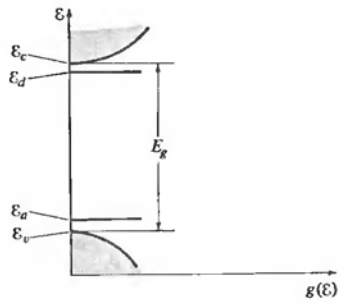

Figure 28.12
Level density for a semiconductor containing both donor and acceptor impurities. The donor levels $\mathcal{E}_d$ are generally close to the bottom of the conduction band, $\mathcal{E}_c$ compared with E_g, and the acceptor levels, $\mathcal{E}_a$, are generally close to the top of the valence 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 ε_d. Therefore (28.30) gives

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

so that[22]

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

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 $2\varepsilon_d$. See Problem 4.

acceptor level). The binding energy of the hole is $\varepsilon_a - \varepsilon_v$, 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 ε_a higher than the two one-electron states. Therefore

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

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

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

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 $N_d \geqslant N_a$. (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 ε_d is just below the conduction band minimum ε_c, and ε_a is just above the valence band maximum, ε_v, we have $\varepsilon_d > \varepsilon_a$ (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,

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

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

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]

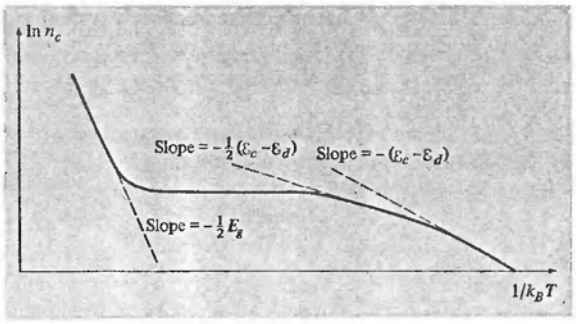

Figure 28.13
Temperature dependence of the majority carrier density (for the case $N_d > N_a$). The two high-temperature regimes are discussed in the text; the very low-temperature behavior is described in Problem 6.

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

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

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 $\mathbf{k}$ to the variable $\mathbf{v}$, and noting that the contribution of the pocket to the carrier density is just $n = \int d\mathbf{v}\, f(\mathbf{v})$.

29
Inhomogeneous Semiconductors

The Semiclassical Treatment of Inhomogeneous Solids

Fields and Carrier Densities in the Equilibrium *p-n* Junction

Elementary Picture of Rectification by a *p-n* Junction

Drift and Diffusion Currents

Collision and Recombination Times

Fields, Carrier Densities, and Currents in the Nonequilibrium *p-n* Junction

As used by music lovers, the term "solid state physics" refers only to the subject of inhomogeneous semiconductors, and it would be more accurate if it were this latter phrase that festooned the brows of countless tuners and amplifiers. The prevailing usage reflects the fact that modern solid state physics has had its most dramatic and extensive technological consequences through the electronic properties of semiconducting devices. These devices use semiconducting crystals in which the concentrations of donor and acceptor impurities have been made nonuniform in a carefully controlled manner. We shall not attempt to survey here the great variety of semiconducting devices, but will only describe the broad physical principles that underlie their operation. These principles come into play in determining how the densities and currents of electrons and holes are distributed in an inhomogeneous semiconductor, both in the absence and in the presence of an applied electrostatic potential.

The inhomogeneous semiconductors of interest are, ideally, single crystals in which the local concentration of donor and acceptor impurities varies with position. One way to make such crystals is to vary the concentration of impurities in the "melt" as the growing crystal is slowly extracted, thus producing a variation in impurity concentration along one spatial direction. Delicate methods of fabrication are needed because it is generally important, for efficient operation, that there be no great increase in electronic scattering associated with the variation in impurity concentration.

