Introductory Electronics for Scientists and Engineers

Second Edition

ROBERT E. SIMPSON
University of New Hampshire

PRENTICE HALL, Englewood Cliffs, New Jersey 07632

CHAPTER 4

Semiconductor Physics

4.1 INTRODUCTION

In this chapter we consider some of the basic features of the physics of semiconductors that will enable us to understand physically why a transistor can be made to amplify signals, why transistor parameters depend so strongly upon temperature and the magnitudes of the various resistances, voltages, and currents in transistor circuits. Most semiconductor devices are made from either germanium or silicon crystals; hence particular emphasis will be placed upon these elements.

4.2 ENERGY LEVELS

In classical physics energy can be thought of as the ability to do work. It has units of joules (in mks units) or ergs (in cgs units) or electron volts. One electron volt (eV) is equal to 1.6×10^{-12} erg or 1.6×10^{-19} joule and is defined as the energy an electron gains when it falls through a potential drop (or voltage drop) of one volt. The electron volt is commonly used to measure energy in nuclear, atomic, and semiconductor physics. A useful number to remember is $1/40\,\text{eV} = 0.025\,\text{eV}$, which is the approximate average translational kinetic energy an atom or free electron has in thermal equilibrium at room temperature.

Classically, there is no restriction on the amount of energy an object may have. However, in quantum physics where we are dealing with very small objects such as nuclei, atoms, electrons, or molecules, the energy an object may have is often restricted by the basic quantum-mechanical laws that apply in such situations. If we calculate the energy possible for a single isolated atom according to the quantum theory, we find that there are a number of sharp or discrete energy values or *levels* possible. These energy levels depend mostly on the orbital electronic properties, not on the nuclear properties, because the nuclear states are usually so stable. The higher levels are usually more closely spaced than the lower levels, but the main point to remember is that the energy levels are discrete. It is impossible for the atom to have any energy between these discrete energy levels, which

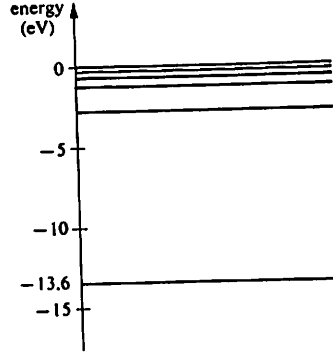
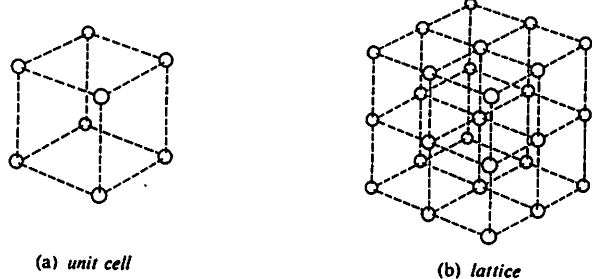


FIGURE 4.1 Energy levels for a single hydrogen atom.

are usually spaced of the order of electron volts apart. The energy level diagram for a hydrogen atom, for example, is given in Fig. 4.1. The horizontal axis is essentially meaningless. Larger atoms containing more electrons in general have more complicated energy level diagrams, but the levels are still discrete with forbidden gaps between them and look basically like those in Fig. 4.1.

4.3 CRYSTALS

A *crystal* is a solid in which the atoms are arranged in a regular periodic way. Most elements, if extremely pure, will form crystals when cooled carefully from the liquid. The exact geometry of the atoms depends on the chemical bonds formed among the outer or *valence* electrons of the atoms. The electrons in the inner shells ordinarily take no part in bonds between atoms because they are so tightly bound to their own nucleus. Thus, the crystal structure depends on the atomic number and the valence of the atom involved. Some compounds, particularly the simpler ones, also form crystals easily: for example, NaCl, KCl, GaAs, and others. If the regular, periodic arrangement of atoms is unbroken throughout the entire piece of material, then we have a perfect "single" crystal. A solid is called *polycrystalline* if it consists of a large number of smaller single crystals oriented randomly with respect to one another. In general, a substance takes on a polycrystalline form upon solidifying from the liquid and special experimental techniques and extremely pure samples must be used to make a single crystal of an appreciable size. In fact, much of the recent progress in transistor and chip technology has come from improvements in the purification techniques for silicon.



(a) *unit cell* (b) *lattice*

FIGURE 4.2 Cubic crystal lattice.

The regular geometrical arrangement of atoms is called a crystal *lattice*. One example of a crystal is a lattice in which all the atoms occur at the corners of cubes of identical size, as shown in Fig. 4.2. Such an arrangement is called a *cubic* lattice. The cube is called a unit cell of this crystal, because the entire crystal lattice can be built up conceptually by a regular "stacking-up" of cubes.

There are many types of imperfections or defects present in any real single crystal: There can be mechanical breaks or discontinuities in the crystal structure extending over many unit cells, atoms can be missing, an extra atom can be present—in an unusual position in a unit cell, or an atom of a different element (with the "wrong" number of valence electrons) may be present in a regular lattice position. The type and number of such imperfections play an extremely important role in determining the mechanical and electrical properties of the crystal. Most transistors and chips today are made from extremely small single crystals of silicon, although some special-purpose semiconductor devices are made from single crystals of compounds such as GaAs. The unit cell of germanium and silicon is a face-centered cubic cell, as shown in Fig. 4.3, in which the atoms

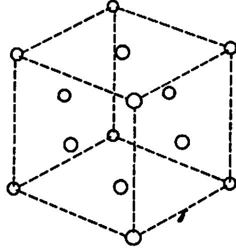


FIGURE 4.3 Face-centered cubic cell (Ge and Si).

occur at the eight corners of a cube and in the centers of the six faces of the cube. The length of one side of the cube is called the *lattice constant*; it is 5.66 Å in germanium and 5.43 Å in silicon (1 Å = 10^{-8} cm = 10^{-10} m).

If several face-centered cubic cells are drawn together, as they occur in either a germanium or a silicon crystal, a very complicated interlocking structure results in which each atom has four nearest neighbors. Both germanium and silicon are Group IV elements in the chemical periodic table and hence have four outer or valence electrons. The four valence electrons enter into covalent electronic bonds with the four nearest-neighbor atoms in the lattice. The resulting covalent chemical bonds literally hold the lattice together, and energy must be supplied from some source to break a valence electron loose from such a bond. The inner electrons are much more tightly bound to the atom's nucleus and are never free to move in the lattice.

4.4 ENERGY LEVELS IN A CRYSTAL LATTICE

If several identical atoms are brought together, the energy levels of the individual isolated atoms split into a closely spaced set of energy levels, as shown in Fig. 4.4(a) and (b).

If we calculate, from quantum theory, the possible energy levels for an electron of an atom in a crystal containing a large number of atoms (e.g., 10^{20} atoms for a tiny crystal composing a transistor), we find that the allowed energy levels are very closely spaced together and that there are large forbidden energy gaps between groups of closely spaced energy levels. A group of many closely spaced energy levels is called an energy *band*. The most important two bands for the purpose of understanding transistors are the *valence* band and the *conduction* band shown in Fig. 4.4. The gap between the valence and conduction bands is 0.72 eV in germanium and 1.09 eV in silicon. Note that these energy gaps are much greater than kT, which is equal to 0.025 eV at room temperature. The electron energy levels in either band are so closely spaced in energy that the band may be regarded as a continuum for most purposes.

The electrons in the conduction band are bound only very weakly to individual atoms in the lattice; hence they are essentially free to move throughout the lattice and take part in conduction of electrical current. Electrons in the valence band are those electrons that form the covalent bonds between atoms of the lattice. They are "valence electrons" in chemical terminology, hence the name "valence" band. Electrons in atomic shells inside the valence electrons lie in energy bands of the order of electron volts below the valence band. These electrons play no part in conduction at temperatures under several thousand degrees, and so from now on we will consider only the valence and conduction band electrons.

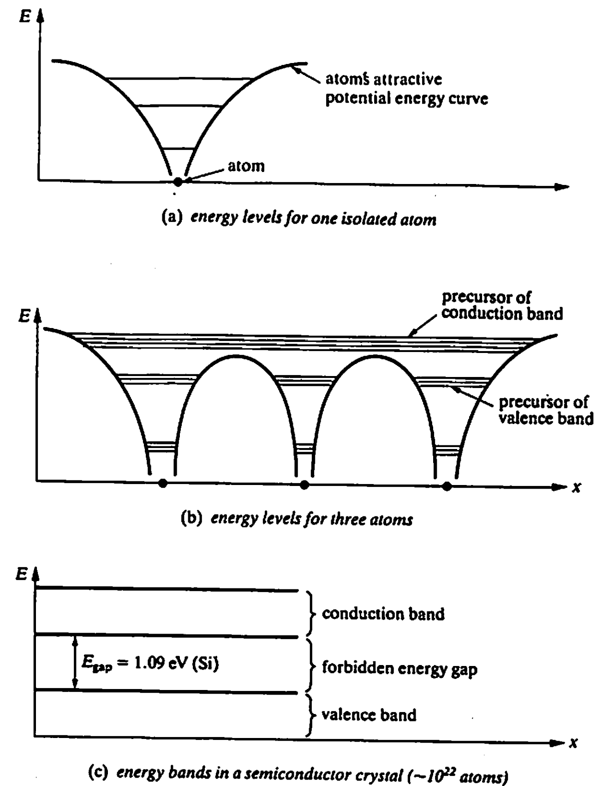


FIGURE 4.4 Atomic electron energy levels.

The physical reason for the energy gap is simply that it takes a certain amount of energy or work, E_{gap}, to pull an electron loose from a bond between lattice atoms and make it free to move through the crystal. Thus, the energy gap is physically the amount of energy that must be given an electron in a bond to free it from that bond. We can now see that at absolute zero, when all the electrons must fall into the bonds between the lattice atoms, all the electrons must lie in the valence band, and *none* lies in the conduction band.

The problem of conduction in a crystal is more complicated than we

have just stated because of two quantum-mechanical laws we have not yet considered: the Pauli exclusion principle and the Fermi–Dirac statistics.

4.5 PAULI EXCLUSION PRINCIPLE

An energy level is characterized by a set of quantum numbers that come from a quantum-mechanical calculation of the physical system's allowed energies. In the simplest case of a hydrogen atom each electron has four quantum numbers: n, l, m_l, and m_s. It can be shown from quantum mechanics that the energy of a hydrogen atom depends only upon n, the principal quantum number according to the formula $E = -\mu e^4/2\hbar^2 n^2$. The electronic charge is e, μ is the reduced mass of the electron and nucleus, $\hbar = h/2\pi$, where h is Planck's constant, l is called the azimuthal quantum number. The electron's orbital angular momentum L is given by $L = \sqrt{l(l+1)}\hbar$, where $l = 0, 1, 2, 3, \ldots, n-1$. The symbol m_l is called the magnetic quantum number; the z component L_z of the orbital angular momentum is given by $L_z = m_l\hbar$, where $m_l = -l, \ldots, +l$. The value of m_s determines the z component of the electron spin angular momentum according to $S_z = m_s\hbar$, and $m_s = -\frac{1}{2}$ or $+\frac{1}{2}$. Hence the $n = 2$ energy level, for example, can contain electrons with quantum numbers n, l, m_l, and m_s equaling 2, 0, 0, $+\frac{1}{2}$ or 2, 0, 0, $-\frac{1}{2}$ or 2, 1, 1, $+\frac{1}{2}$ or 2, 1, 1, $-\frac{1}{2}$ or 2, 1, 0, $+\frac{1}{2}$ or 2, 1, 0, $-\frac{1}{2}$ or 2, 1, -1, $+\frac{1}{2}$ or 2, 1, -1, $-\frac{1}{2}$, a total of eight possible states or sets of quantum numbers.

The Pauli exclusion principle says that no two (or more) electrons can have exactly the same set of quantum numbers. Thus, in an energy level characterized by the quantum numbers n, l, m_l, m_s, if $n = 2$ there can be at most eight electrons. We say that the $n = 2$ level is "filled" with eight electrons. A hydrogen atom never has eight electrons, but a hydrogen-like energy level can exist in many heavy atoms. For an energy band in a crystal, each of the closely spaced energy levels has its own set of quantum numbers. Electrons in the crystal will naturally tend to fill up the lowest energy levels first but always subject to the restriction that no two electrons can occupy the same state; that is, no two can have the same set of quantum numbers.

The most important implication of the exclusion principle is that a completely full energy band of electrons cannot conduct. This can be seen by realizing that conduction implies that an electron is raised from a lower to a slightly higher energy state. However, in a completely full band the higher energy state is already occupied by an electron, and the exclusion principle prevents a second electron from jumping up to this higher state. Thus for conduction to occur, an energy band must be only partially full; there must be empty higher energy states for the electron to jump up to. Another implication of the exclusion principle is that electrons for conduction come mainly from near the top of the partially filled band because these electrons need only a small increase in energy to jump up to an unfilled level in the band.

