19. Electronics

19.1 Introduction

Electronics is the art of controlling the flow of electrons. It began with the discovery of the ancient Greeks that a piece of amber (ἤλεϰτρον) could attract and hold small objects after being rubbed. The path leading to electrons' controlled removal from matter included the observation early in the eighteenth century that the electrical conductivity of air increases in the vicinity of a hot poker, continued with Franklin's kite experiment relating lightning to static electricity, and culminated in the middle of the nineteenth century with experiments of Crookes who passed electricity between high-voltage plates in evacuated tubes. Edison noticed in 1883 that if he placed a metal plate inside a light bulb, current would flow to the plate if it was at a positive voltage with respect to the filament, but not otherwise. Edison did not think this observation of *rectification* particularly significant, but it has turned out to have as many consequences as his electric lights.

Credit for the discovery of the electron is given to J. J. Thomson, whose experiments on the flow of electricity from heated filaments in evacuated tubes isolated it as a particle with a definite ratio of charge to mass. The name "electron" was proposed by G. J. Stoney in 1894 for the unit of charge equal to 10^{-19} coulombs, and this term gradually superseded Thomson's term "corpuscle" for the new particle. The first practical electronic device was built by J. A. Fleming, who built upon the work of Thomson and Edison to create a *cathode ray tube* with a heated filament capable of rectifying oscillating currents. He called it a "valve," but it is now better known as the *diode*, and is depicted in Figure 19.1. Commercial radio

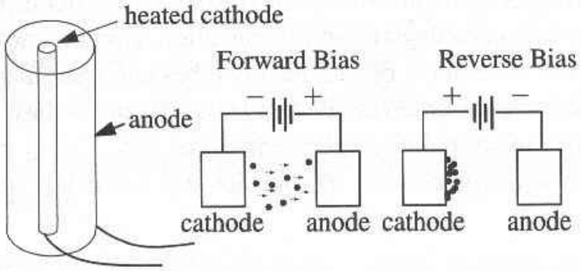

Figure 19.1. The essential action of a diode is to send current in one direction in response to an applied voltage, and not the other. The origin of this asymmetry is the fact that metals at elevated temperatures emit electrons long before they emit positively charged ions. A heated cathode therefore sends off an appreciable current toward a positively charged anode, but almost none toward a negatively charged anode.

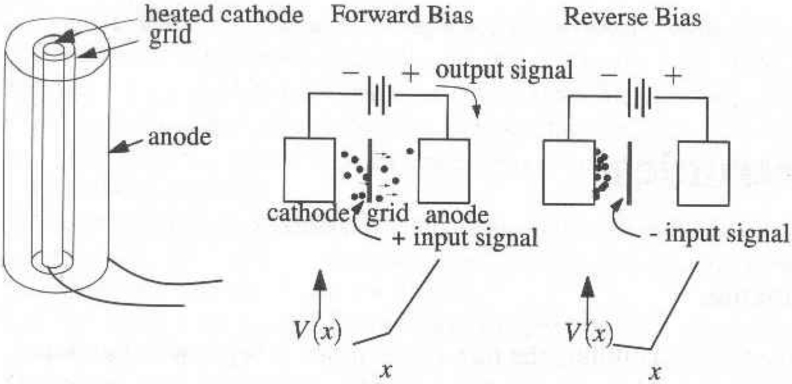

Figure 19.2. The triode serves not only to rectify current, but also to amplify small signals. It accomplishes this task through the interposition of a grid between cathode and anode. The potential between cathode and grid determines whether electrons begin the journey between cathode and grid at all. However, once electrons arrive at the grid, they discover that a much larger positive potential awaits them at the anode and accelerate toward it. Because only a small proportion of the electrons enters the grid, small grid currents control large cathode→anode currents.

transmission became feasible soon after with the invention by L. De Forest of the *triode*, shown in Figure 19.2. Rectification is essential to practical radio transmission because the time average electrical field of a propagating radio wave is zero, and even if the amplitude of such a wave is modulated to encode sound, it cannot directly drive a speaker. Once the signal is passed through a rectifier, time averages no longer vanish, and the signal can easily be decoded. The triode was essential not only because it allowed amplification of weak signals, but also because by feeding a portion of the output back in to the control grid, it could be made into a powerful and stable source of radio-frequency oscillations.

Up through the 1970s much of electronics consisted in the study of cathode ray tubes. They have now almost entirely been superseded by semiconductor devices, which are much more reliable, and have slowly managed to capture even high-power and high-frequency applications that at first seemed out of reach. However, the basic concepts of controlling current, rectification, amplification, and switching all first developed in the context of cathode ray tubes and were then taken over and further developed by semiconductor descendants. Even the basic physics of the various devices has many points of similarity. For this reason, it is advisable to begin the study of electronics with the physics that made the cathode ray tubes possible.

19.2 Metal Interfaces

As sketched in Figure 19.1, the cathode ray tube diode relies upon the fact that a heated piece of metal emits electrons, but not positively charged ions. This effect becomes most clearly visible when air is evacuated from the region in which the electrons are to travel, and for this reason cathode ray tubes are also known as *vacuum tubes* or *electron tubes.* The physical question that needs to be answered is how a metal surface in contact with vacuum emits electrons as a function of temperature and electrical potential.