We shall illustrate the physics of inhomogeneous semiconductors by considering the simplest example, the *p-n* junction. This is a semiconducting crystal in which the impurity concentration varies only along a given direction (taken to be the x-axis) and only in a small region (taken to be around $x = 0$). For negative x the crystal has a preponderance of acceptor impurities (i.e., it is p-type) while for positive x it has a preponderance of donor impurities (i.e., it is n-type) (Figure 29.1). The manner in which the densities of donors and acceptors $N_d(x)$ and $N_a(x)$ vary with position is

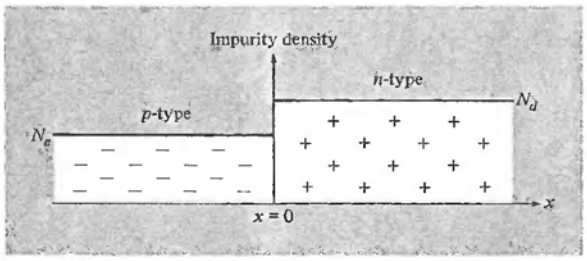

Figure 29.1
The impurity densities along a *p-n* junction in the case of an "abrupt junction," for which donor impurities dominate at positive x, and acceptor impurities at negative x. The donors are represented by (+) to indicate their charge when ionized, and the acceptors by (−). For a junction to be abrupt, the region about $x = 0$ where the impurity concentrations change must be narrow compared with the "depletion layer" in which the carrier densities are nonuniform. (Typical plots of the carrier densities are superimposed on this figure in Figure 29.3.)

called the "doping profile." The term "junction" is used to refer both to the device as a whole and, more specifically, to the transition region about $x = 0$ in which the doping profile is nonuniform.

As we shall see below, the nonuniformity in impurity concentrations induces a nonuniformity in the densities $n_c(x)$ and $p_v(x)$ of conduction band electrons and valence band holes, which in turn gives rise to a potential $\phi(x)$. The region in which these carrier densities are nonuniform is known as the "depletion layer" (or "space-charge region"). The depletion layer can extend for a range of about 10^2 to 10^4 Å around the (generally more narrow) transition region in which the doping profile varies, as we shall see below. Within the depletion layer, except near its boundaries, the total density of carriers is very much less than it is in the homogeneous regions farther away from the transition region. The existence of a depletion layer is one of the crucial properties of the p-n junction. One of our main concerns will be to explain why such a layer is induced by the variation in impurity concentrations, and how its structure changes with the application of an external potential V.

For simplicity we shall consider here only "abrupt junctions," in which the transition region is so sharp that variation in impurity concentrations[1] can be represented by a single discontinuous change at $x = 0$:

$$N_d(x) = \begin{cases} N_d, & x > 0 \\ 0, & x < 0 \end{cases},$$
$$N_a(x) = \begin{cases} 0, & x > 0 \\ N_a, & x < 0 \end{cases}. \tag{29.1}$$

Abrupt junctions are not only conceptually the simplest, but also the type of greatest practical interest. How sharp the actual transition region must be made for (29.1) to give a reasonable model of a physical junction will emerge in the analysis below. We shall find that a junction may be regarded as abrupt if the transition region in the actual doping profile is small in extent compared with the depletion layer. In most cases this permits the transition region to extend for 100 Å or more. A junction that cannot be treated as abrupt is called a "graded junction."

THE SEMICLASSICAL MODEL

To calculate the response of an inhomogeneous semiconductor to an applied electrostatic potential, or even to compute the distribution of electric charge in the absence of an applied potential, one almost always uses the semiclassical model of Chapter 12. When a potential $\phi(x)$ is superimposed on the periodic potential of the crystal, the semiclassical model treats the electrons in the nth band as classical particles (i.e., as wave packets) governed by the Hamiltonian

$$H_n = \mathcal{E}_n\left(\frac{\mathbf{p}}{\hbar}\right) - e\phi(x). \tag{29.2}$$

[1] It is not essential that there be only donor impurities in the n-type region and only acceptor impurities in the p-type region. It suffices for each impurity type to be the dominant one in its own region. In what follows N_d may be viewed as the excess density of donors over acceptors and N_a as the excess density of acceptors over donors.

Such a treatment is valid provided that the potential $\phi(x)$ varies sufficiently slowly. How slow this variation must be is, in general, a very difficult question to answer. At the very least, one requires that the change in electrostatic energy $e\Delta\phi$ over a distance of the order of the lattice constant be small compared with the band gap E_g, but the condition may well be even more stringent than this.[2] In the case of the p-n junction the potential ϕ has almost all of its spatial variation within the depletion layer. There, as we shall see, the energy $e\phi$ changes by about E_g, over a distance that is typically a few hundred angstroms or more (so that the field in the depletion layer can be as large as 10^6 volts per meter). Although this satisfies the minimum necessary condition for the validity of the semiclassical model (the change in $e\phi$ over a lattice constant is no more than a fraction of a percent of E_g), the variation is strong enough that one cannot exclude the possibility that the semiclassical description may break down in the depletion layer. Thus one should bear in mind the possibility that the field in the depletion layer may be strong enough to induce tunneling of electrons from valence band to conduction band levels, leading to a conductivity considerably in excess of the semiclassical prediction.