4.6 FERMI–DIRAC STATISTICS

To understand the electrical behavior of a semiconductor, we clearly must know how many electrons are in the conduction band and how many are in the valence band, because only these electrons in the conduction band can move and create a current flow in response to an applied voltage. The basic problem is then to calculate the distribution of the electrons in the crystal among the various allowable energy levels. We have already stated that a quantum-mechanical calculation of the energy levels in a crystal lattice yields the result that the allowable energy levels lie in two bands (conduction and valence), separated by a forbidden energy region or gap. The next thing to calculate is the probability $F(E)$ that a given energy level E somewhere in a band is actually occupied by an electron. If we let $N(E)\,dE$ equal the number of electrons per unit volume between E and $E + dE$, then we write $N(E)\,dE$ as the product of the number $\rho(E)\,dE$ of allowable or available energy states per unit volume in the range E to $E + dE$ [where $\rho(E)$ is often called the *density of states*] and the probability $F(E)$ that the energy level E is actually occupied.

$$N(E)\,dE = \rho(E)\,dE\,F(E) \tag{4.1}$$

The $\rho(E)$ and therefore $N(E)\,dE$ will be zero in the energy gap and positive in the conduction and valence bands. The exact form of $\rho(E)\,dE$ comes from the quantum-mechanical solution to the problem of calculating the allowable energy levels for electrons in a periodic crystal lattice. The result is

$$\rho(E) = \frac{2^{7/2} m^{3/2} \pi}{h^3} E^{1/2} \tag{4.2}$$

for electrons not too near the top of the band. $E = 0$ represents the energy of the bottom of the band,[†] m is the mass of the electron, and h is Planck's constant.

The Fermi function $F(E)$ tells us the probability that the energy level E is actually occupied by an electron. It seems intuitively reasonable that $F(E)$ should depend on temperature, because the higher the temperature, the more thermal energy is available to be distributed among the electrons

[†] Near the top of the band electrons undergo Bragg reflections from atoms in the lattice, and the net result is that fewer energy levels are allowable than are predicted from equation (4.2).

and the nuclei of the lattice atoms. And the more energetic the electrons, the greater the probability that a higher energy state is occupied. At absolute zero when there is zero† thermal energy available, all vibrational motion of the atoms in the lattice should cease; and all the electrons should be lying in the lowest possible energy levels, subject only to the restriction of the Pauli exclusion principle. Thus we would expect the Fermi function $F(E)$ to be temperature dependent and, at absolute zero, to be a constant value from $E = 0$ up to some maximum energy, which is the energy of the highest filled state.

$F(E)$ is calculated by an involved statistical argument. The problem is to calculate the probability that an energy level E is occupied subject to four conditions: (1) The total number of electrons is constant (conservation of charge). (2) The total energy of all the electrons remains constant (conservation of energy). (3) No two electrons can be distinguished from one another. (4) No two (or more) electrons can lie in the same quantum state. This total energy depends on the temperature, which is assumed to be constant (i.e., the particles are assumed to be in thermal equilibrium). Once the probability is known, the maximum probability is calculated, because we know nature in equilibrium always assumes the most probable configuration. There are clearly many ways to distribute a fixed total amount of energy among N electrons; this calculation gives us the most *likely* or most probable distribution of electrons among the various energy levels.

The net result of the calculation is that the Fermi function is given by

$$F(E) = \frac{1}{e^{(E-E_F)/kT} + 1} \tag{4.3}$$

where E_F is a constant energy called the Fermi energy, k is Boltzmann's constant equal to 1.38×10^{-16} ergs/deg $= 1.38 \times 10^{-23}$ J/deg, and T is the temperature in kelvins. To see the physical meaning of $F(E)$, let us graph it for various temperatures (see Fig. 4.5). At absolute zero the Fermi function

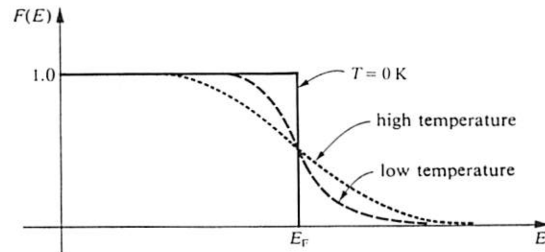


FIGURE 4.5 The Fermi function vs. E.

†Actually at absolute zero, there is a nonzero, fixed minimum value of energy available to the electrons and atoms; this is called the *zero-point* energy and comes from a careful consideration of the uncertainty principle of quantum mechanics.

$F(E)$ is 1.0 from $E = 0$ up to $E = E_F$; for energies above E_F, $F(E) = 0$. This says that at absolute zero all the allowable energy levels from 0 up to E_F are filled and that no levels above E_F are filled, which is just what we expect physically. E_F represents the highest energy level filled at absolute zero.

The physical meaning may become clearer if we refer to Fig. 4.6,

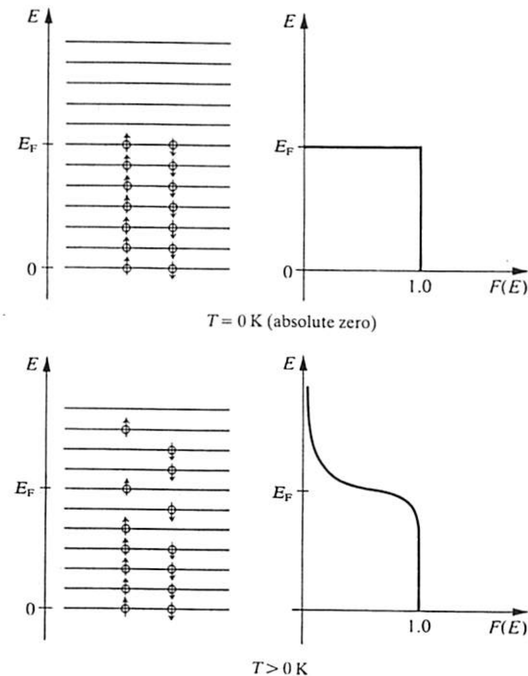


FIGURE 4.6 Simple energy level picture and Fermi function.

which is a simple energy level picture of a hypothetical solid with evenly spaced, allowed energy levels. The density of states $\rho(E)$ gives the number of these energy levels per unit energy. At absolute zero the electrons have cascaded down to the lowest energy states possible, subject to the restriction of the Pauli exclusion principle, namely, that no more than one electron can occupy any one energy level with a specific set of quantum numbers. In Fig. 4.6 we have assumed that the spin quantum number does not affect the

energy; that is, the energy does not depend upon the quantum number m_s. Thus two electrons, one with spin "up" ($m_s = +\frac{1}{2}$) and one with spin "down" ($m_s = -\frac{1}{2}$), can occupy the same energy level. The Fermi energy will be at the topmost occupied energy level at 0 K. Notice that the only way to change the Fermi energy would be to add more electrons or to change the spacing of the energy levels. Figure 4.6(b) shows the distribution of the electrons at a temperature above absolute zero. The density of the electrons per unit energy increases for energies above the Fermi level as the temperature is raised.

4.7 ELECTRON ENERGY DISTRIBUTION

Before we can multiply the density of states $\rho(E)\,dE$ by the Fermi function to obtain the number of electrons actually between E and $E + dE$, we must first know E_F. If we note that at absolute zero all the electrons in the lattice will be bound in the covalent bonds between the atoms, we see that the conduction band must be empty and the valence band full. Therefore, since the Fermi level is the maximum energy an electron can have at absolute zero, the Fermi energy must be somewhere above the valence band and below the bottom of the conduction band. An exact treatment gives the result that

$$E_F = \frac{E_{gap}}{2} \tag{4.4}$$

where $E = 0$ is the top of the valence band and E_{gap} is the energy gap. In other words, the Fermi energy lies exactly halfway between the valence and conduction bands. If we take $E = 0$ to be the top of the valence band, then from (4.2) the density of states in the conduction band must be

$$\rho(E) = \frac{2^{7/2} m^{3/2} \pi}{h^3} (E - E_{gap})^{1/2} \tag{4.5}$$

Thus the number $N(E)\,dE$ of electrons in the conduction band between E and $E + dE$ will be given by $N(E)\,dE = \rho(E)\,dE\,F(E)$. $F(E)$, $\rho(E)$, and $N(E)$ are graphed on an energy level diagram in Fig. 4.7.

Substituting for $\rho(E)$ and $F(E)$, we obtain

$$N(E)\,dE = \frac{2^{7/2} m^{3/2} \pi}{h^3} (E - E_{gap})^{1/2} \frac{1}{e^{(E-E_F)/kT} + 1}\,dE \tag{4.6}$$

For pure silicon or germanium with E in the conduction band, $E - E_F \gg kT$. Hence,

$$e^{(E-E_F)/kT} + 1 \cong e^{(E-E_F)/kT}$$

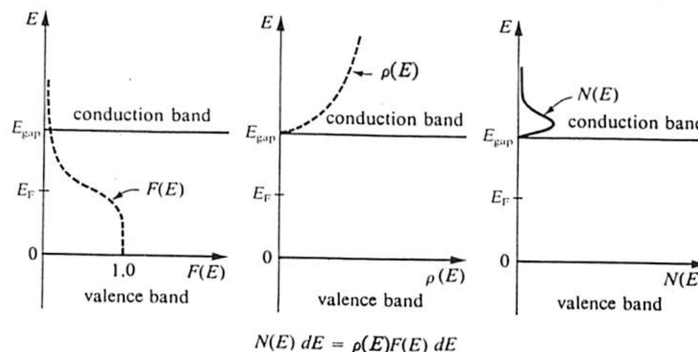


FIGURE 4.7 Density of electrons in conduction band.

So we may write

$$N(E)\,dE \cong \frac{2^{7/2} m^{3/2} \pi}{h^3} (E - E_{gap})^{1/2}\, e^{-(E-E_F)/kT}\,dE \tag{4.7}$$

The total number N_{cb} of electrons in the conduction band can be obtained by integrating $N(E)\,dE$ from the bottom of the conduction band to the top, which can be taken to be $E = \infty$ for all practical purposes. The result is

$$N_{cb} = \int_{E_{gap}}^{\infty} N(E)\,dE = \frac{2^{5/2}(m\pi kT)^{3/2}}{h^3} e^{-E_{gap}/2kT} \tag{4.8}$$

where we have set $E_F = E_{gap}/2$. It is useful to rewrite N_{cb} as

$$N_{cb} = AT^{3/2} e^{-E_{gap}/2kT} \tag{4.9}$$

where

$$A \equiv \frac{2^{5/2}(m\pi k)^{3/2}}{h^3} = 4.6 \times 10^{15} \frac{\text{electrons}}{\text{cm}^3} (\text{deg})^{-3/2}$$

The most important thing to remember is that the total number of electrons in the conduction band depends on the temperature according to $T^{3/2}$ times $e^{-E_{gap}/2kT}$. Table 4.1 gives some numerical values for various temperatures. Note that the exponential factor $e^{-E_{gap}/kT}$ is much smaller for silicon because of the larger value of E_{gap}. Therefore, the number of electrons thermally excited to the conduction band is much less for silicon than for germanium.

Also notice that the exponential factor $e^{-E_{gap}/2kT}$ is a much stronger

TABLE 4.1. N_{cb} (cm^{-3}) Equals the Number of Thermally Excited Electrons/cm^3 in the Conduction Band for Various Temperatures in Silicon and Germanium

	T	$T^{3/2}$	$e^{-E_{gap}/2kT}$	$AT^{3/2}e^{-E_{gap}/2kT} = N_{cb}$
Ge ($E_{gap} = 0.72$ eV)	20°C = 293 K	5000	6.21×10^{-7}	1.51×10^{13}
	30°C = 303 K	5280	10.3×10^{-7}	2.65×10^{13}
	40°C = 313 K	5560	16.2×10^{-7}	4.39×10^{13}
	50°C = 323 K	5800	28.8×10^{-7}	8.15×10^{13}
Si ($E_{gap} = 1.1$ eV)	20°C = 293 K	5000	3.52×10^{-10}	0.858×10^{10}
	30°C = 303 K	5280	7.09×10^{-10}	1.82×10^{10}
	40°C = 313 K	5560	13.56×10^{-10}	3.67×10^{10}
	50°C = 323 K	5800	26.1×10^{-10}	7.36×10^{10}

function of temperature than $T^{3/2}$. Therefore, the gap energy E_{gap} mainly determines the number of electrons in the conduction band for pure germanium and silicon.

From the graphs in Fig. 4.8 we see that, starting at 20°C, the electron

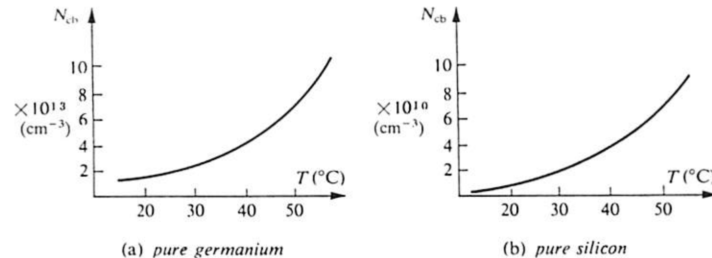


FIGURE 4.8 Conduction band electron density N_{cb} as a function of temperature.

density in the conduction band N_{cb} doubles for a 13°C temperature rise for germanium; for silicon N_{cb} doubles for only an 8°C rise. However, there are far fewer electrons in absolute number in silicon for the same temperature.

4.8 CONDUCTION IN SEMICONDUCTORS

The Fermi level in pure semiconductors lies halfway between the top of the valence band and the bottom of the conduction band. Hence, at absolute zero all the electrons are in the valence band, which is then full and thus can carry no current. At higher temperatures some electrons acquire

enough thermal energy to break loose from a valence bond in the lattice, thus jumping from the valence band up to the conduction band. The conductivity of a semiconductor then increases with increasing temperature.