19.2.1 Work Functions

All the calculations of electronic energy levels up until now have been carried out relative to one another. For example, the Fermi level was sometimes calculated relative to the energy of the lowest single-particle electronic level, and in the band structure diagrams of Section 10.4, the Fermi level was defined to be zero. None of these calculations answers the question of the amount of energy needed to remove an electron from an electrically neutral solid in vacuum. This energy is defined to be the *work function* and is often denoted by ϕ. It can be measured by optical methods to be discussed in Section 23.6.1, or by properties of thermal emission to be discussed immediately below. To calculate it requires understanding what happens when an electron is dragged through the interface between metal and vacuum.

The work function must be contrasted with the chemical potential, which is the energy required to take an electron from the bulk and remove it to infinity. Inspection of Table 23.2 shows that the experimentally measured energy required to pass an electron through one crystal surface typically differs by 10–20% from the energy required to move it through an inequivalent one. As the energy required for transit from bulk to infinity cannot depend upon path, it is necessary to establish carefully what the experiments actually determine.

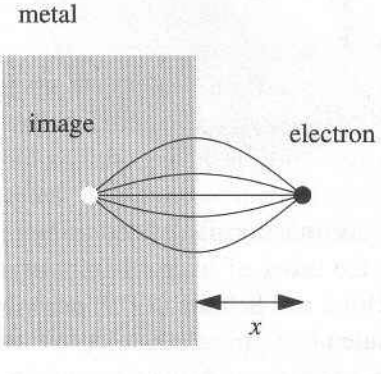

Figure 19.3. An electron at a distance x from a metal surface is attracted to the surface by an image charge of opposite sign which guarantees that electric field lines will be normal to the surface.

The electrostatic potentials contributing to the work function operate on three separate length scales:

Atomic. In passing through the surface of a crystal, an electron passes through a highly inhomogeneous environment where the crystal terminates. Although the surface is almost completely electrically neutral, there is always a strong dipole layer. The electron is buffeted by strong forces as it passes through this layer, but the range over which these forces operate is only on the order of a few lattice spacings, due to the effectiveness of screening in a metal.

Micron. After the electron exits the metal, it interacts with an image charge, whose presence enforces the boundary condition that the tangential electric field vanish on the surface, as shown in Figure 19.3. The force F at distance x from the surface is

$$F = \frac{e^2}{(2x)^2}, \tag{19.1}$$

which implies that the electron has a potential energy

$$U(x) = -\frac{e^2}{4x} = -\frac{1}{x}3.3\cdot 10^{-2}\mu\text{m eV}. \tag{19.2}$$

Work functions are on the order of several electron volts, and therefore further changes due to the image charge energy become negligible by the time an electron has traveled a few microns from the surface.

Macroscopic. Because different crystal faces have different dipole layers, the regions outside them must be at different electrical potentials, produced by minute shifts in electron density near the crystal surface. The existence of this potential is absolutely necessary, because there is no other way to bring an electron out of the bulk through one surface, return it through another surface, and have it return to the original bulk energy. The spatial scale for the variation of this potential is the size of the crystal itself.

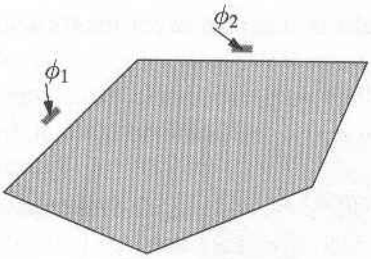

Figure 19.4. The work function is defined as the energy needed to remove an electron from the bulk of a metal, and bring it within about a micron of a particular surface.

Therefore, as indicated in Figure 19.4 the work function is defined to be the energy of electrons brought out to distances on the order of microns from crystals whose dimensions are larger than microns. Hölzl and Schulte (1979) describe many additional complications that can arise in attempting to calculate or to measure work functions, such as what happens when surfaces are rough, or contain a layer of adsorbate atoms.

19.2.2 Schottky Barrier

Equation (19.2) fails when an electron is too close to the crystal surface, because the potential energy U diverges, and more realistic calculations require explicit description of the electronic surface states of a metal. However, it is adequate for the purpose of estimating the effect of an externally applied electrical potential on the work function. Suppose that a positively charged metal plate is placed at a large distance to the right of the metal surface, creating a linear electric field of strength $-E$. Now the potential energy $U(x)$ of the electron is

$$U(x) = -\frac{e^2}{4x} - e|E|x, \tag{19.3}$$

which creates a barrier, shown in Figure 19.5, at distance

$$x_0 = \sqrt{\frac{e}{4|E|}} \Rightarrow U(x_0) = -e\sqrt{e|E|}. \tag{19.4}$$

Therefore an externally applied electric field changes the barrier restraining an electron within the metal by $e\sqrt{e|E|}$.