Having issued this warning, however, we shall follow the general practice of assuming the validity of the semiclassical description, so that we may explore its consequences. Before describing the semiclassical theory of the currents that flow in a p-n junction in the presence of an applied potential, we first examine the case of the p-n junction in thermal equilibrium, in the absence of applied potentials and current flow.

THE p-n JUNCTION IN EQUILIBRIUM

We wish to determine the carrier densities and the electrostatic potential $\phi(x)$ induced by the nonuniform doping. We assume that nondegenerate conditions hold throughout the material, so that the carrier densities at each position x have the "Maxwellian" forms analogous to the densities (28.12) we found in the uniform case. In the nonuniform case, to derive the carrier densities at position x along the junction in the presence of a potential $\phi(x)$, the semiclassical procedure is simply to repeat the analysis for the uniform case, but using the semiclassical one-electron energy (29.2), in which each level is shifted by $-e\phi(x)$. Using the forms (28.3) of the $\mathcal{E}(\mathbf{k})$ appropriate to levels near the conduction band minimum or valence band maximum, we see that the effect of this is simply to shift the constants $\mathcal{E}_c$ and $\mathcal{E}_v$ by $-e\phi(x)$. Thus Eq. (28.12) for the equilibrium carrier densities is generalized to

$$n_c(x) = N_c(T) \exp\left\{-\frac{[\mathcal{E}_c - e\phi(x) - \mu]}{k_B T}\right\},$$
$$p_v(x) = P_v(T) \exp\left\{-\frac{[\mu - \mathcal{E}_v + e\phi(x)]}{k_B T}\right\}. \tag{29.3}$$

The potential $\phi(x)$ must be determined self-consistently, as that potential arising (via Poisson's equation) when the carrier densities have the forms (29.3). We examine

[2] A crude argument appropriate to metals is given in Appendix J. Analogous arguments (of comparable crudity) can be developed for semiconductors.

this problem in the special case (again, the case of major practical interest) in which far from the transition region on either side extrinsic conditions prevail, in which the impurities are fully "ionized" (see pages 583–584). Thus far away on the n-side the density of conduction band electrons is very nearly equal to the density N_d of donors, while far away on the p-side the density of valence band holes is very nearly equal to the density N_a of acceptors:

$$N_d = n_e(\infty) = N_c(T) \exp\left\{-\frac{[\mathcal{E}_c - e\phi(\infty) - \mu]}{k_B T}\right\},$$
$$N_a = p_v(-\infty) = P_v(T) \exp\left\{-\frac{[\mu - \mathcal{E}_v + e\phi(-\infty)]}{k_B T}\right\}. \tag{29.4}$$

Since the entire crystal is in thermal equilibrium, the chemical potential does not vary with position. In particular, the same value of μ appears in either of Eqs. (29.4); this immediately requires that the total potential drop across the junction be given by[3]

$$e\phi(\infty) - e\phi(-\infty) = \mathcal{E}_c - \mathcal{E}_v + k_B T \ln\left[\frac{N_d N_a}{N_c P_v}\right], \tag{29.5}$$

or

$$\boxed{e\,\Delta\phi = E_g + k_B T \ln\left[\frac{N_d N_a}{N_c P_v}\right].} \tag{29.6}$$

An alternative way of representing the information in (29.3) and (29.6) is sometimes helpful. If we define a position-dependent "electrochemical potential" $\mu_e(x)$ by

$$\mu_e(x) = \mu + e\phi(x), \tag{29.7}$$

then we may write the carrier densities (29.3) as

$$n_e(x) = N_c(T) \exp\left\{-\frac{[\mathcal{E}_c - \mu_e(x)]}{k_B T}\right\},$$
$$p_v(x) = P_v(T) \exp\left\{-\frac{[\mu_e(x) - \mathcal{E}_v]}{k_B T}\right\}. \tag{29.8}$$

These have precisely the form of the relations (28.12) for a homogeneous semiconductor, except that the constant chemical potential μ is replaced by the electrochemical potential $\mu_e(x)$. Thus $\mu_e(\infty)$ is the chemical potential of a homogeneous n-type crystal whose properties are identical to those of the inhomogeneous crystal far on the n-side of the transition region, while $\mu_e(-\infty)$ is the chemical potential of a homogeneous p-type crystal identical to the inhomogeneous crystal far on the p-side. The relation (29.6) can equally well be written as[4]

$$e\,\Delta\phi = \mu_e(\infty) - \mu_e(-\infty). \tag{29.9}$$

---

[3] The derivation of (29.5) requires the validity of (29.3) only far from the depletion layer, where ϕ is indeed slowly varying. It therefore holds even when the semiclassical model fails in the transition region.