In a metal the Fermi level lies *in* the conduction band, which is only partially full. Hence the metal is a good conductor at any temperature because of the many empty energy levels in the conduction band above the filled levels. There is no energy gap that the electrons must jump to get into the conduction band.

In a crystalline insulator the energy level picture is qualitatively the same as for a semiconductor, but the energy gap is much larger. For example, in diamond the energy gap is 5 eV, which is so large that an enormous temperature (approximately 60,000°C!) is required to thermally excite an electron from the valence band all the way up to the conduction band, where it may then contribute to conduction (see Fig. 4.9).

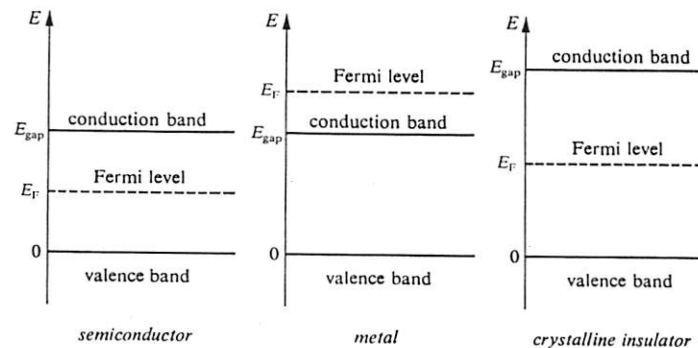


FIGURE 4.9 Location of Fermi level in various materials.

If we are to use semiconductors as practical components in electrical circuits, then they must be able to carry a reasonable amount of current. From the preceding section the total number N_{cb} of electrons per unit volume in the conduction band of a pure semiconductor was given by

$$N_{cb} = AT^{3/2}e^{-E_{gap}/2kT}$$

Substituting numerical values gives us

1. $N_{cb} = 1.5 \times 10^{13}$ electrons/cm^3 for Ge at room temperature.
2. $N_{cb} = 8.6 \times 10^{9}$ electrons/cm^3 for Si at room temperature.

Are there enough electrons in either pure germanium or pure silicon to carry a practical amount of current? The answer is no, for the following reasons. The maximum current I that can be passed is the number of electrons flowing per second multiplied by the charge per electron. The number flowing per second is the number per unit volume times the area times the speed. Thus, $I = N_{cb}Aev$, where e is the electronic charge. It can be shown that the average speed v of electrons flowing in germanium is of the order of 4×10^4 cm/s for a 10 V/cm electric field (voltage gradient) applied.† If we assume a generous cross-sectional area of 1 mm × 1 mm, then the current I is

$$I = N_{cb}Aev = (1.5 \times 10^{13}\ \text{cm}^{-3})(0.1\ \text{cm})^2(1.6 \times 10^{-19}\ \text{C})(4 \times 10^4\ \text{cm/s})$$
$$= 9.6 \times 10^{-4}\ \text{C/s} = 960\ \mu\text{A} = 0.96\ \text{mA}$$

A value of 0.96 mA is too small a current for many practical applications. In addition, the actual current will be less than that because not every electron in the conduction band will contribute to conduction. The current passed by a silicon cube is even less because the number of electrons in the conduction band is smaller and also because, for a given applied electric field the electrons move more slowly than in germanium.

Because of the temperature dependence of the number of electrons available for conduction, one might try to increase the number of electrons by warming the semiconductor. However, a quick numerical calculation shows that even for a temperature of 100°C, the maximum current passed is still too small for practical use in circuits.

There is another way of exciting electrons up into the conduction band, and that is by the absorption of electromagnetic radiation. An electron in the valence band can jump up to the conduction band if it absorbs an electromagnetic photon of energy greater than the gap, $h\nu \geqq E_{gap}$, where ν is the frequency of the photon. This process is called *photoconduction*. Some devices (photocells) detect light using this process, and the photovoltaic cell converts sunlight into electrical energy by photoconduction.

The practical way of increasing the number of electrons in the conduction band so that the semiconductor can carry a reasonable amount of current is to effectively add more electrons by introducing "impurity" or "donor" atoms to the lattice. The process of adding impurity atoms is called *doping* the crystal. Consider a silicon lattice at absolute zero. All the valence electrons are locked in covalent bonds between the atoms, so none

†The speed is calculated from the electron mobility μ, which is defined as the electron drift speed per unit electric field.

$$\mu = \frac{v(\text{cm/s})}{|\mathbf{E}|(\text{V/cm})}$$

	Ge	Si	
$\mu =$	3600	1200	for electrons
$\mu =$	1700	250	for holes

is available for conduction. Each atom has four valence electrons that bond to the four nearest atoms in the lattice. If we now introduce a donor atom with *five* valence electrons in a regular lattice site in place of a silicon atom, then there clearly will be one electron left over (i.e., not bound in a covalent bond with the four nearest silicon neighbors). This one electron is only very *loosely* bound to its atom and therefore can easily be pulled loose. Once freed it can migrate through the crystal under the influence of an applied electric field and contribute to conduction. In other words, only a small amount of energy is necessary to free this fifth electron from its donor atom and raise it up into the conduction band.

On the energy level diagram for silicon the fifth electron when loosely attached to its donor atom then lies in an energy level in the gap only slightly below the bottom of the conduction band (see Fig. 4.10). The

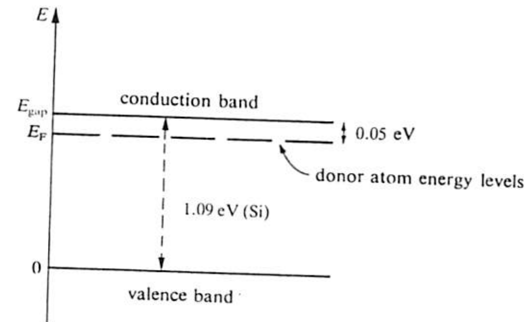


FIGURE 4.10 n-type semiconductor energy level diagram.

concentration of these donor atoms is such that they are spaced of the order of 30 Å apart in the lattice. Thus, each donor atom has four silicon atoms as its nearest neighbors in the lattice. Notice that the donor atom, once it loses its fifth electron, is a positively charged ion and is still locked in the crystal lattice by the four remaining valence electrons—being bonded to the adjacent four silicon atoms. Thus, it is only the electrons, not the positive ions, that contribute to conduction. The atom that gives up its electron is called a *donor* atom because it *donates* an electron to the conduction band. The donor energy level in silicon lies only about 0.05 eV below the conduction band, so at room temperature (300 K), where the average thermal agitation energy $kT = 0.025$ eV, almost all the donor atoms are ionized and contribute an electron to the conduction band. A typical concentration of donor impurity atoms is $10^{16}/\text{cm}^3$, so there will be of the order of 10^{16} electrons/cm^3 in the conduction band from the ionized donor

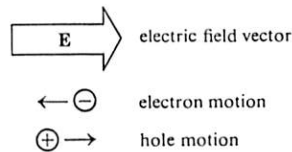


FIGURE 4.12 Electron and hole motion.

semiconductor points from left to right, as shown in Fig. 4.12, then electrons will move from right to left against the electric field, and holes will move from left to right. A positive hole moving from left to right actually involves electrons moving from right to left, but from now on we will speak of the holes as if they were positively charged particles capable of moving in the direction of the applied electric field. It can be shown that one can treat the motion of the holes exactly by treating them as positively charged particles with an effective mass slightly larger than the electron mass. The larger hole mass simply means that for a given applied electric field, a hole will accelerate less rapidly than will an electron.

To sum up: In an n-type semiconductor the impurity or donor atoms have one more valence electron than the atoms composing the crystal lattice. These donor atoms readily donate their extra valence electron to the conduction band, thus producing mobile electrons in the conduction band to carry current and positively charged donor atoms or ions locked in the lattice. In a p-type semiconductor the impurity or acceptor atoms have one less valence electron than the atoms composing the lattice. These acceptor atoms readily attract electrons up out of the valence band, thus producing mobile positive holes in the valence band to carry current and negatively charged acceptor atoms or ions locked in the lattice. In both n- and p-type semiconductors, at room temperature, there are always present a few (10^{13} cm^{-3} in Ge and 10^{10} cm^{-3} in Si) holes in the valence band and electrons in the conduction band caused by thermal breaking of lattice bonds.

4.9 p-n JUNCTIONS

If a p-type semiconductor is joined to an n-type semiconductor to form a "good" junction, that is, a junction at which there are few breaks or imperfections in the lattice structure, then this junction will act as a rectifier—it will conduct current readily in one direction and only very poorly in the other direction. Such a junction cannot be formed by merely pressing together a p-type semiconductor and an n-type semiconductor; this procedure would produce a poor junction with gaps at the junction larger than the lattice spacing. A good junction is usually made by changing over

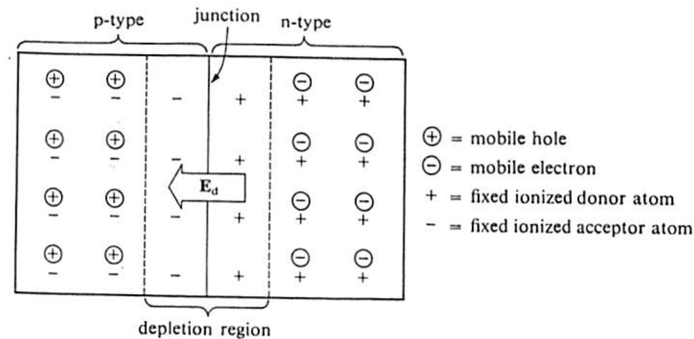


FIGURE 4.13 p-n junction (only majority carriers shown).

the type of impurity from p- to n-type while the crystal is being grown. We will consider a junction in which the type of impurity changes abruptly as we cross the junction; this type of junction is called an *abrupt* junction.

An abrupt p-n junction is shown in Fig. 4.13. Let us first consider the behavior of the majority carriers on both sides of the junction. Majority carriers are mobile and are continually diffusing around in the lattice, due to thermal motion. If holes in the p-type material diffuse away from the junction back into the p-type material, then a concentration gradient will build up, thus tending to diffuse the holes back toward the junction. Diffusion always results in a flow from regions of higher concentration towards regions of lower concentration. Similarly, if electrons in the n-type material diffuse away from the junction, the concentration gradient thus established will tend to diffuse the electrons back towards the junction. Near the junction some holes will diffuse across the junction from the p-type material into the n-type material where they will recombine with the mobile electrons of the n-type material. Similarly some electrons will diffuse across the junction from the n-type material into the p-material and will recombine with the mobile holes of the p-type material. Recombination of a hole and an electron merely means that the electron drops from the conduction band down into the hole in the valence band. Notice that after recombination neither the electron nor the hole can conduct because the electron is no longer in the conduction band, and the valence band hole is filled.

The result of this recombination is the formation of a thin region at the junction where there are essentially no mobile charge carriers. This is called the *depletion region* because the mobile charge carriers have been depleted here. The only charges remaining in the depletion region are the ionized acceptor and donor impurity atoms that are fixed in the crystal lattice and,

of course, the electrons in the filled valence band. These fixed charges produce an electric field $\mathbf{E}_d$ in the depletion region pointing from the n-type toward the p-type material. This field tends to sweep any electrons near the junction back toward the n-type material, and any holes back toward the p-type material. Thus, the depletion region has no free charge carriers, and the electric field produced tends to keep the depletion region free of any mobile charge carriers that are either created in it by thermal excitation or that may diffuse into it. In effect, the depletion region is a thin slab of insulator sandwiched between the p- and n-type materials. A typical thickness d for the depletion region is about one micron, which is 10^{-4} cm or 10,000 Å. If the doping of the p- and n-type materials is equal—that is, if the number of acceptor atoms per cubic centimeter, [p], equals the number of donor atoms per cubic centimeter, [n]—then the depletion region extends equally into the p- and n-type materials on either side of the p-n interface. However, if $[p] \neq [n]$, then the depletion region extends unequally into the p- and n-type materials. It can be shown that $d_n/d_p = [p]/[n]$, where d_n is the depletion region thickness in the n-type material, and d_p is the depletion region thickness in the p-type material. The total depletion region thickness $d = d_n + d_p$. This result is reasonable because the electrons and holes *diffuse* across the p-n interface and then recombine. If there are many more donor atoms per cm^3 in the n-type material than there are acceptor atoms per cm^3 in the p-type material ($[n] \gg [p]$), then the electrons diffusing into the p-type material will have to diffuse a long way before they all recombine with the holes. Thus we expect a large d_p. Conversely, the few holes from the p-type material that diffuse over into the n-type material very quickly recombine with the plentiful electrons there; thus we get a small d_n.

The effective voltage V_c developed between the p- and n-type materials is called the *contact potential* and, if we assume parallel plate geometry, is given by $V_c = E_d d$, where d is the thickness of the depletion region. V_c is about 0.2 V for a germanium junction and 0.5 V for a silicon junction. The electric field in the depletion region points from the n-type material toward the p-type material. This field thus tends to keep the majority carriers *out* of the depletion region, but notice that any *minority* carriers thermally generated in the depletion region will be swept across the depletion region by this same electric field.

When equilibrium is established for an isolated p-n junction, there is no net current flowing across the junction. For every majority hole from the p-type material diffusing across against the electric field to the n-type material, there will be a minority hole from the n-type material accelerated back by the electric field to the p-type material. A similar two-way flow will occur for electrons.