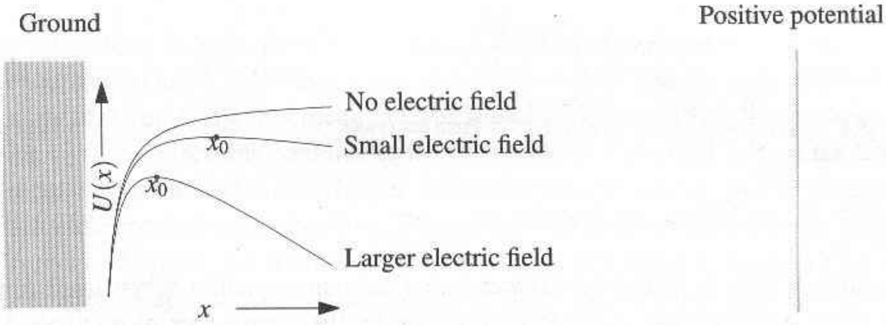

Figure 19.5. An externally applied electric field, created by placing at a distance a large plate at an elevated voltage, lowers the barrier an electron must surpass in order to exit a metal.

Having first found the effect of applied electric fields, the next goal is to examine the *thermionic emission* of electrons that results from heating the metal. This task is accomplished by considering the electrons in the metal to be in equilibrium with a dilute gas of electrons hovering outside it. For simplicity the electrons are treated within the semiclassical approximation, which makes it possible to speak of the probability for an electron to have wave number $\vec{k}$ at position x outside the metal

$$f_{x\vec{k}} = \frac{1}{e^{\beta(\mathcal{E}^0_{\vec{k}} + U(x) - \mu)} + 1}. \tag{19.5}$$

Compare with Eq. (17.16). Here it is more convenient to keep μ constant and describe the spatial change in potential through $U(\vec{r})$.

The probability of finding electrons does not vanish as x travels far from the metal. Finding the vacuum full of electrons may seem unacceptable, but is an inevitable consequence of the fact that no solid or liquid can ever be in equilibrium with a vacuum at nonzero temperature. Entropy always favors total evaporation. However, a solid can exist in equilibrium with a dilute vapor of a particular concentration, and that is what Eq. (19.5) implies for the electrons in a metal. The properties of the electron vapor are fixed by the observation that because it is in equilibrium with the metal, the two must have the same chemical potential. Taking "far from the metal" to denote distances on the order of a micron, the chemical potential must be replaced by the negative of the work function, $\phi = -\mu > 0$. Work functions are typically on the order of several electron volts, as shown in Table 23.2, so for temperatures much less than 10 000 K, Eq. (19.5) can be replaced by

$$f_{x\vec{k}} \approx e^{-\beta(\mathcal{E}^0_{\vec{k}}+U(x)+\phi)}. \tag{19.6}$$

In order to find the current drawn from the metal in the presence of an applied electric field, one should write down the Boltzmann equation and calculate the nonequilibrium function $g_{x\vec{k}}$. However, it is adequate to consider a simple approximation, which is to assume that the electron gas is in equilibrium at all points to the left of x_0 and that all electrons that reach x_0 and are traveling to the right escape over the barrier and run off as a current. This idea predicts a current

$$j = -e \exp\{-\beta[\phi + U(x_0)]\} \int [d\vec{k}]\, \frac{\hbar k_x}{m} \theta(k_x) e^{-\beta\hbar^2 k^2/2m} \quad \text{For } [d\vec{k}]\text{, see Eq. (6.15).} \tag{19.7}$$

$$= -\mathcal{A}T^2 \exp\left\{-\beta\Big[\phi - e\sqrt{e|E|}\Big]\right\}, \quad \text{From Eq. (19.4).} \tag{19.8}$$

where

$$\mathcal{A} = \frac{em}{2\pi^2\hbar^3} k_B^2 = 120.2 \text{ A cm}^{-2}\text{ K}^{-2}. \tag{19.9}$$

Equation (19.8) is called the *Richardson–Dushman* equation when used for $E = 0$, while the reduction of the work function by the square root of an applied field is called the *Schottky* effect. The current does not vanish when the electric field E goes away, which means that if one places a cold grounded plate at some distance from the heated metal and provides a path for the electrons departing from the metal to return to it, current will flow through the vacuum even in the absence of a voltage difference. The factors outside the exponential in Eq. (19.8) are not particularly to be trusted, but the exponential scaling with temperature and electric field can be verified experimentally and can be used to measure the work function ϕ. Equation (19.9) was derived by Schottky (1938), and it provided theoretical underpinning for the development of cathode-ray tube electronics.

19.2.3 Contact Potentials

Whenever two dissimilar materials are brought together, charge moves between them. The reason is that they have in general different work functions, and electrons from the material with the smaller work function rush into the material with the larger one. As this process occurs, charge builds up in the second material, and at some point Coulomb repulsion brings the charge transfer to a halt. The effects of Coulomb repulsion can, however, be minimized if the electrons that flow to the second material are located as close as possible to the (positive) holes flowing to the first material. For this reason, the electrons and holes arrange themselves as surface charges along the interface between the two materials, the electrons on the side of the second material, the holes on the side of the first. Variations on this basic scenario follow mainly from the widths of the regions with surface charge. When two metals are brought in contact, the regions with excess charge have atomic dimensions. When semiconductors are brought into contact, charge densities are much smaller than in metals, and the length scales over which charges build up turn out to be much larger, as will be shown in Section 19.4.2.