[4] This follows directly from (29.7). Equation (29.9) is sometimes summarized in the rule that the total potential drop is such as to bring the "Fermi levels at the two ends of the junction" into coincidence. This point of view is evidently inspired by the representation of Figure 29.2b.

Figure 29.2a shows the electrochemical potential plotted as a function of position along the p-n junction. We have assumed (as will be demonstrated below) that ϕ varies monotonically from one end to the other. Figure 29.2b shows an alternative representation of the same information in which the potential ϕ giving the position dependence in (29.3) is regarded as shifting $\mathcal{E}_c$ (or $\mathcal{E}_v$) rather than μ. In either case, the significance of the diagrams is that at any particular position x along the junction, the carrier densities are those that would be found in a piece of homogeneous material with the impurity concentrations prevailing at x, and with a chemical potential that is positioned with respect to the band edges as shown in the vertical section of the diagrams at x.

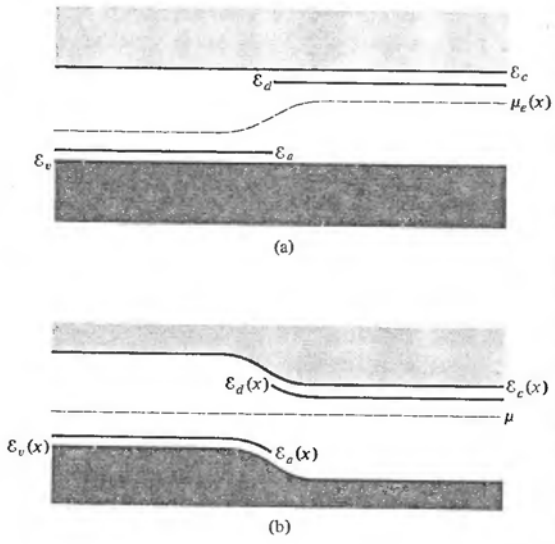

Figure 29.2
Two equivalent ways of representing the effect of the internal potential $\phi(x)$ on the electron and hole densities of a p-n junction. (a) The electrochemical potential $\mu_e(x) = \mu + e\phi(x)$ is plotted along the p-n junction. The carrier densities at any point x are those that would be found in a uniform semiconductor characterized by the fixed band and impurity energies $\mathcal{E}_c$, $\mathcal{E}_v$, $\mathcal{E}_d$, and $\mathcal{E}_a$, at a chemical potential equal to $\mu_e(x)$. (b) Here $\mathcal{E}_c(x) = \mathcal{E}_c - e\phi(x)$ is the energy of an electron wave packet localized about x formed from levels very near the conduction band minimum, and similarly for $\mathcal{E}_v(x)$. The energies of the local impurity levels are $\mathcal{E}_d(x) = \mathcal{E}_d - e\phi(x)$ and $\mathcal{E}_a(x) = \mathcal{E}_a - e\phi(x)$. The (constant) chemical potential is also shown. The carrier densities at any point x are those that would be found in a uniform semiconductor characterized by band and impurity energies equal to $\mathcal{E}_c(x)$, $\mathcal{E}_d(x)$, $\mathcal{E}_a(x)$, and $\mathcal{E}_v(x)$ at the fixed chemical potential μ.