If a battery is connected to a p-n junction with a polarity to make [see Fig. 4.14(a)] the p-type material negative with respect to the n-type

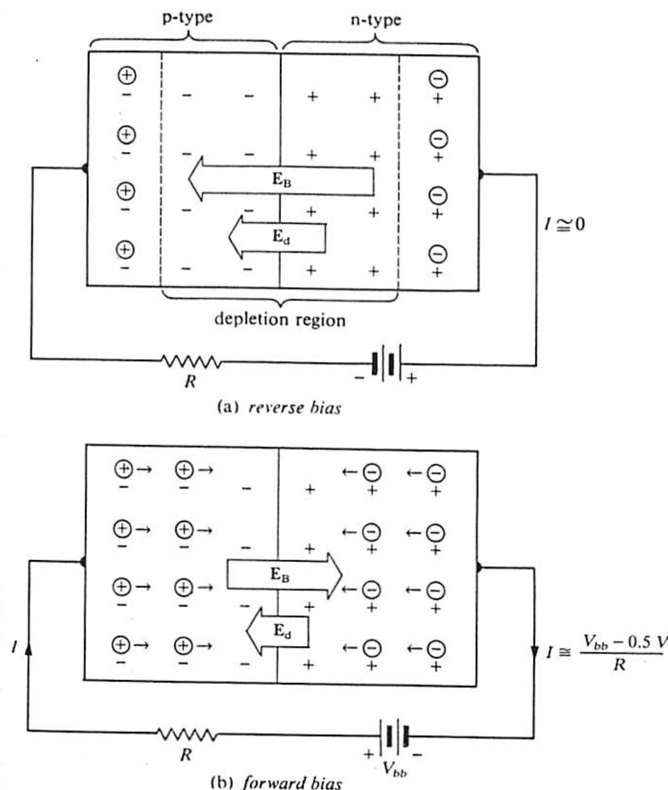


FIGURE 4.14 Biased p-n junction.

material, only a very small current will flow, and the junction is said to be *reverse* biased.

It is only in the depletion region that the p-n junction has a high resistance due to the absence of mobile charge carriers; hence the battery will apply an electric field $\mathbf{E}_B$ across the depletion region only. In the reverse bias configuration $\mathbf{E}_B$ is in the same direction as $\mathbf{E}_d$ and will tend to pull the mobile charge carriers away from the junction, thus increasing the thickness of the charge depletion region, which acts like an insulating slab, and little current will flow through the junction. Actually the preceding

argument applies only to the majority charge carriers. The minority carriers (electrons in the p-type and holes in the n-type) will be attracted *toward* the junction by the electric field introduced by the battery. Thus at the junction the minority carriers will recombine, and the junction will pass current. More minority carriers are continually being created by thermal excitation on both sides of the junction, so this current passed due to the minority carriers will continue to be passed. This is called the *reverse current*. In a germanium junction at room temperature the reverse current is on the order of several microamperes (10^{-6} A), whereas in a silicon junction it is on the order of 10 nA (10^{-8} A). Because the number of minority carriers created depends on thermal excitation across the energy gap, we expect the reverse current of a p-n junction to be temperature dependent. At room temperature a 10°C increase in temperature will approximately double the reverse current in a germanium junction; a 6°C increase will approximately double the reverse current in a silicon junction. At 70°C a typical reverse current for a germanium junction is 100 μA, but for a silicon junction only 1 μA. Thus, if low reverse currents are important, particularly at high temperatures, one *must* use silicon. Virtually all modern transistors and chips are silicon.

Because of the fixed charges present in the depletion layer due to the ionized donor and acceptor atoms, the depletion layer acts like a charged capacitor. Even though the depletion region as a whole is electrically neutral, there is positive charge on the n-type side of the p-n interface due to ionized donor atoms and negative charge on the p-type side due to the ionized acceptor atoms. Thus we have two layers of charge in a parallel plate configuration, which makes up a capacitor. The situation is somewhat different from a parallel plate capacitor where the insulating region between the plates contains no charge. In the reverse biased junction the charges are distributed throughout the volume of the depletion region. As the reverse bias increases in magnitude, the depletion region grows thicker (the capacitor "plates" become farther apart), and the capacitance decreases. An exact calculation shows that the junction capacitance for an abrupt junction is given by

$$C_J = \frac{\mathcal{K}}{\sqrt{V}} \tag{4.10}$$

where $\mathcal{K}$ is a constant depending on the transistor material and geometry, and V is the reverse bias voltage across the junction. C_J may vary from several pF for a low-power transistor especially designed to operate at high frequencies, to several hundred pF for a high-power transistor. In general, with transistors, as with any device, the higher the power handling capability is, the worse the high-frequency response is, because higher powers mean physically larger structures for heat dissipation, which in turn mean larger capacitances.

If a battery is connected across the p-n junction with a polarity to make

the p-type material positive with respect to the n-type material [see Fig. 4.14(b)], a large current will flow, and the junction is said to be *forward biased*. The resistance R is present only as a precaution to limit the current flow. In this case, the electric field from the battery $\mathbf{E}_B$ set up across the depletion layer is in a direction to pull the mobile charge carriers *across* the depletion layer; that is, electrons will be attracted from the n-type material over into the p-type material, and holes from the p-type material over into the n-type material. Recombination will take place when the electrons reach the p-type material and when the holes reach the n-type material; thus a current will flow through the junction. More electrons will continually be injected into the n-type material by the wire connected to the negative terminal of the battery, and electrons will continually be taken out of the p-type material by the wire connected to the positive terminal of the battery. The taking out of electrons from the p-type material is electrically equivalent to injecting positive holes into the p-type material. Thus, a continual flow of electrons and holes moves toward the junction from the n-type and p-type materials, respectively, and the junction passes current. Notice that the electric field $\mathbf{E}_d$ in the depletion layer due to the contact potential opposes the forward bias electric field of the battery, so no current will flow through the junction until the battery voltage applied to the junction exceeds the contact potential. In other words, it takes a finite voltage to "turn on" a forward biased p-n junction, about 0.5 V for a silicon junction. Once the turn-on voltage is exceeded, the forward biased junction presents essentially a short circuit to the flow of current; that is, $I \cong (V_{bb} - V_{\text{turn on}})/R$, where R is the external resistance in series with the forward biased junction.

The turn-on voltage $V_{\text{turn on}}$ is slightly temperature dependent because of the weak temperature dependence of the Fermi energy. In a pure or *intrinsic* semiconductor with no doping, the Fermi level is in the energy gap halfway between the valence and the conduction band. In a doped n-type semiconductor the Fermi level is between the donor levels and the bottom of the conduction band (see Fig. 4.10).

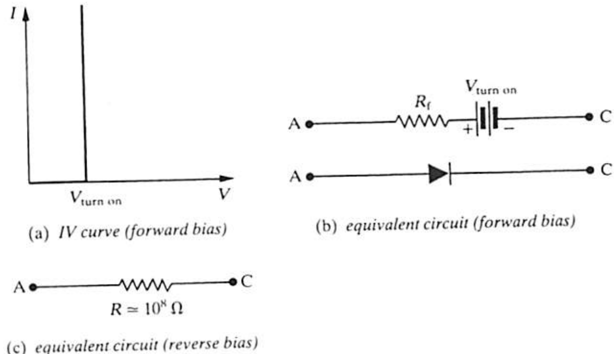
Thus, an increase in temperature lowers the Fermi energy for n-type material, because at higher temperatures a larger fraction of the mobile charge carriers are from electrons thermally excited across the gap from the valence band to the conduction band. A similar argument shows that the Fermi level is raised for an increase in temperature in p-type material. The net result is that if the temperature increases, a *smaller* forward bias voltage is required to maintain a constant current flowing through a forward biased p-n junction. If the temperature rises and the forward bias voltage is maintained constant by the circuit, then the current flowing through the junction will increase. In other words, a temperature rise will lower the effective turn-on voltage. Empirically, for silicon p-n junctions near room temperature, the turn-on voltage *decreases* by approximately 2.5 mV for every degree Celsius temperature rise.

In the reverse biased configuration the electric field within the semiconductor exists only in the depletion region. Thus, the motion of the mobile charge carriers in the rest of the semiconductor is governed mainly by diffusion—there simply is little or no electric field present to "hurry" the charges along. Charge will flow through the semiconductor because of the concentration gradients set up by the depletion layer absorbing charge carriers and the wire contacts injecting more carriers. The charge carriers will then diffuse from regions of greater concentration toward regions of lesser concentration. Thus, there is a certain definite lag in the propagation of a current through a semiconductor, but this lag or *transit time* does not become important until frequencies of the order of tens or hundreds of megahertz or higher are considered. And the diode manufacturers make the physical size of the semiconductor material used as small as possible to minimize the time necessary for a charge carrier to diffuse from the connecting wire to the depletion layer. In some diodes the conductivity of the semiconductor material is deliberately made less so as to have an electric field exist inside the material; this field moves the charge carriers faster than by diffusion alone.

In summary, we have seen that a p-n semiconductor junction will conduct current very well in one direction (the forward direction) and very poorly in the other direction (the backward direction). This one-way type of conduction is called *rectification*, and the device is called a *diode*.

An ideal or perfect diode would present zero resistance in the forward direction and infinite resistance in the backward direction. If we plot a graph of current I passed by an ideal diode versus V the voltage difference across it, we would get a 90° break in the curve as is shown in Fig. 4.15(a).

An equivalent circuit for a forward biased diode is thus a resistance R_f,



(a) *IV curve (forward bias)*

(b) *equivalent circuit (forward bias)*

(c) *equivalent circuit (reverse bias)*

FIGURE 4.15 Ideal diode.

representing the forward dc resistance of the diode, in series with a battery representing the turn-on voltage $V_{turn\ on}$, as shown in Fig. 4.15(b).

The equivalent circuit for a reverse biased diode is simply a large resistance about $10^8\ \Omega$. This simply means that if a reverse bias of several volts exists across a diode, it will conduct a current of about 0.01 μA.

The value of R_f depends on how strongly the diode is conducting. A small Si signal diode might conduct 10 mA with a 0.7-V drop across it. If $V_{turn\ on} = 0.6$ V then $R_f = 0.1\ \text{V}/10\ \text{mA} = 10\ \Omega$. If it conducted 20 mA with a 0.75-V drop across it, then $R_f = 0.15\ \text{V}/20\ \text{mA} = 7.5\ \Omega$.

The current–voltage curve for a real semiconductor diode differs from the ideal diode curve in several ways: The reverse current is not exactly zero because the thermally generated minority carriers will pass current even when the diode is reverse biased; the forward current is finite because the semiconductor material composing the diode has some resistance; and a minimum voltage (the turn-on voltage) must exist across the diode before it will pass appreciable current. The reverse current increases with increasing temperature because there are more minority carriers at higher temperatures. Typical curves for germanium and silicon diodes are shown in Fig. 4.16. It can be shown that the current–voltage equation for a p-n junction is

$$I = I_0(e^{eV_B/kT} - 1) \tag{4.11}$$

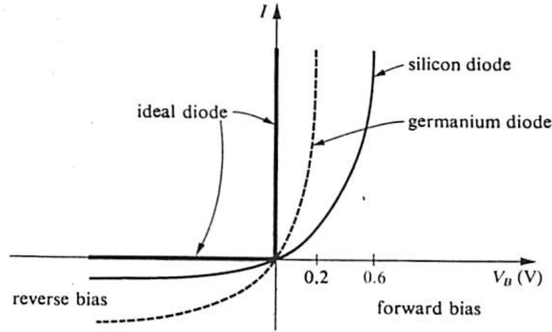


FIGURE 4.16 Current–voltage curves for real diodes.

where V_B is the applied bias voltage, e is the electronic charge, k is Boltzmann's constant, and T is the absolute temperature. V_B is positive for forward bias and negative for reverse bias. I_0 is the reverse current for large reverse bias. Notice that the silicon diode passes much less reverse current than does the germanium diode, and that the turn-on voltage differs

appreciably between the two types. Therefore, in applications requiring an extremely small reverse current, silicon is used, and if one has only a small voltage to turn on the diode, germanium is used. Silicon diodes can be used at temperatures up to about 200°C, whereas germanium diodes can be used only up to about 85°C. Silicon diodes are now almost universally used instead of germanium.