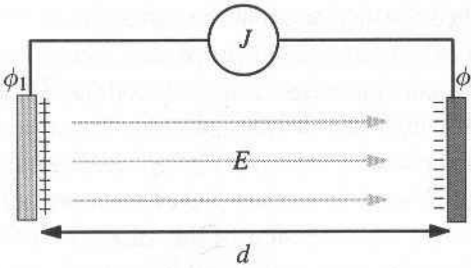

Figure 19.6. Two metals with work functions $\phi_1 < \phi_2$ are arrayed as plates of a capacitor, and charge is allowed to pass between them through an ammeter. By measuring the current J and from it deducing the total charge transfered, the difference between the metals' work functions is determined using Eq. (19.11).

For now, consider the case of two metals in contact. To bring an electron from one metal to its neighbor costs energy ϕ_1, but recovers energy $-\phi_2$, where ϕ_1 is the work function of the first metal, and ϕ_2 is the work function of the second. Assuming $\phi_2 > \phi_1$, electrons continue to flow from metal 1 to metal 2, until the Coulomb repulsion of the additional charges added to metal 2 cancels out the advantages of the difference in work function. This difference in electrical potential is called a *contact potential.* As shown in Figure 19.6, by using two different metals as two plates of a capacitor and then connecting them with a wire, one can measure the difference in their work functions. Because the metals are arrayed as a capacitor, the electrical potential difference V between them is

$$V = Ed = 4\pi\sigma d, \tag{19.10}$$

where σ is the magnitude of the surface charge on each of the metals, and d is the spacing between them. In equilibrium, the potential energy $-eV$ needed to bring an electron from one plate to another equals the difference in work functions, so

$$\phi_2 - \phi_1 = 4\pi e\sigma d. \tag{19.11}$$

One way to measure the difference in work functions is simply to measure the total current that flows between two metals at known spacing after they are connected by a wire. A more accurate procedure is to find an external potential difference imposed between the two metals so that no current flows when the spacing between the two metals is changed slightly. This potential difference must be just the difference in work functions shown in Eq. (19.11).

Double Layers and Reconstruction. Expressions (19.4) and (19.11) provide relations for metals in contact with vacuum or each other by cleverly evading questions of what happens at short length scales. Equation (19.4) must break down when electrons come within a few angstroms of a metal surface, while Eq. (19.11) should fail when two metals come closer than within a few angstroms of each other. Qualitatively, however, each of them is correct. For angstrom-scale separations between metals, Eq. (19.11) predicts that a *double layer* of charge will build up, with charge density on the order of $5 \cdot 10^{-3}$electrons/Å^2. Compared to the the normal density of electrons along any surface of a metal, this number is not particularly large. However, the electric fields involved are on the order of 1eV/Å, and they are enormous compared to fields normally generated in the laboratory. The double layer of surface charges is a dipole layer, and one can view the work function of a metal generally as arising from the presence of such layers at the surface.

Band structure programs are able to calculate detailed properties of surfaces with a fair degree of success, and they find such quantities as work functions. Early work along these directions was described by Lang (1973) and Appelbaum and Hamann (1976), and a more recent review is given by Zangwill (1988). Because the computer programs depend upon using Bloch's theorem, they must have a periodic crystal in which to carry out the calculations. One solution of this difficulty is to carry out calculations with a unit cell such as depicted in Figure 19.7.

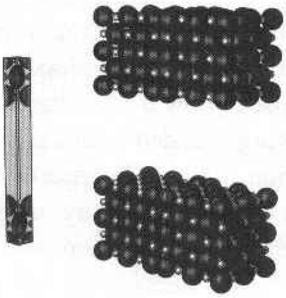

Figure 19.7. Band structure programs study surfaces by creating a unit cell (left) that upon repetition in all directions produces an array of slabs (right). The thicker the slabs, the more realistic an account of surface and bulk states the program can provide. The figure does not show any surface reconstruction, which often occurs and whose analysis is a frequent aim in the calculations.

19.3 Semiconductors

The beginnings of modern electronics lay in the control of current rectification by the cathode-ray diode. However, the cathode-ray tube did not provide the first case in which rectification was observed. It was seen independently by Braun (1874) and by Schuster (1874). Braun conducted experiments in which a crystal such as ferrous sulfide was contacted with a very thin wire, and the resistance was measured as a function of the direction in which current was flowing. Such point junctions do rectify current, although the effect is quite small and had no immediate practical consequences.