Equation (29.6) (or its equivalent form, (29.9)) serves as the boundary condition in a differential equation determining the potential $\phi(x)$. The differential equation is simply Poisson's equation,[5]

$$-\nabla^2\phi = -\frac{d^2\phi}{dx^2} = \frac{4\pi\rho(x)}{\epsilon}, \tag{29.10}$$

[5] Here ϵ is the static dielectric constant of the semiconductor. The use of the macroscopic equation is possible because ϕ varies over the depletion layer, which is large on the interatomic scale.

relating the potential $\phi(x)$ to the charge distribution $\rho(x)$ giving rise to it. To express $\rho(x)$ in terms of ϕ and get a closed equation, we first note that if (as we have assumed) the impurities are fully ionized far from the junction, then they will remain fully ionized[6] at all x. Consequently the charge density due to the impurities and the carriers is[7]

$$\rho(x) = e[N_d(x) - N_a(x) - n_c(x) + p_v(x)]. \tag{29.11}$$

When the carrier and impurity densities (29.3) and (29.1) are substituted into the form (29.11) for the charge density, and the result is substituted into Poisson's equation (29.10), one finds a nonlinear differential equation for $\phi(x)$ whose exact solution usually requires numerical techniques.[8] However, a quite reasonable description of $\phi(x)$ may be had by exploiting the fact that the total change in $e\phi$ is of order $E_g \gg k_B T$. The relevance of this fact emerges when we combine (29.3) and (29.4) to write

$$\begin{aligned} n_c(x) &= N_d e^{-e[\phi(\infty)-\phi(x)]/k_B T}, \\ p_v(x) &= N_a e^{-e[\phi(x)-\phi(-\infty)]/k_B T}. \end{aligned} \tag{29.12}$$

Suppose that the change in ϕ occurs within a region $-d_p \leqslant x \leqslant d_n$. Outside of this region, ϕ has its asymptotic value, and therefore $n_c = N_d$ on the n side, $p_v = N_a$ on the p side, and $\rho = 0$. Within the region, except quite near the boundaries, $e\phi$ differs by many $k_B T$ from its asymptotic value, so $n_c \ll N_d$, $p_v \ll N_a$. Thus, except in the vicinity of $x = -d_p$ and $x = d_n$, the charge density (29.11) between $-d_p$ and d_n is quite accurately given by $\rho(x) = e[N_d(x) - N_a(x)]$, there being no appreciable carrier charge to cancel the charges of the "ionized" impurities. The points $x = -d_p$ and $x = d_n$ therefore mark the boundaries of the depletion layer.

Combining these observations, and using the form (29.1) for the impurity densities, we find that except for x just greater than $-d_p$ or just less than d_n, Poisson's equation is well approximated by

$$\phi''(x) = \left\{ \begin{array}{ll} 0, & x > d_n, \\ \dfrac{-4\pi e N_d}{\epsilon}, & d_n > x > 0, \\ \dfrac{4\pi e N_a}{\epsilon}, & 0 > x > -d_p, \\ 0, & -d_p > x. \end{array} \right\} \tag{29.13}$$

---

[6] If ϕ is monotonic (as we shall find below) this follows from the fact that the degree of ionization of an impurity increases, the farther the chemical potential is from the impurity level. See Figure 29.2 and Eqs. (28.32) and (28.34).

[7] The density of holes on the far n-side has the very small value $p_v(\infty) = n_i^2/N_d$ required by the law of mass action. However, the density of electrons on the far n-side actually exceeds N_d by this same small amount so as to insure that $n_c(\infty) - p_v(\infty) = N_d$. In computing the total charge density, if we ignore this small correction in n_c (as we have done in writing (29.4)), then we should also ignore the small compensating density of holes on the far n-side. Similar remarks apply to the small concentration of electrons on the far p-side. These "minority carrier densities" have negligible effect on the total balance of charge. We shall see below, however, that they play an important role in determining the flow of currents in the presence of an applied potential.

[8] Some aspects of that equation are investigated in Problem 1.

This immediately integrates to give

$$\phi(x) = \begin{cases} \phi(\infty), & x > d_n, \\ \phi(\infty) - \left(\dfrac{2\pi e N_d}{\epsilon}\right)(x - d_n)^2, & d_n > x > 0, \\ \phi(-\infty) + \left(\dfrac{2\pi e N_a}{\epsilon}\right)(x + d_p)^2, & 0 > x > -d_p, \\ \phi(-\infty), & x < -d_p. \end{cases} \tag{29.14}$$