The explanation for the turn-on voltage and the equation for the current–voltage curve for a p-n junction can be obtained from a more detailed energy level picture of both sides of the junction. We recall that

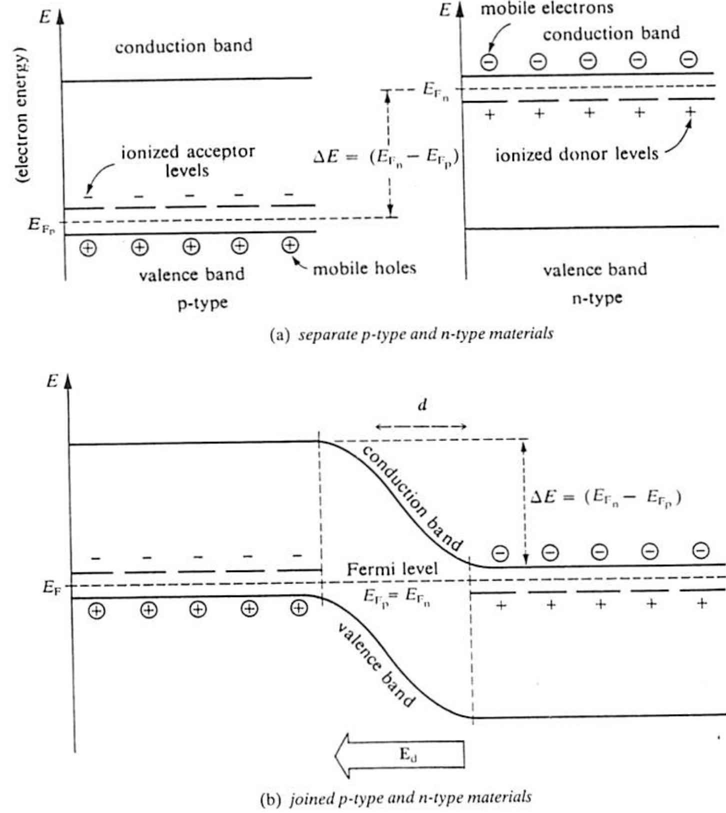


FIGURE 4.17 Energy levels in an unbiased p-n junction.

the Fermi energy for n-type material lies just above the donor levels and in p-type material lies just below the acceptor levels. When any two materials—metal, semiconductor, or insulator—are brought together, their Fermi levels must equalize. If the Fermi levels are not initially equal, then electrons will flow from the material having the higher Fermi level into the other material until the Fermi levels are equalized, as shown in Fig. 4.17(b).

Referring to Fig. 4.17, we can draw some conclusions. The conduction band of the p-type material is higher in energy than that of the n-type material by an amount ΔE equal to the difference in the Fermi energies, which is approximately equal to the gap energy E_{gap}. The electric field $\mathbf{E}_d$ in the depletion region (pointing from n-type to p-type) simply arises from the slope of the bottom of the conduction band as we go across the junction, $E_d \cong \Delta E/ed$. Thus, an electron at the bottom of the conduction band in the n-type material must somehow get ΔE energy to climb up the potential hill and arrive in the conduction band of the p-type material. The turn-on voltage $V_{\text{turn on}}$ of the junction is approximately given by $eV_{\text{turn on}} = \Delta E$, because for any electrons to flow from the n-side to the p-side, they must attain a minimum of ΔE energy, which means a forward bias of at least $V_B = \Delta E/e$ volts must be applied. The reason for the larger turn-on voltage of silicon as compared to germanium p-n junctions is simply that the energy gap (and therefore ΔE) is larger for silicon than for germanium.

There are four types of current which can flow across the junction: two from majority carriers and two from minority carriers. Let J represent current densities due to majority carriers and j current densities due to minority carriers.

Any electrons in the conduction band of the n-type material are majority carriers and are due to the ionized donor atoms. They will flow over to the p-type material only if their total energy is greater than E_V, the energy of the bottom of the conduction band in the p-type material. The number of such electrons is represented by the shaded area in Fig. 4.18 and will provide a current density J_e, which is given by

$$J_e \propto \int_{E_V}^{\infty} N_{cb}(E)\, dE \tag{4.12}$$

where $N_{cb}(E)$ is the density of electrons in the n-type conduction band as a function of energy. If the junction is forward biased (see Fig. 4.18) by an applied voltage V_B, then all energy levels on the n-type side will be raised by an amount eV_B if we assume the p-side is grounded. (Grounding the p-side merely fixes all the energy levels in the p-type material.) Therefore,

$$J_e \propto \int_{E_V - eV_B}^{\infty} N_{cb}(E)\, dE \propto e^{-(E_V - eV_B)/kT} \propto e^{eV_B/kT} \tag{4.13}$$

Let J_{e_0} be the current density when no bias is applied ($V_B = 0$). Then J_e can

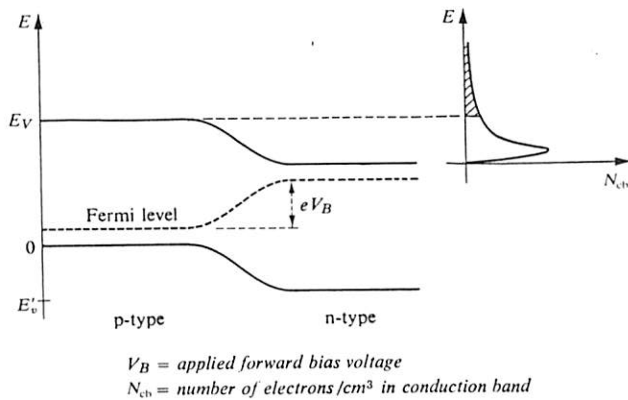


FIGURE 4.18 Forward biased p-n junction energy level diagram.

be written simply as

$$J_e = J_{e_0} e^{eV_B/kT} \tag{4.14}$$

Any electrons in the p-type material thermally excited from the valence band to the conduction band will be minority carriers and will tend to "fall down" the potential hill, thus flowing from the p-type to the n-type material. The resulting current density j_e will be proportional only to the number of minority carriers and, hence, to the temperature so long as there is a downhill potential slope over to the n-type material. The net current flow from these two *electron* currents will then be $J_e - j_e$, where positive current means effective positive charge flowing from the p-type side to the n-type side of the junction.

Any holes in the valence band of the p-type material are majority carriers and are due to the ionized acceptor atoms. They will flow over to the n-type material only if their energy is less than E'_v of Fig. 4.18. An argument similar to that given for the majority carrier electrons in the n-type material shows that the resulting current density J_h from these majority carrier holes is given by

$$J_h = J_{h_0} e^{eV_B/kT} \tag{4.15}$$

Any holes in the valence band of the n-type material are minority carriers and are produced by thermal excitation. They give rise to a voltage-independent current density j_h because the holes spontaneously flow up the potential hill from the n-type side to the p-type side of the

junction. The net current flow from these two *hole* currents is then $J_h - j_h$ with the same sign convention.

The net current density flowing across the junction is the algebraic sum of these four current densities:

$$J_{net} = J_h - j_h + J_e - j_e \tag{4.16}$$

If we now realize that the net current J must equal zero when no bias is applied ($V_B = 0$), we must have

$$J_{h_0} - j_h + J_{e_0} - j_e = 0 \tag{4.17}$$

It can be shown that the hole and electron currents must separately equal zero in equilibrium, so

$$J_{h_0} = j_h \quad \text{and} \quad J_{e_0} = j_e \tag{4.18}$$

Therefore,

$$J_{net} = J_{h_0} e^{eV_B/kT} - J_{h_0} + J_{e_0} e^{eV_B/kT} - J_{e_0} \tag{4.19}$$

Rewriting yields

$$J = (J_{h_0} + J_{e_0})(e^{eV_B/kT} - 1) \tag{4.20}$$

Letting $J_0 = J_{h_0} + J_{e_0}$, we have

$$J = J_0(e^{eV_B/kT} - 1) \tag{4.21}$$

Equation (4.21) is called the *diode equation* or the *rectifier equation* and tells us that as the forward bias $V_B > 0$ is increased, the total current passed through the junction increases rapidly with V_B. And for large reverse bias ($V_B < 0$) the current will decrease to a value of $-J_0$. For large forward bias the increase of J will be somewhat slower than implied by equation (4.21) because of the inherent resistance of the semiconductor material, which we have neglected in this treatment.

Diodes can also be made with a sharp junction between a metal and a semiconductor, the metal acting like a p-type material. Such diodes are called *point contact* diodes. They are superior to junction diodes at high frequencies but can carry far less current than either Ge or Si junction diodes.

A Schottky barrier diode is another example of a metal-semiconductor diode. It is usually made with n-type silicon and aluminum in integrated circuits and is the basis of Schottky and low-power Schottky integrated circuits, which will be discussed in Chapter 12.

The aluminum is evaporated on the n-type silicon, and a depletion region or *Schottky barrier* region is formed (in the silicon), containing only positive ionized donor atoms. The Fermi levels of the aluminum and the n-type silicon equalize, and a rectifying metal-semiconductor junction is formed with a low turn-on voltage of 0.3 V. If the doping of the silicon is very heavy, the junction is ohmic—it conducts almost equally well for both polarities of bias voltages. This technique is used in manufacturing integrated circuits.

All metal-semiconductor contacts do not act as rectifying junctions. Only extremely sharp p-n junctions will rectify. The gradual junctions found between the metal leads and the semiconductor material of diodes and transistors do not rectify; they pass current equally well in both directions. These contacts are usually made by depositing a thin metal coating on the semiconductor and soldering the thin wire to the metal coating.

Another type of diode, called the *zener* or *avalanche* or *reference* diode, can be made to break down at a specified reverse voltage V_z. The breakdown process occurs, briefly, because an electron or hole obtains enough energy (from the reverse bias electric field) between collisions to break a covalent bond in the crystal lattice and thereby create two new charge carriers, which in turn are accelerated and create more charge carriers. This process occurs very sharply at a certain voltage V_z. If one tries to increase the voltage drop across the diode above V_z, then enough additional charge carriers are produced in the diode to increase the voltage drop across whatever resistance is in series with the diode. The net result is that the zener diode draws just enough current to keep the voltage drop across it constant at essentially V_z volts. The zener diode circuit symbol and current–voltage curve are shown in Fig. 4.19. Its principal use is in regulating voltages, for if its reverse current changes from I_A to I_B, the voltage across the diode will remain essentially constant at the zener

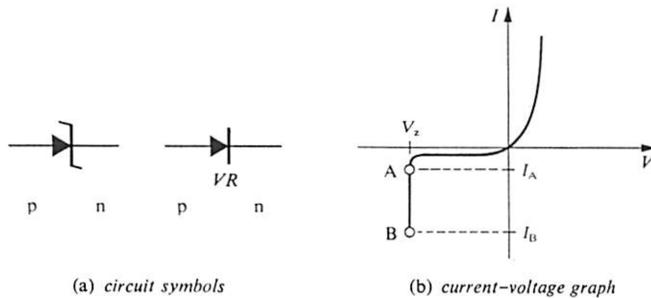


FIGURE 4.19 The zener or voltage regulator diode.

voltage V_z. In a typical zener diode a 1-mA change in current will result in only a 1-mV change in the voltage across the diode, corresponding to a dynamic resistance of only about 1 Ω. Special zener diodes can be made with a very low temperature coefficient to provide an extremely stable reference voltage.

If the amount of impurities is increased to the order of 10^{19} cm^{-3}, then the impurity energy levels merge to form a small band, and the diode can conduct due to a quantum-mechanical "tunneling" process for small forward biases on the order of several tenths of a volt. The tunneling falls to zero for larger forward biases, yielding the current–voltage curve shown in Fig. 4.20. This device is, appropriately, called a *tunnel* diode and is useful

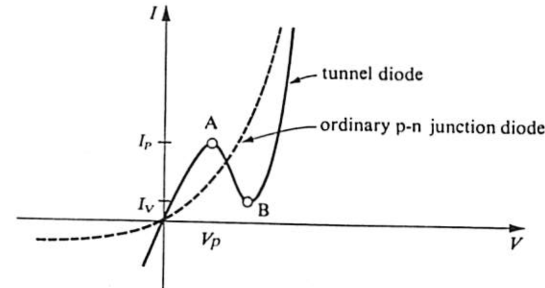


FIGURE 4.20 Tunnel diode current–voltage curve.

not as a rectifier, but because it exhibits a negative dynamic resistance from point A to point B on the current–voltage curve. The peak current I_P is typically from 1 to 10 mA; the valley current I_V is 1 mA or less, and V_P is only about 50 mV. Oscillators and amplifiers utilizing this negative resistance phenomenon can be made. Because the tunneling occurs with majority rather than minority carriers, it can be shown that the tunnel device will operate at extremely high frequencies—up to 10^{10} Hz (10 GHz), which is well up in the so-called microwave or radar region. The tunnel diode is also useful because from points 0 to A its current increases more rapidly with voltage than for ordinary p-n diodes.