The first diodes were produced by placing a whisker of metal in contact with a semiconductor crystal, and are described by Henisch (1957). Early devices could rarely compete with cathode-ray tubes, because they were still comparatively inefficient and unpredictable. In order to make them work at all, it was sometimes necessary to slide the whisker around until a region of good contact was found at random. The progress of basic research into solid-state physics in the 1930s and late 1940s found the cause of the apparent unpredictability of semiconductors: the presence of certain crucial impurities in extremely small quantities. Once the role of these impurities was understood, and methods developed to control them, diodes and triodes based upon semiconductors took part in a remarkable development that eventually displaced the cathode-ray tubes that had inspired them, and they led electronics to a level of extraordinary complexity.

The discussion will begin with the simplest basic physics, and gradually decorate it with additional effects, until the mechanisms responsible for semiconductor electronics emerge. The starting point is the statistical mechanics of pure semiconductor crystals, followed by statistical mechanics of semiconductor crystals doped with small quantities of impurities, and finally the theory of conductivity in junctions between differently doped semiconductors.

19.3.1 Pure Semiconductors

Preliminaries. Semiconductors are bad insulators. At zero temperature all electrons lie within completely filled valence bands separated from conduction bands by an energy gap of magnitude $\mathcal{E}_g$. Important features of the bands of silicon, germanium, and gallium arsenide appear in Figure 19.8. One would expect these materials simply to be insulators, except that the energy gap is small, on the order of 2 eV or less. At room temperature the occupation of the conduction band is proportional to

$$e^{-\beta\mathcal{E}_g/2} \sim 10^{-10}\,. \quad \begin{array}{l}\text{For } \beta = 1/k_BT = 1/40 \text{ eV and } \mathcal{E}_g = 1 \text{ eV. The factor}\\ \text{of 1/2 is a bit surprising, but will emerge from analysis.}\end{array} \tag{19.12}$$

Because thermal excitation provides exponentially growing numbers of mobile charge carriers, the electrical conductivity of semiconductors grows exponentially

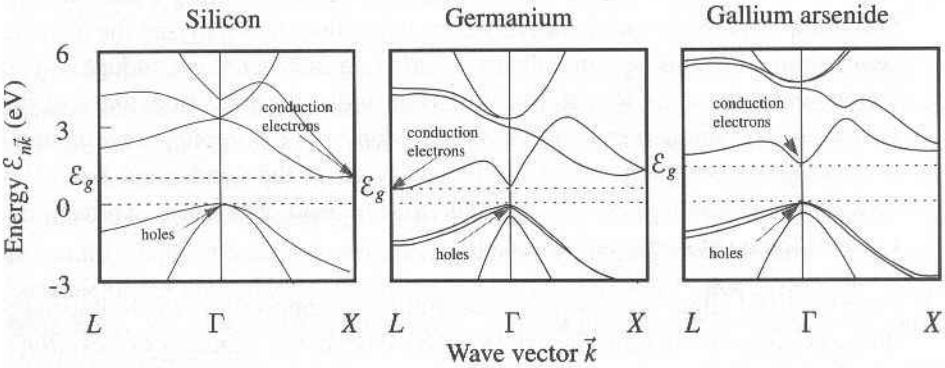

Figure 19.8. Essential features of band structures of silicon, germanium, and gallium arsenide. All have band gaps on the order of 1 eV. The bottom of the conduction band for silicon and germanium does not lie at Γ, so these materials have an indirect gap. Gallium arsenide, by contrast, has a direct gap. These diagrams are extracted from Figures 23.15 and 23.16, which contain information on how they were obtained.

with temperature, in contrast with metals where scattering generally reduces conductivity as temperature goes up. As the band gap $\mathcal{E}_g$ sinks below 1 eV, thermal excitation becomes a sufficiently important source of carriers that the semiconductors conduct at room temperature. More important is the fact that the electrical properties of semiconductors are enormously sensitive to the presence of certain types of impurities, which make their presence felt even at concentrations on the order of one part in 10^{10}. Before the role of impurities was understood, semiconductors seemed capricious and unreliable. Now that they are not only understood but can be controlled, the impurities are employed to give semiconductors tremendously interesting and variable electrical transport properties, with which the electronics industry has developed and grown for over four decades. The word "impurity" connotes something undesirable, so one stops using it in reference to elements intentionally added to semiconductors and refers to "dopants" and "doping" instead.

Band Structure of Semiconductors. Because of the great importance of the energy gap in semiconductors, a few words are in order on how it is measured and calculated. In fact, neither experimental measurement nor theoretical calculation is straightforward. The most precise experimental technique is optical absorption. According to a simple band-theory picture, light falling upon a semiconductor should pass through unimpeded until the energy of a photon is adequate to create an excitation of energy $\mathcal{E}_g$, after which absorption should rapidly increase. The actual story of what happens in such experiments is sufficiently complex and interesting that it is deferred to Chapters 21 and 22. Some of the effects should, however, be mentioned now.