The boundary conditions (continuity of ϕ and its first derivative) are explicitly obeyed by the solution (29.14) at $x = -d_p$ and $x = d_n$. Requiring them to hold at $x = 0$ gives two additional equations that determine the lengths d_n and d_p. Continuity of ϕ' at $x = 0$ implies that

$$N_d d_n = N_a d_p, \tag{29.15}$$

which is just the condition that the excess of positive charge on the n-side of the junction be equal to the excess of negative charge on the p-side. Continuity of ϕ at $x = 0$ requires that

$$\left(\frac{2\pi e}{\epsilon}\right)(N_d d_n^2 + N_a d_p^2) = \phi(\infty) - \phi(-\infty) = \Delta\phi. \tag{29.16}$$

Together with (29.15) this determines the lengths d_n and d_p:

$$d_{n,p} = \left\{\frac{(N_a/N_d)^{\pm 1}}{(N_d + N_a)}\frac{\epsilon\Delta\phi}{2\pi e}\right\}^{1/2}. \tag{29.17}$$

To estimate the sizes of these lengths we may write Eq. (29.17) in the numerically more convenient form

$$d_{n,p} = 105\left\{\frac{(N_a/N_d)^{\pm 1}}{10^{-18}(N_d + N_a)}[\epsilon e\ \Delta\phi]_{\text{eV}}\right\}^{1/2} \text{ Å}. \tag{29.18}$$

The quantity $\epsilon e\ \Delta\phi$ is typically of order 1 eV, and since typical impurity concentrations are in the range from 10^{14} to 10^{18} per cubic centimeter, the lengths d_n and d_p, which give the extent of the depletion layer, will generally be from 10^4 to 10^2 Å. The field within the depletion layer is of order $\Delta\phi/(d_n + d_p)$, and for d's of this size is therefore in the range from 10^5 to 10^7 volts per meter, for an energy gap of 0.1 eV.

The resulting picture of the depletion layer is shown in Figure 29.3. The potential ϕ varies monotonically through the layer, as asserted above. Except at the boundaries of the layer, the carrier concentrations are negligible compared with the impurity concentrations, so the charge density is that of the ionized impurities. Outside of the depletion layer the carrier concentrations balance the impurity concentrations, and the charge density is zero.

The mechanism establishing such a region of sharply reduced carrier densities is relatively simple. Suppose that one initially were able to impose carrier concentrations that gave charge neutrality at every point in the crystal. Such a configuration could not be maintained, for electrons would begin to diffuse from the n-side (where their concentration was high) to the p-side (where their concentration was very low), and

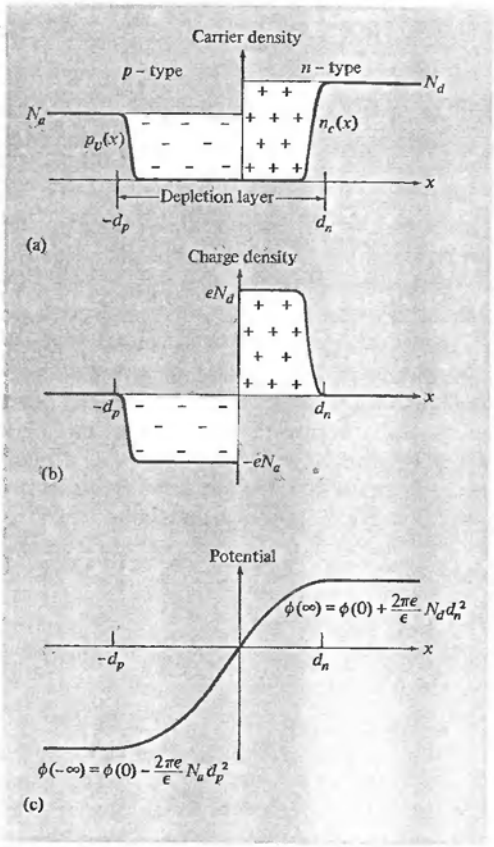

Figure 29.3
(a) Carrier densities, (b) charge density, and (c) potential $\phi(x)$ plotted vs. position across an abrupt p-n junction. In the analysis in the text the approximation was made that the carrier densities and charge density are constants except for discontinuous changes at $x = -d_p$ and $x = d_n$. More precisely (see Problem 1), these quantities undergo rapid change over regions just within the depletion layer whose extent is a fraction of order $(k_B T/E_g)^{1/2}$ of the total extent of the depletion layer. The extent of the depletion layer is typically from 10^2 to 10^4 Å.

holes would diffuse in the opposite direction. As this diffusion continued, the resulting transfer of charge would build up an electric field opposing further diffusive currents, until an equilibrium configuration was reached in which the effect of the field on the currents precisely canceled the effect of diffusion. Because the carriers are highly mobile, in this equilibrium configuration the carrier densities are very low wherever the field has an appreciable value. This is precisely the state of affairs depicted in Figure 29.3.