Another useful semiconductor device is the *silicon controlled rectifier* (SCR). The SCR is a four-layer device (p-n-p-n or n-p-n-p) with three terminals called the *anode*, the *cathode*, and the *gate*, as shown in Fig. 4.21. The SCR will not conduct in the forward direction until the anode–cathode voltage V_{AC} exceeds a certain value V_C, which depends on the (small) gate current I_G. Once the SCR conducts or "fires," V_{AC} drops to a very small value, and the SCR current I_A is independent of the gate current—the SCR acts just like a forward biased diode. The SCR current I_A drops to zero only

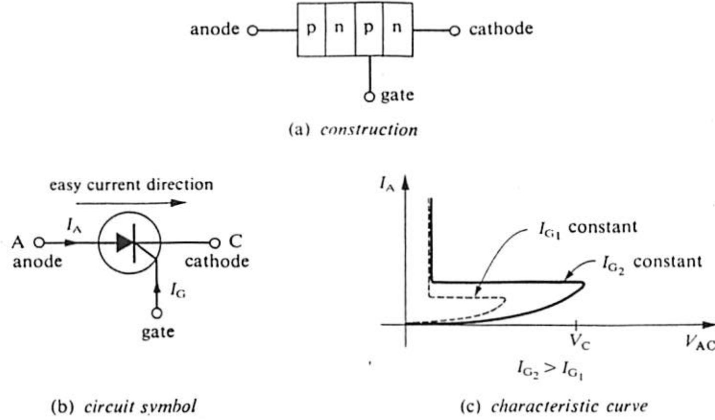


FIGURE 4.21 Silicon controlled rectifier (SCR).

when the polarity of V_{AC} is reversed. The diode can be made to conduct again only when the anode is made positive with respect to the cathode; the exact value that will fire the diode depends on the gate current. The SCR is commonly used to control high-power ac circuits such as electric motors. One SCR can control currents of tens of amperes. The SCR is a solid-state

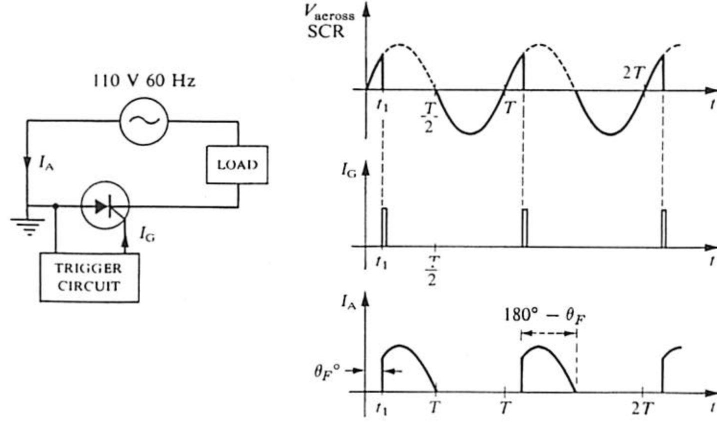


FIGURE 4.22 SCR operation.

equivalent of a thyratron vacuum tube. A gate current of 50 mA can turn on an SCR, which can then carry a current of 20 A.

The basic operation of an SCR in controlling an ac current is to fire the SCR at various phases of the ac voltage across the SCR, as shown in Fig. 4.22. If a trigger circuit supplies a pulse of current to the SCR gate at t_1, the SCR will fire and conduct so long as its anode voltage remains positive with respect to its cathode. The phase angle corresponding to t_1 is often called the *firing angle*, θ_F; the number of electrical degrees during which the SCR conducts equals $180° - \theta_F$. The trigger pulse must last approximately 30 μs or more. The time t_1 of the trigger pulse will determine at what voltage the SCR will fire. The current passed by the SCR will then last from t_1 to $T/2$ on each cycle. By varying the time t_1 at which the trigger pulses occur, the average current passed by the SCR can be adjusted. SCR trigger circuits commonly use small neon lamps that will not conduct until the voltage across the lamp terminals reaches approximately 80 V. One such circuit is shown in Fig. 4.23. The capacitor C is charged up

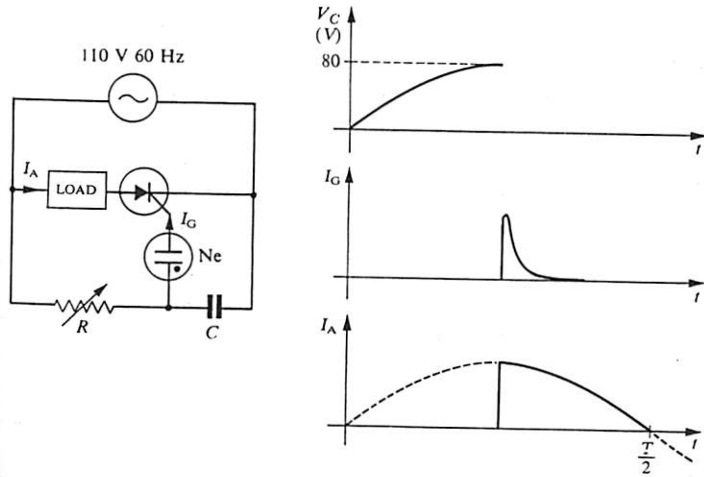
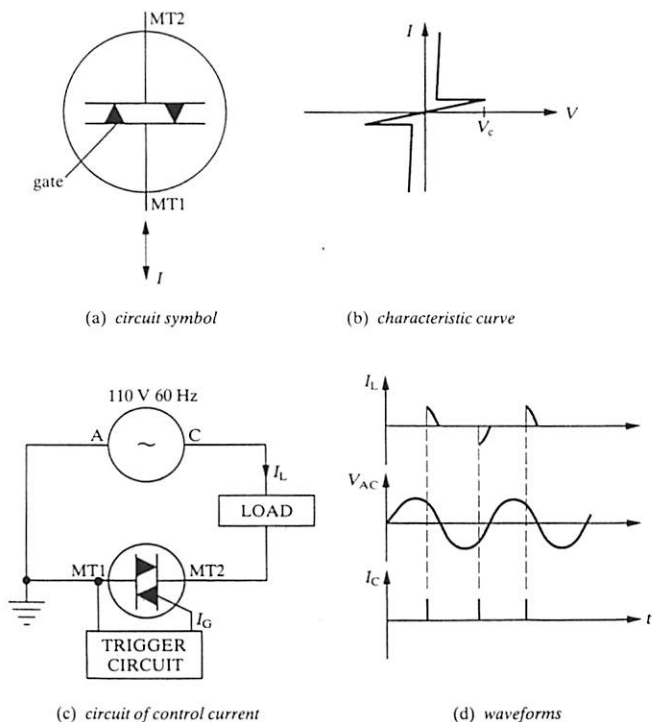


FIGURE 4.23 SCR controlled load.

positively through R by the 110-V, 60-Hz supply. When V_C reaches the firing voltage of the neon lamp, it fires and C is discharged through the SCR gate, thus providing a pulse of gate current to turn on the SCR. The capacitor should be large enough so that the I_G pulse is large enough to turn on the SCR even for a very low anode–cathode voltage, which would

be the situation when the firing angle is close to 180°. When the 110-V ac reverses polarity, a gate pulse is again applied to the SCR, but the SCR does not fire because its anode is now negative with respect to its cathode. As R is increased, it takes a longer time for C to charge up to the neon bulb firing voltage, and θ_F is larger, resulting in a smaller average current through the load.

The *triac* is useful to control ac current. It is basically two SCRs connected so that it can conduct current in either direction, unlike the SCR. The triac symbol and characteristic curves are shown in Fig. 4.24. The two triac terminals are called main terminal #1 (MT1) and main terminal #2 (MT2) instead of anode and cathode. When the triac gate is triggered by a brief current pulse I_C (longer than approximately 50 μs), it



(a) *circuit symbol* (b) *characteristic curve*

(c) *circuit of control current* (d) *waveforms*

FIGURE 4.24 The triac.

conducts current regardless of the polarity of the voltage between MT1 and MT2. The current drops to zero only when this voltage drops to zero, as shown in Fig. 4.24(d). Thus, the phase of the gate current pulse relative to the phase of the ac voltage determines the average current through the load.

4.10 THE PHOTOVOLTAIC DIODE

A p-n junction can be used to convert the solar energy in sunlight directly into electrical energy; the device is called a *photovoltaic diode* or a *photovoltaic cell* or a *solar cell*. Many of the space satellites presently in orbit use arrays of photovoltaic cells to supply power to their electronic circuits.

The basic operation of a solar cell is simply that a photon is absorbed by a semiconductor atom, and an electron is raised across the energy gap from the valence band to the conduction band, leaving a hole behind in the valence band. For silicon the energy gap is 1.09 eV, so the minimum photon energy for this type of absorption is 1.09 eV, corresponding to a photon wavelength of 1140 nm, which is in the near infrared. All visible photons are more energetic than this because the visible range of wavelengths is from approximately 350 to 700 nm.

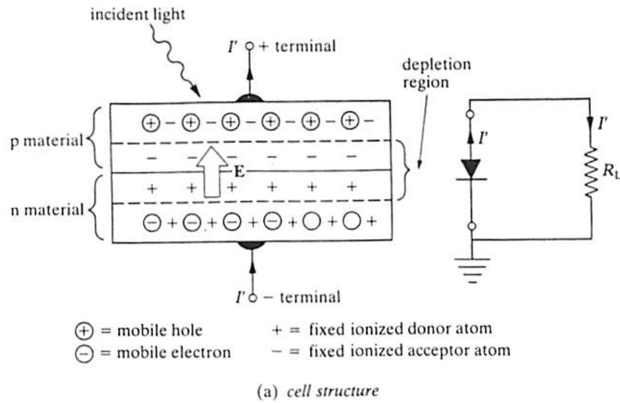
The electric field in the depletion region of the junction then accelerates the electron and hole and thereby produces a *photocurrent* I', which is a *reverse* current, flowing out of the p material, as shown in Fig. 4.25(a). Neglecting recombination of the electrons and holes (which is appreciable in real cells), we determine the cell current to be

$$I' = I_0(e^{eV/kT} - 1) - I_L$$

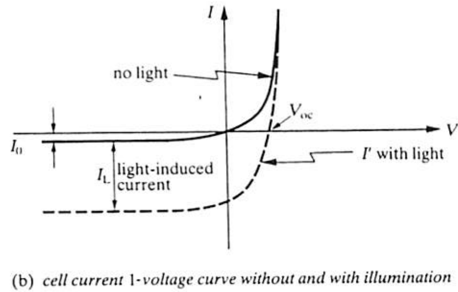
where I_L is the light-induced current.

The maximum voltage of a silicon cell is 1.09 V, the gap energy in volts, because any electron excited by photon absorption to a level above the bottom of the conduction band quickly returns to the bottom of the conduction band by emitting phonons (quanta of vibrational energy). These low-energy phonons produce heat not electrical energy. A practical silicon solar cell has an open-circuit voltage V_{oc} of approximately 0.7 V.

The maximum theoretical power efficiency of a silicon single-crystal solar cell is 27%; real commercial cells have a 12%–14% efficiency. The incident solar-power density at the earth's surface on a clear sunny day (winter or summer) is approximately 900 W/m², so a 12% efficient cell would give approximately 108 W power for each square-meter area. Processes contributing to the inefficiency are reflection from the silicon surface (30% from untreated silicon, much less from a surface with an antireflection coating), recombination of electron-hole pairs, I^2R losses due to bulk silicon resistance and contact resistance, phonon production, blocking of



(a) *cell structure*



(b) *cell current I-voltage curve without and with illumination*

FIGURE 4.25 Photovoltaic cell.

some incident light by the electrical connections at the top surface of the cell, and thinness of the cell (i.e., too thin to absorb all the incident light).

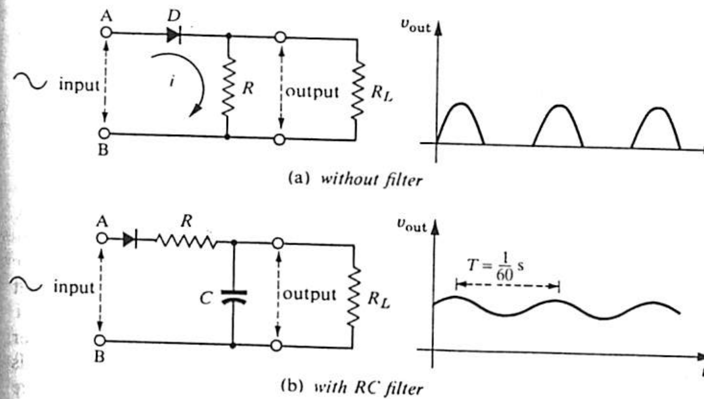
The physics of such a cell is rather complicated. The incoming photons are exponentially absorbed, so a thick cell is better. But a thick cell is much more expensive because of the expense of the high-purity single-crystal material, and recombination can be higher in thick cells. There is a continuous recombination of the electron-hole pairs, so an equilibrium is reached between the photoproduction of the pairs and the recombination. The electrons and holes move by diffusion outside the depletion region, and they drift in response to the electric field inside the depletion region. Finally, Poisson's equation must be satisfied.

An interesting recent development is the manufacture of inexpensive *amorphous* silicon solar cells. The amorphous silicon does not have the long-range crystalline order of a single crystal, and small microcrystals are separated from each other by small *microvoids*. The result of this structure is that there are many (unwanted) energy levels in the forbidden 1.09-eV gap, thus rendering the material useless for a good p-n junction solar cell.

However, when an amorphous silicon film from a glow discharge of SiH_4 is deposited on a substrate, some of the hydrogen is trapped in the amorphous silicon film and effectively quenches or saturates the unwanted energy levels in the gap. The resultant energy level picture is similar to that of single-crystal silicon. Such amorphous silicon can be doped in the usual way to form a p-n junction that can be made into a solar cell. Amorphous silicon solar cells have recently been made with about the same efficiency as single-crystal silicon cells and are widely used to power small solar-powered pocket calculators. Considerable research in amorphous cell technology is presently being conducted.