1. Any transition involving a photon must conserve not only energy but also momentum. The momentum carried by a photon turns out to be negligible compared with that of typical electron states. In Figure 19.8, the lowest-energy spot in the conduction band of silicon lies at about $8/10$ of the way toward X, while the highest-energy spot in the valence band lies at Γ. An electron occupying a state near X cannot transfer to Γ simply by emitting a photon. The transition is therefore comparatively rare, with phonons supplying the missing momentum. For this reason, silicon is called an *indirect semiconductor,* as it has an *indirect gap.* Germanium is also an indirect semiconductor, and the bottom of its conduction band lies at L. Many optical applications demand a *direct semiconductor,* where the lowest point of the conduction band lies directly above the highest point of the valence band. For these applications, GaAs is the most important material.

2. Near the band edge, where optical absorption is supposed to vanish, it usually displays one or more thin sharp peaks. These peaks are signatures of *excitons,* which are bound electron–hole pairs whose energy can sit slightly below any states describable in the one-electron picture.

3. Photons whose energy lies below the band gap and out of range of excitons continue to be absorbed, at a rate that decreases exponentially the farther they lie below the band edge. This absorption is due to impurities and fluctuations.

Table 19.1. Semiconductor data

Com- pound		$\mathcal{E}_g$ (eV)	$d\mathcal{E}_g/dT$ (eV/K)	n_i (cm^{-3})	ϵ^0	$m_n^\star$ (m)	$m_{ph}^\star$ (m)	$m_{pl}^\star$ (m)	μ_n (cm^2/V s)	μ_p (cm^2/V s)
Si	i	1.11	$-9.0\cdot10^{-5}$	$1.02\cdot10^{10}$	11.9	1.18	0.54	0.15	1350	480
Ge	i	0.74	$-3.7\cdot10^{-4}$	$2.33\cdot10^{13}$	16.5	0.55	0.3	0.04	3900	1800
GaAs	d	1.43	$-3.9\cdot10^{-4}$	$2\cdot10^{6}$	12.5	0.067	0.50	0.07	7900	450
SiC	i	2.2	$-5.8\cdot10^{-4}$		9.7	0.82	1		900	50
AlAs	i	2.14	$-4\cdot10^{-4}$	$2\cdot10^{17}$	10.0	0.5	0.5	0.26	294	
AlSb	i	1.63	$-4\cdot10^{-4}$		12.0	0.3	1	0.5	200	400
GaN	d	3.44	$-6.7\cdot10^{-4}$	$2\cdot10^{17}$	12.0	0.3	1		440	
GaSb	d	0.7	$-3.7\cdot10^{-4}$	10^{14}	15.7	0.05	0.3	0.04	7700	1600
InP	d	1.34	$-2.9\cdot10^{-4}$	$1.2\cdot10^{8}$	15.2	0.073	0.6	0.12	5400	150
InAs	d	0.36	$-3.5\cdot10^{-4}$	$1.3\cdot10^{15}$	15.2	0.027	0.4	0.03	30 000	450
InSb	d	0.18	$-2.8\cdot10^{-4}$	$2.0\cdot10^{16}$	16.8	0.013	0.4	0.02	77 000	850

Data on whether a compound has a direct (d) or indirect (i) gap, energy gap, static dielectric constant, effective masses, and mobilities, for some semiconductors. The electron effective mass $m_n^\star$ is the density of states effective mass defined in Eq. (19.23). The data refer to room temperature, and to samples with donor and acceptor impurities at densities of 10^{15} cm^{-3} or less. Source: Landolt and Börnstein (New Series) vol. 17 and Pierret (1996).

Despite these experimental complications the experimental determination of band gaps can be made rather precisely. Not only the energy gap, but also the structure of the energy bands in the neighborhood of valence band maxima and conduction band minima, is important. One can fit the energy to a quadratic form and write

$$\mathcal{E}_{\vec{k}} = \mathcal{E}_c + \frac{\hbar^2}{2}\vec{k}^*\mathbf{M}^{-1}\vec{k} \qquad \text{For electrons in the conduction band.} \tag{19.13a}$$

$$\mathcal{E}_{\vec{k}} = \mathcal{E}_v - \frac{\hbar^2}{2}\vec{k}^*\mathbf{M}^{-1}\vec{k}, \qquad \text{For holes in the valence band.} \tag{19.13b}$$

where **M** is the effective mass tensor. For silicon, germanium, and gallium arsenide, the bands at the valence maximum would be threefold degenerate in the absence of spin. The spin–orbit interaction splits off one of the bands, leaving two above it that still are degenerate at Γ. The two bands have, however, different curvatures near Γ, leading to *heavy holes* (low curvature) and *light holes* (high curvature), both of which contribute to the transport properties of semiconductors. Because of the great degree of symmetry associated with Γ, the energy surfaces of these holes are spherically symmetrical, and the effective mass tensors are multiples of the unit matrix.

The conduction band minimum in gallium arsenide is nondegenerate and spherical. In silicon and germanium, the conduction band minima are quite anisotropic, and consist in a number of symmetrically arrayed pockets of electrons, as shown in Figure 19.9. The effective mass tensors have been measured by the technique of cyclotron resonance, to be discussed in Section 21.2.