ELEMENTARY PICTURE OF RECTIFICATION BY A p-n JUNCTION

We now consider the behavior of a p-n junction when an external voltage V is applied. We shall take V to be positive if its application raises the potential of the p-side with respect to the n-side. When $V = 0$ we found above that there is a depletion layer some 10^2 to 10^4 Å in extent about the transition point where the doping changes from p-type to n-type, in which the density of carriers is reduced greatly below its value in the homogeneous regions farther away. Because of its greatly reduced carrier

density, the depletion layer will have a much higher electrical resistance than the homogeneous regions, and the whole device can therefore be viewed as a series circuit in which a relatively high resistance is sandwiched between two relatively low resistances. When a potential V is applied across such a circuit, almost all of the potential drop will occur across the region of high resistance. Thus even in the presence of an applied potential V, we expect that the potential $\phi(x)$ along the device will vary appreciably only within the depletion layer. When $V = 0$ we found that $\phi(x)$ rose from the p-side of the depletion layer to the n-side by the amount (which we now denote by $(\Delta\phi)_0$) given by Eq. (29.6), so we conclude that when $V \neq 0$ the change in potential across the depletion layer is modified to

$$\Delta\phi = (\Delta\phi)_0 - V. \tag{29.19}$$

Associated with this change in potential drop across the depletion layer, there is a change in the size of the layer. The lengths d_n and d_p giving the extent of the layer on the n- and p-sides of the junction are determined by Eqs. (29.15) and (29.16), which use only the value of the total potential drop across the layer, and the assumption that the carrier densities are greatly reduced throughout almost all of the layer. We shall find below that this assumption remains valid when $V \neq 0$, and therefore d_n and d_p continue to be given by Eq. (29.17) provided that we take the value of $\Delta\phi$ to be $(\Delta\phi)_0 - V$. Since d_n and d_p vary as $(\Delta\phi)^{1/2}$ according to Eq. (29.17), we conclude that when $V \neq 0$,

$$d_{n,p}(V) = d_{n,p}(0)\left[1 - \frac{V}{(\Delta\phi)_0}\right]^{1/2}. \tag{29.20}$$

This behavior of ϕ and the extent of the depletion layer are illustrated in Figure 29.4.

To deduce the dependence on V of the current that flows when a p-n junction is "biased" by the application of an external voltage, we must consider separately the currents of electrons and holes. Throughout the discussion that follows we shall use the symbol J for number current densities and j for electrical current densities, so that

$$j_e = -eJ_e, \quad j_h = eJ_h. \tag{29.21}$$

When $V = 0$, both J_e and J_h vanish. This does not, of course, mean that no individual carriers flow across the junction, but only that as many electrons (or holes) flow in one direction as in the other. When $V \neq 0$, this balance is disrupted. Consider, for example, the current of holes across the depletion layer. This has two components:

1. A hole current flows from the n- to the p-side of the junction, known as the hole *generation current*. As the name indicates, this current arises from holes that are generated just on the n-side of the depletion layer by the thermal excitation of electrons out of valence band levels. Although the density of such holes on the n-side ("minority carriers") is minute compared with the density of electrons ("majority carriers"), they play an important role in carrying current across the junction. This is because any such hole that wanders into the depletion layer is immediately swept over to the p-side of the junction by the strong electric field