4.11 DIODE APPLICATIONS

One of the most common uses of the diode is in power supplies where the standard 110-V, 60-Hz ac line voltage is converted into a dc voltage. A simple diode-resistance circuit shown in Fig. 4.26(a) will convert the input ac voltage $v_{in} = V_{AB} = V_0 \sin \omega t$ into a pulsating dc voltage. The explanation is simply that the diode conducts only on the positive half-cycles and



(a) *without filter*

(b) *with RC filter*

FIGURE 4.26 Half-wave diode rectification.

not on the negative half-cycles of the input. We assume in this section that the diode has an infinite resistance in the reverse direction and has zero turn-on voltage. Thus the current i passed by the diode is unidirectional; that is, it is pulsating direct current always flowing in the direction shown in Fig. 4.26(a). The positive output voltage is the iR voltage drop across R. Notice that when the 60-Hz input voltage makes point A negative with respect to point B, the diode is reverse biased, and essentially all the 60-Hz voltage appears across the diode because $i = 0$. For the circuit to function properly the diode must be capable of withstanding this reverse voltage without breaking down. The appropriate diode rating is the *peak reverse voltage* (PRV) or *peak inverse voltage* (PIV). Modern silicon rectifier diodes come with PIV ratings up to 500 V or more. Diodes are also rated according to how much current they may safely pass when forward biased. The maximum current rating usually refers to the average dc current that the diode may safely pass. The peak or surge current rating is generally much higher and refers to the maximum peak current the diode may safely pass in a very short interval of time. Typical modern silicon rectifier diodes cost less than $1 each, can carry average currents of one ampere and peak currents of tens of amperes, and have peak inverse ratings of 500 V.

If a steady dc output voltage is desired, an RC low-pass filter is added, as shown in Fig. 4.26(b). The resistance R also limits the current surge through the diode when the circuit is turned on, when C is completely uncharged. The capacitance C is charged and serves to smooth out the amplitude variations in the output. Another view of the filter is that it passes dc and the lower-frequency Fourier components of the pulsating voltage of Fig. 4.26(a) and attenuates the higher-frequency components. The lower the breakpoint frequency $\omega_B = 1/RC$, the more effective is the filtering. Thus, the larger RC is, the better the filtering will be. The remaining ac variation in the output voltage is called *ripple*. Too large a resistance R cannot be used because the output dc voltage will fall if an appreciable dc current is drawn from the output. For example, if $f = \omega/2\pi = 60$ Hz, then we desire $\omega_B = 1/RC \ll 2\pi \times 60$ or $1/RC \ll 120\pi = 377$. The maximum dc output current usually fixed an upper limit for R. Thus, if we desire the output voltage to fall by less than 1 V as I_{out} increases from 0 to 100 mA, then the maximum voltage drop across R is $(I_{out\,max})R = 1$ V or $R_{max} = 1\text{ V}/100\text{ mA} = 10\ \Omega$. Thus C is determined from

$$C \gg \frac{1}{377R} = \frac{1}{3770} = 2.65 \times 10^{-4}\ \text{F} = 265\ \mu\text{F}$$

We would therefore try to use $C = 1000\ \mu$F or $2000\ \mu$F if the budget permits.

It should also be noted that a larger dc output current implies a smaller load resistance R_L. When the diode is not conducting, the capacitance C is discharging through R_L. Thus, the smaller R_L, the more rapid the discharge of C and the more the output voltage falls before the next positive

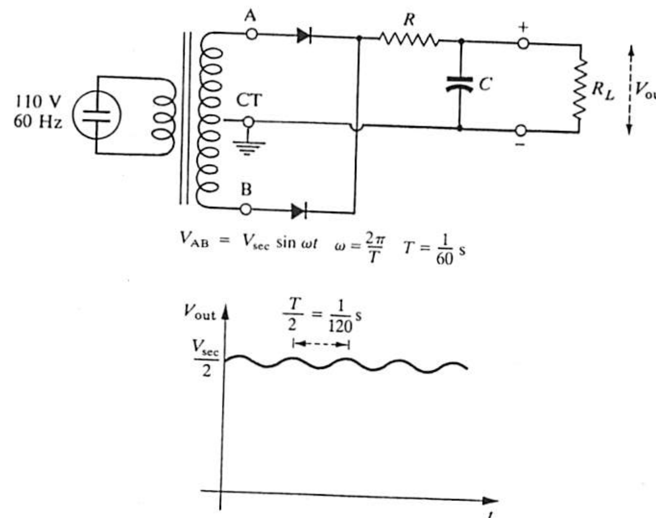


FIGURE 4.27 Full-wave rectifier.

half-cycle, when D conducts and charges C again. In other words, the larger the dc output current, the larger the ripple.

Practical power supply circuits are considerably more complicated than the half-wave circuit of Fig. 4.26. Usually, both halves of the input ac voltage are utilized, either in a full-wave circuit shown in Fig. 4.27 or in the full-wave bridge circuit of Fig. 4.28. Usually a transformer is used to change the 110-V, 60-Hz line voltage to the approximate desired output voltage. V_{AB} is the secondary voltage of the transformer. Notice that the half-wave circuit ripple is 60 Hz, while the full-wave ripple is 120 Hz. Also, the full-wave circuit requires a center tap "CT" on the transformer, whereas the full-wave bridge circuit does not. For the same transformer secondary voltage the half-wave and the full-wave bridge circuits give twice the output voltage that the full-wave circuit does. More current can be drawn from the half-wave circuit than the full-wave circuit without overheating the transformer because current flows through the half-wave circuit transformer only during 50% of the time; that is, the *duty cycle* is 50%. A complete schematic for a practical regulated power supply is given in Appendix A.

A simple AM detector can be made with a diode as shown in Fig. 4.29. The diode rectifies the amplitude-modulated input so that only the positive peaks appear at the input to the low-pass RC filter. If the time constant of the filter is chosen so that $T_c \ll RC \ll T_m$, the carrier oscillations will be

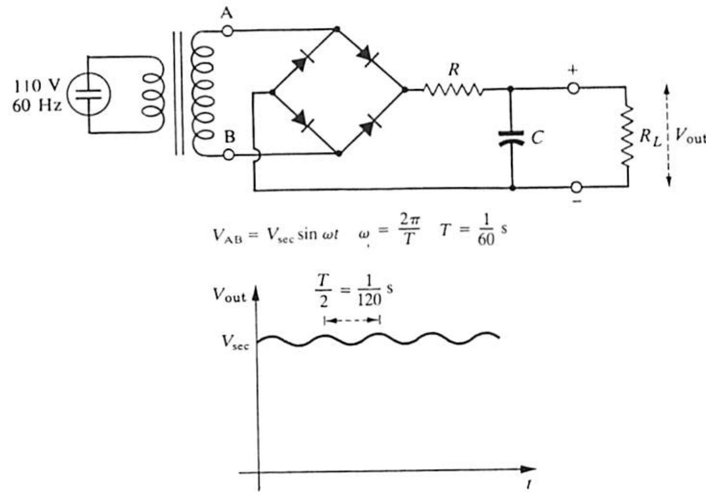


FIGURE 4.28 Full-wave bridge rectifier.

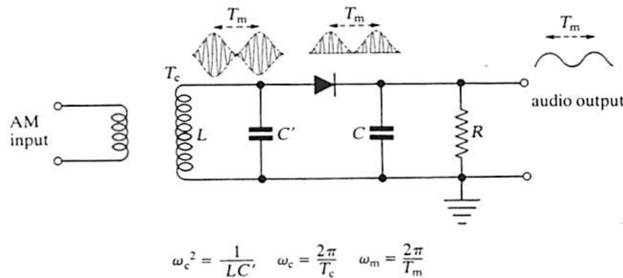


FIGURE 4.29 AM diode detector.

filtered out but not the modulation. Thus, the output voltage across C will follow the audio modulation envelope of the AM wave; the output will be the desired audio. For AM radio the carrier frequency is from 550 to 1500 kHz, so $T_c \cong 1\ \mu$s. The audio modulation is from 20 Hz to 10 kHz, so T_m is from 50 ms to 100 μs. Thus, a reasonable choice for the time constant would be $RC \cong 10\ \mu$s.

A simple FM (frequency modulation) detector can be constructed from diodes using the circuit of Fig. 4.30. The top half of the transformer

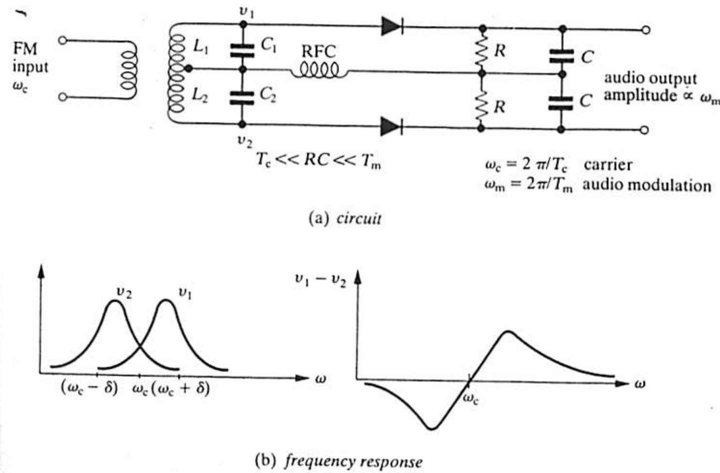
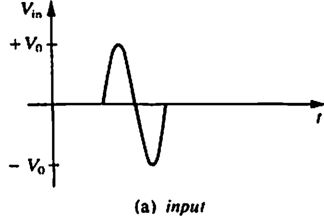


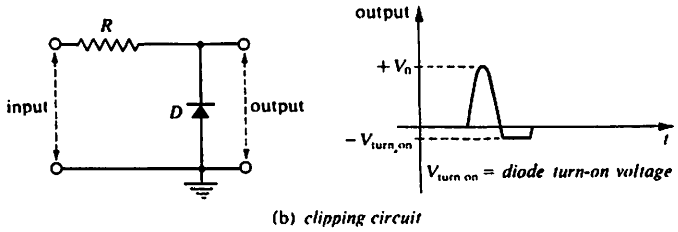
FIGURE 4.30 FM discriminator or detector.

secondary (L_1 and C_1) is turned to a frequency slightly higher than the carrier frequency ($\omega_c + \delta$), and the bottom half (L_2 and C_2) is tuned to a slightly lower frequency ($\omega_c - \delta$). The response of the tuned transformer secondary is shown in Fig. 4.30(b). The difference voltage ($v_1 - v_2$) across the output terminals is proportional to ($\omega - \omega_c$), that is, to how far the frequency has been modulated from the unmodulated carrier frequency ω_c. Thus, if RC is large enough to filter out the carrier ($RC \gg T_c = 2\pi/\omega_c$) and small enough to pass the audio modulation ($RC \ll T_m = 2\pi/\omega_m$), the output ($v_1 - v_2$) will be proportional to the audio modulation signal. In other words, the FM signal will have been *demodulated*.

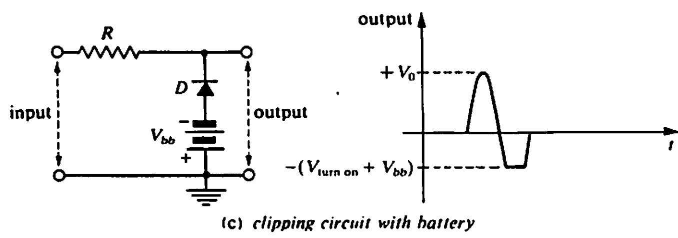
Another common use of diodes is in clipping circuits where the amplitude of a signal must be limited or clipped. In Fig. 4.31(b) the diode conducts only on the negative half-cycle of the input; thus for the negative portion of the input the output is limited to the turn-on voltage. The positive portion of the input is passed through to the output, provided only that $R \ll R_r$, where R_r is the reverse resistance of the diode. If a battery V_{bb} is added, as shown in Fig. 4.31(c), then the diode will not start to conduct



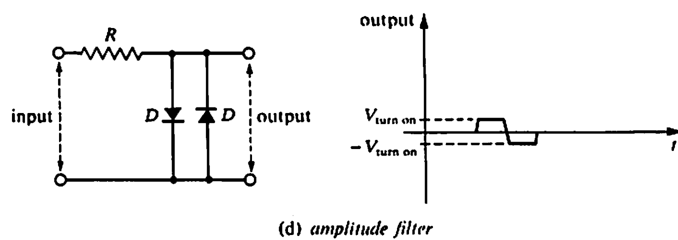
(a) *input*



(b) *clipping circuit*



(c) *clipping circuit with battery*



(d) *amplitude filter*

FIGURE 4.31 Diode clipping circuits.

until the input is more negative than $-(V_{bb} + V_{turn\ on})$. Thus the negative excursion of the output is clipped off at $-(V_{bb} + V_{turn\ on})$ volts. The circuit in Fig. 4.31(d) will pass voltages of either polarity only if the amplitude is less than approximately the diode turn-on voltage; that is, this circuit rejects large-amplitude inputs and passes small-amplitude inputs.

High-voltage transformers are expensive and impractical at voltages much above several thousand volts. A useful diode power supply circuit with no high-voltage transformer is a voltage doubler, shown in Fig. 4.32.

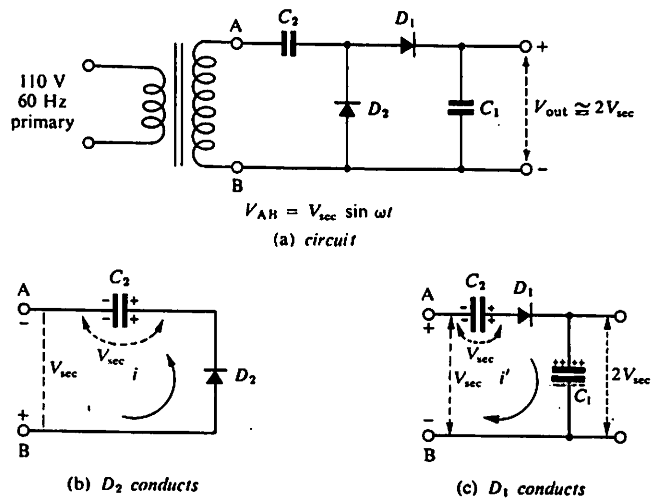


FIGURE 4.32 Voltage doubler.