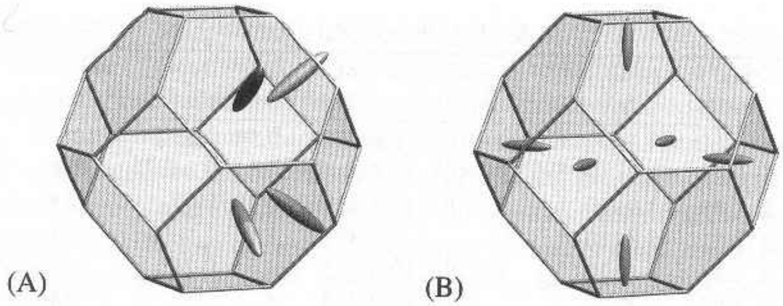

Figure 19.9. (A) The conduction band minima in germanium lie along (111) and straddle the zone boundary, producing four inequivalent pockets of electrons with a highly anisotropic effective mass. (B) In silicon, the conduction band minima lie 8/10 of the way toward (100), producing six pockets of electrons, but only three with distinct symmetries.

19.3.2 Semiconductor in Equilibrium

Electron and Hole Densities. In equilibrium, the numbers of mobile charge carriers in a semiconductor are given by the Fermi function. The volume density of electrons n above the conduction band edge is given by

$$n = \int_{\mathcal{E}_c}^{\infty} d\mathcal{E}\; D(\mathcal{E}) \frac{1}{e^{\beta(\mathcal{E}-\mu)}+1}, \tag{19.14}$$

while the density of holes p below the valence band edge is

$$p = \int_{-\infty}^{\mathcal{E}_v} d\mathcal{E}\; D(\mathcal{E}) \left\{1 - \frac{1}{e^{\beta(\mathcal{E}-\mu)}+1}\right\} \tag{19.15a}$$

$$= \int_{-\infty}^{\mathcal{E}_v} d\mathcal{E}\; D(\mathcal{E}) \frac{1}{e^{-\beta(\mathcal{E}-\mu)}+1}. \tag{19.15b}$$

Nondegenerate Semiconductors. These expressions simplify for a *nondegenerate semiconductor*, which is one for which the probability of occupying states near the band edge is exponentially small: that is,

$$\mathcal{E}_c - \mu \gg k_B T \quad \text{and} \quad \mu - \mathcal{E}_v \gg k_B T. \;\; \text{A practical criterion is } \mathcal{E}_c - \mu > 3k_B T. \tag{19.16}$$

When these conditions hold, the semiconductor is quite different from most metals. Whereas in metals carrier concentrations are on the order of 10^{22} electrons per cubic centimeter, for nondegenerate semiconductors carrier concentrations are on the order of 10^{19} cm^{-3} or less. Whether a semiconductor lies in the nondegenerate limit or not will depend upon the density of dopants (impurities) added to it. In semiconductor devices, dopant densities are frequently great enough to cause violation of inequalities (19.16). Nevertheless, the nondegenerate limit is of great utility because the transport properties of semiconductor devices are largely determined by the regions with light doping, while the regions with heavy doping act like short circuits and can often be ignored.

Given conditions (19.16), the Fermi functions (19.14) and (19.15) can be replaced by Boltzmann factors, and the equations for electron and hole concentration in the nondegenerate case become

$$n = \mathcal{N}_c e^{-\beta(\mathcal{E}_c - \mu)}, \quad p = \mathcal{N}_v e^{-\beta(\mu - \mathcal{E}_v)} \tag{19.17}$$

with

$$\mathcal{N}_c = \int_{\mathcal{E}_c}^{\infty} d\mathcal{E}\, D(\mathcal{E}) e^{-\beta(\mathcal{E} - \mathcal{E}_c)} \tag{19.18a}$$

$$\mathcal{N}_v = \int_{-\infty}^{\mathcal{E}_v} d\mathcal{E}\, D(\mathcal{E}) e^{-\beta(\mathcal{E}_v - \mathcal{E})}. \tag{19.18b}$$

Effective Masses. With reasonable approximations, one can calculate $\mathcal{N}_c$ and $\mathcal{N}_v$. It is not sufficient to take the density of states $D(\mathcal{E})$ just to be a constant. In Eq. (19.18) the exponential factor places heavy emphasis on states just at the edges of the bands where the density of states vanishes, so there is an interplay between the two terms in the integrand. Still, only states within a narrow strip near the valence maximum or conduction minimum are important, and one can use the quadratic approximations (19.13) to evaluate the density of states. For the conduction band, one has

$$D(\mathcal{E}) = \int [d\vec{k}]\, \delta\Big(\mathcal{E} - \mathcal{E}_c - \frac{1}{2}\hbar^2 \vec{k}^* \mathbf{M}^{-1} \vec{k}\Big) \tag{19.19}$$

For $[d\vec{k}]$, see Eq. (6.15).

$$= \int [d\vec{k}]\, \delta\Big(\mathcal{E} - \mathcal{E}_c - \frac{1}{2}\hbar^2 \sum_l k_l^2/m_l\Big). \tag{19.20}$$

Changing to a $\vec{k}$ basis in which $\mathbf{M}$ is diagonal, with elements m_l.