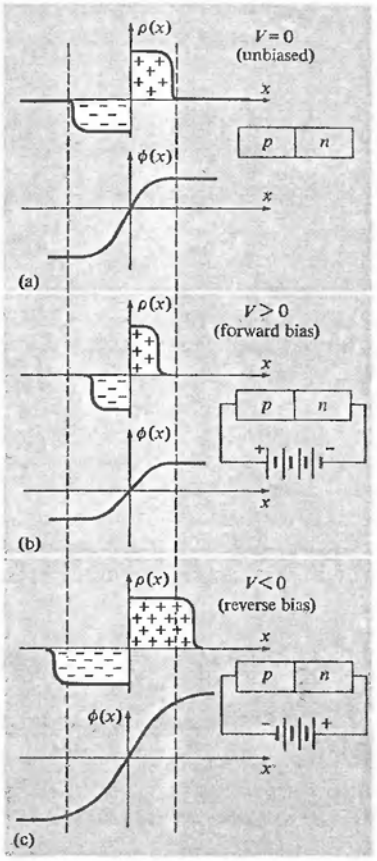

Figure 29.4
The charge density ρ and potential ϕ in the depletion layer (a) for the unbiased junction, (b) for the junction with $V > 0$ (forward bias), and (c) for the junction with $V < 0$ (reverse bias). The positions $x = d_n$ and $x = -d_p$ that mark the boundaries of the depletion layer when $V = 0$ are given by the dashed lines. The depletion layer and change in ϕ are reduced by a forward bias and increased by a reverse bias.

that prevails within the layer. The resulting generation current is insensitive to the size of the potential drop across the depletion layer, since any hole, having entered the layer from the n-side, will be swept through to the p-side.[9]

2. A hole current flows from the p- to the n-side of the junction, known as the hole *recombination current*.[10] The electric field in the depletion layer acts to oppose such a current, and only holes that arrive at the edge of the depletion layer with a thermal energy sufficient to surmount the potential barrier will contribute to

[9] The density of holes giving rise to the hole generation current will also be insensitive to the size of V, provided that eV is small compared with E_g, for this density is entirely determined by the law of mass action and the density of electrons. The latter density differs only slightly from the value N_c outside of the depletion layer when eV is small compared with E_g, as will emerge from the more detailed analysis below.

[10] So named because of the fate suffered by such holes upon arriving on the n-side of the junction, where one of the abundant electrons will eventually drop into the empty level that constitutes the hole.

the recombination current. The number of such holes is proportional to $e^{-e\Delta\phi/k_BT}$, and therefore[11]

$$J_h^{\text{rec}} \propto e^{-e[(\Delta\phi)_0 - V]/k_BT}. \tag{29.22}$$

In contrast to the generation current, the recombination current is highly sensitive to the applied voltage V. We can compare their magnitudes by noting that when $V = 0$ there can be no net hole current across the junction:

$$J_h^{\text{rec}}\big|_{V=0} = J_h^{\text{gen}}. \tag{29.23}$$

Taken together with Eq. (29.22), this requires that

$$J_h^{\text{rec}} = J_h^{\text{gen}} e^{eV/k_BT}. \tag{29.24}$$

The total current of holes flowing from the p- to the n-side of the junction is given by the recombination current minus the generation current:

$$J_h = J_h^{\text{rec}} - J_h^{\text{gen}} = J_h^{\text{gen}}(e^{eV/k_BT} - 1). \tag{29.25}$$

The same analysis applies to the components of the electron current, except that the generation and recombination currents of electrons flow oppositely to the corresponding currents of holes. Since, however, the electrons are oppositely charged, the electrical generation and recombination currents of electrons are parallel to the electrical generation and recombination currents of holes. The total electrical current density is thus:

$$\boxed{j = e(J_h^{\text{gen}} + J_e^{\text{gen}})(e^{eV/k_BT} - 1).} \tag{29.26}$$

This has the highly asymmetric form characteristic of rectifiers, as shown in Figure 29.5.

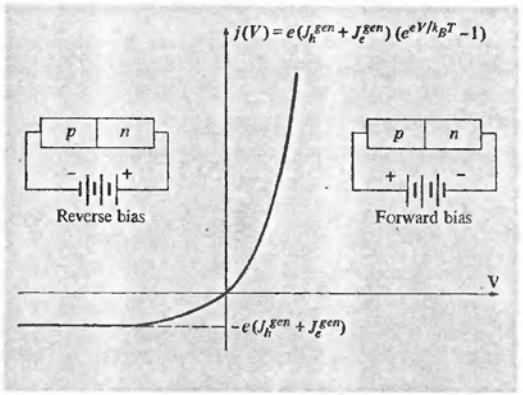

Figure 29.5
Current vs. applied voltage V for a p-n junction. The relation is valid for eV small compared with the energy gap, E_g. The saturation current $(eJ_h^{\text{gen}} + eJ_e^{\text{gen}})$ varies with temperature as e^{-E_g/k_BT}, as established below.

[11] In assuming that (29.22) gives the dominant dependence of the hole recombination current on V, we are assuming that the density of holes just on the p-side of the depletion layer differs only slightly from N_a. We shall find that this is also the case provided that eV is small compared with the energy gap E_g.