The dc output voltage will be equal to twice the zero-to-peak input voltage between terminals A and B, which is usually the voltage from the secondary of a transformer. This can be seen by the following argument. If A is *negative* with respect to B, then diode D_2 conducts and diode D_1 does not. Current i flows through D_2 to charge up C_2, as shown in Fig. 4.32(b). The voltage across C_2 will be V_{sec}, the peak secondary transformer voltage. On the next half-cycle when A is *positive* with respect to B, D_1 conducts and D_2 does not. Thus current i' flows as shown in Fig. 4.32(c). C_1 is thus charged up to a voltage $2V_{sec}$ because the peak transformer secondary voltage $V_{AB} = V_{sec}$ and the voltage on $C_2 = V_{sec}$ *add*. C_2 can be an electrolytic capacitor because the voltage drop across it is always of the

same polarity. The value of V_{sec} depends on the turns ratio of the transformer; it can be up to several kilovolts.

Higher output voltages can be achieved with similar circuits. A voltage tripler is possible, and a voltage quadrupler is shown in Fig. 4.33. In

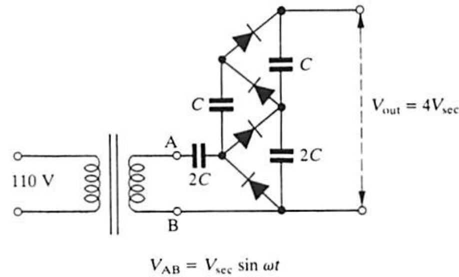


FIGURE 4.33 Voltage quadrupler.

general, these circuits can be extended to yield higher output voltages limited only by the breakdown of the capacitors and the diodes (when reverse biased). The first successful atom-smashing experiment was performed in 1932 by Cockroft and Walton in the Cavendish Laboratory, England, using a voltage multiplier containing vacuum tube diodes and capacitors as the source of high dc voltage. Protons were accelerated through a voltage drop of 250,000 V dc and struck a lithium target, producing disintegration of the lithium nuclei.

A simple diode clipping circuit provides for dc baseline restoration as mentioned in Chapter 2. Recall that when a positive pulse is passed through an RC coupling network (an RC high-pass filter), the output pulse area above and below the zero voltage axis must be equal. For negligible distortion of the pulse shape, the time constant of the circuit should be much larger than the pulse width T. The output pulse, shown in Fig. 4.34(a), then has a long negative tail that decays with the characteristic time constant RCs. If many pulses come along at the input in a time of the order of RCs or less, then the cumulative effect of all the negative pulse tails is to appreciably depress the dc voltage level or baseline of the output as shown in Fig. 4.34(b). The addition of a diode across the output as shown in Fig. 4.34(c) and (d) will essentially cure the problem by providing a low-impedance path to ground for negative outputs. There are many other possible diode clipping or limiting circuits. Several diode circuits useful in computers are discussed in Chapter 12.

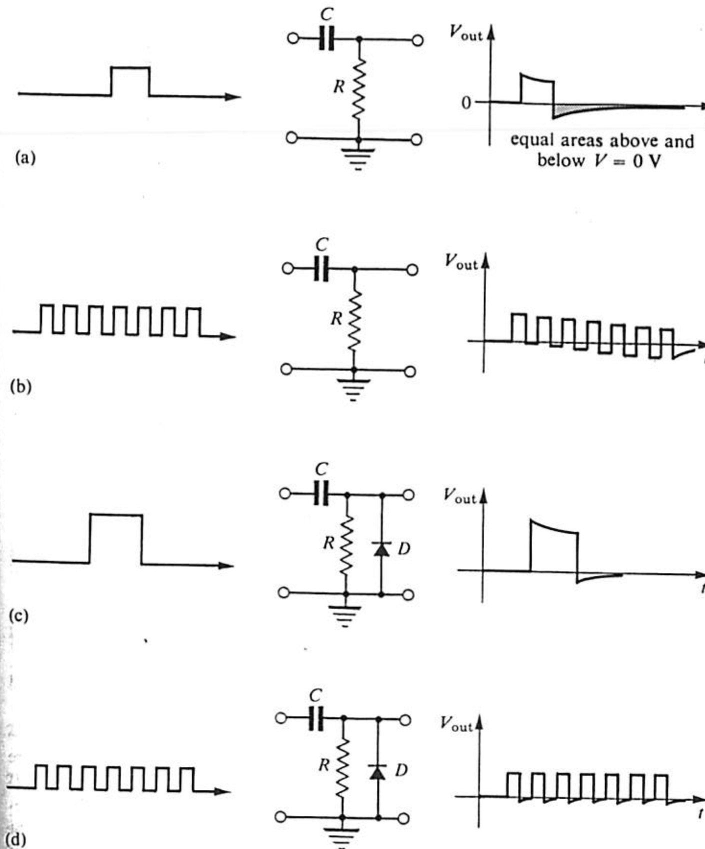
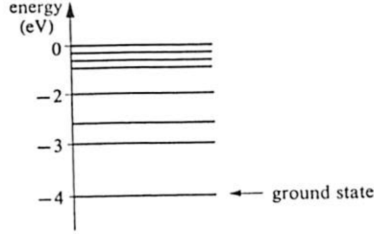


FIGURE 4.34 Elementary diode baseline restoration circuit.

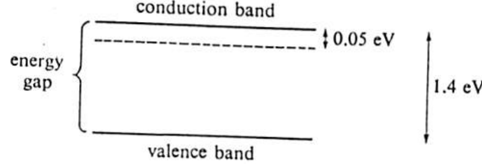
PROBLEMS

1. A new atom, confusium, is discovered with the following energy level diagram. How much work must be done (how much energy must be expended) to ionize a confusium atom in the ground state?

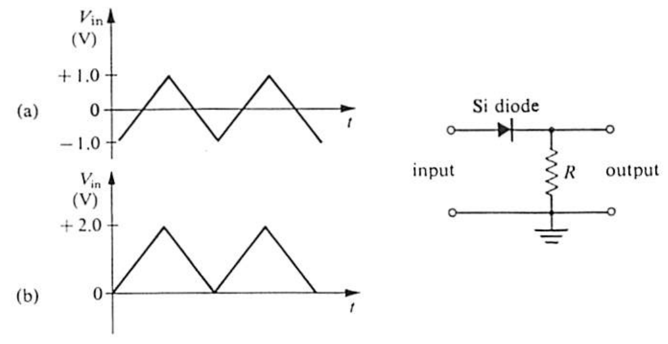


2. In the energy level diagram of confusium in Question 1, at what temperature (approximately) would you expect to have the energy level at −3 eV populated?
3. Approximately how large would the energy gap in a pure semiconductor have to be for it to be a nonconductor at room temperature? What is the physical significance of the energy gap?
4. Calculate an approximate speed for (a) an electron in thermal equilibrium with a silicon lattice at room temperature, and (b) an electron in thermal equilibrium with a germanium lattice at room temperature. The electron mass $m = 9 \times 10^{-31}$ kg.
5. Calculate an approximate speed for the random thermal motion of a silicon atom (atomic weight = 28) in thermal equilibrium at room temperature. Repeat for germanium (atomic weight = 73) at room temperature. 1 amu = 1.7×10^{-27} kg.
6. Calculate the total number of mobile charge carriers in a pure semiconductor if 10^{10} valence bonds/cm^3 are broken on the average due to thermal agitation. Does one distinguish between majority and minority charge carriers in a pure semiconductor?
7. At roughly what temperature would you expect large numbers of electrons to be excited from the valence band to the conduction band in pure silicon? In pure germanium?
8. A silicon crystal is doped with pentavalent arsenic. Is the resulting semiconductor p-type or n-type? If the crystal is doped with trivalent phosphorus, is the resulting semiconductor p-type or n-type?
9. Calculate the average distance between arsenic doping atoms in a silicon crystal if there are 10^{16} As atoms per cm^3.
10. Briefly explain physically why the donor atom energy levels in an n-type semiconductor lie so close to the conduction band.
11. Briefly give a physical interpretation of the Fermi energy. Why can't the Fermi energy lie in the conduction band for a pure semiconductor? What does the Fermi factor mean physically?
12. Briefly explain why the Fermi energy cannot lie below the donor energy levels in an n-type semiconductor.

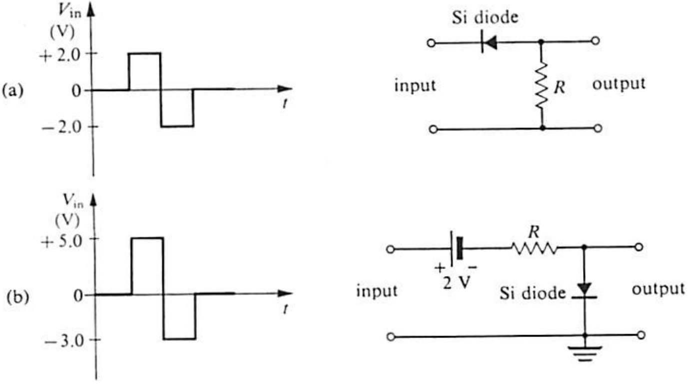
13. Calculate an approximate temperature at which most of the donor atoms would be ionized in a doped n-type semiconductor with the following energy level diagram.



14. At roughly what temperature would you expect a conventional silicon transistor to cease operating as it is cooled down? Explain briefly.
15. Explain why a positively ionized donor atom does not contribute to conduction in an n-type semiconductor.
16. In an abrupt p-n junction with no applied bias, why will there be a small depletion region formed at the junction?
17. Why does an n-type doped semiconductor have many more mobile negative charge carriers than positive carriers? Give typical approximate numbers for silicon and germanium.
18. Carefully state the Pauli exclusion principle. Why does this principle imply that a completely filled band cannot conduct?
19. Describe in words and sketch a diagram of the motion of minority charge carriers in a reverse biased abrupt p-n junction. Include the depletion region.
20. Distinguish among a metal, a semiconductor, and an insulator on the basis of the location of the Fermi level.
21. Set up (do not evaluate) an expression for the number of electrons/cm^3 in the conduction band of an n-type semiconductor with energies more than 0.1 eV above the bottom of the conduction band. Define all symbols carefully.
22. Sketch the energy level diagram (to scale) for an abrupt p-n silicon junction with a 0.3-V forward bias applied. Include the positions of the Fermi levels and the various bands.
23. Sketch a full-wave diode rectifier circuit to produce a −12-V dc output voltage relative to ground. Include approximate numerical values for the transformer and the other components used.
24. Repeat Problem 23, using a full-wave bridge rectifier circuit.
25. Design a diode clipping circuit to clip off negative-going pulses so that the output is never more negative than −3 V.
26. Sketch the output waveform to scale for the following.



27. Sketch the output waveform to scale for the following.



28. Explain why the ripple amplitude in a power supply increases as the load current drawn from it is increased.

29. Diagram a simple half-wave power supply with RC filter to supply an output voltage of approximately −5 V. Specify the transformer secondary voltage and the RC values.

30. Repeat Problem 29 for a full-wave power supply to supply an output voltage of +5 V.

31. (a) Approximately how large an area of solar cells would be required to charge up a 12-V battery with a charging current of 1.0 A? Assume a typical solar cell power efficiency of 12% and 0.7 V per solar cell. [*Hint:* You will need a charging voltage of approximately 14 V to charge a 12-V battery.] (b) Would the solar cells be hooked up in series or in parallel? How much area would each cell have?

CHAPTER 5

The Bipolar Transistor

5.1 INTRODUCTION

In this chapter we will consider the physical construction of the bipolar transistor, explain transistor properties on the basis of how forward biased and reverse biased p-n junctions work, and design a common emitter transistor amplifier circuit.

5.2 TRANSISTOR CONSTRUCTION

A transistor consists of three layers of doped semiconductor material in the order p-n-p, which is called a *pnp* transistor, or in the order n-p-n, which is called an *npn* transistor. Figure 5.1 shows the construction and the circuit symbols used for pnp and npn transistors. The center layer is always called the *base*. The outside layers are called the *emitter* and the *collector*. They look identical in Fig. 5.1, but they are not. Briefly, the collector is thermally connected to the outside world for better heat dissipation, whereas the emitter is not; also, the emitter is more heavily doped.

The boundary between the different types of semiconductors must be abrupt; that is, the type of impurity atom must suddenly change from donor to acceptor as we go from n-type to p-type and from p-type to n-type. Also, the crystal lattice structure must be undistorted or unbroken as we go from p-type to n-type material, or vice versa. One cannot make a transistor by gluing or clamping together separate pieces of doped semiconductor because there would be too many imperfections in the crystal structure at the interface. The entire transistor must be grown from one single crystal with the type of impurity changed abruptly to create the p-n junctions.

Starting with an n-type single crystal of silicon for the collector, a group III p impurity atom such as boron is diffused downward from the top to form the p-type base. Then a group V n impurity atoms such as phosphorus is diffused downward to form the n-type emitter. A thin layer of insulating SiO_2 is formed on the top by exposing the silicon to oxygen gas at a high temperature. A stencil of organic polymer deposited directly on the top surface of the silicon determines where the impurity atoms diffuse