Defining

$$m_n^\star = [m_1 m_2 m_3]^{1/3} \quad \text{and} \quad \vec{q} = (k_1/m_1,\, k_2/m_2,\, k_3/m_3) \tag{19.21}$$

gives

$$D(\mathcal{E}) = 2 \int m_n^{\star 3/2} \frac{d\vec{q}}{(2\pi)^3} \delta\Big(\mathcal{E} - \mathcal{E}_c - \frac{1}{2}\hbar^2 q^2\Big) \tag{19.22}$$

$$= \sqrt{2(\mathcal{E} - \mathcal{E}_c)}\, \frac{m_n^{\star 3/2}}{\hbar^3 \pi^2} \mathcal{M}_c. \tag{19.23}$$

Where $\mathcal{M}_c$ is the number of symmetrically equivalent minima in the conduction band, equaling six for silicon and eight for germanium.

Because $m_n^\star$ is defined so as to bring the density of states $D(\mathcal{E})$ into a simple form, it is called the *density of states effective mass;* experimental values for several semiconductors appear in Table 19.1. In the case of holes, one can repeat the steps leading to Eq. (19.23) for heavy and light holes separately and define $m_p^{\star 3/2}$ to be the sum $(m_{pl}^\star)^{3/2} + (m_{ph}^\star)^{3/2}$ of the light and heavy effective hole masses. Then

$$\mathcal{N}_c = \frac{1}{4} \left(\frac{2 m_n^\star k_B T}{\pi \hbar^2}\right)^{3/2} \mathcal{M}_c \tag{19.24a}$$

$$\mathcal{N}_v = \frac{1}{4} \left(\frac{2 m_p^\star k_B T}{\pi \hbar^2}\right)^{3/2}. \tag{19.24b}$$

In order to find equilibrium densities of electrons and holes from Eq. (19.17), one needs to determine the chemical potential μ. However, there is a convenient relation independent of it, the *law of mass action*, obtained by multiplying together the expressions for n and p, to find

$$np = \mathcal{N}_c\mathcal{N}_v e^{-\beta\mathcal{E}_g}. \text{ The energy gap } \mathcal{E}_g = \mathcal{E}_c - \mathcal{E}_v. \tag{19.25}$$

19.3.3 Intrinsic Semiconductor

An *intrinsic semiconductor* is a pure single crystal. For every electron excited into the conduction band, a hole must be left behind in the valence band, so the intrinsic electron density n_i is

$$n_i = \sqrt{\mathcal{N}_c\mathcal{N}_v}\, e^{-\beta\mathcal{E}_g/2} \text{ From Eq. (19.25), setting } n = p. \tag{19.26a}$$

$$= 2.510 \cdot 10^{19}\text{cm}^{-3} \left(\frac{m_n^\star m_p^\star}{m^2}\right)^{3/4} \mathcal{M}_c^{1/2} \left(\frac{T}{300\text{ K}}\right)^{3/2} e^{-\beta\mathcal{E}_g/2}. \tag{19.26b}$$

Solving Eq. (19.17) for the chemical potential gives immediately the intrinsic chemical potential μ_i

$$\mu_i = k_BT \ln \frac{n_i}{\mathcal{N}_c} + \mathcal{E}_c = \mathcal{E}_v + \frac{\mathcal{E}_g}{2} + \frac{3}{4}k_BT \ln(m_p^\star/m_n^\star). \tag{19.27}$$

Table 19.2. Binding energies of common donors and acceptors in some semiconductors at room temperature

		Group V donors, $\mathcal{E}_c - \mathcal{E}_d$ (meV)				
Host	Eq. (18.30)	N	P	As	Sb	Bi
Si	113	140	45	53.7	42.7	70.6
Ge	28		12.9	14.2	10.3	12.8

		Group III acceptors, $\mathcal{E}_a - \mathcal{E}_v$ (meV)				
Host	Eq. (18.30)	B	In	Ga	Al	Tl
Si	48	45	155	74	67	25
Ge	15	9.73	12.0	11.3	10.8	13.5

		Donors, $\mathcal{E}_c - \mathcal{E}_d$ (meV)					
Host	Eq. (18.30)	Pb	Se	Si	S	Ge	C
GaAs	5.8	5.8	5.8	5.8	5.9	5.9	5.9

		Acceptors, $\mathcal{E}_a - \mathcal{E}_v$ (meV)								
Host	Eq. (18.30)	Be	Mg	Zn	Cd	C	Si	Ge	Sn	Mn
GaAs	23	28	29	31	35	27	35	40	167	113
InP	21	31	31	46	57	41		210		270

Apart from the case of donors in GaAs, the simple theory of Eq. (18.30) gives no more than the order of magnitude of the binding energy. Improvements on the theory, more properly incorporating anisotropy of the effective mass, and corrections due to the strong potential in the central cell near the impurity are discussed by Yu and Cardona (1996), Chapter 4. Source: Landolt and Börnstein (New Series), vol. 17.

The logarithm in Eq. (19.27) is of order unity or even zero if holes and electrons have the same effective mass, so because $k_BT \sim 1/40$ eV at room temperature and band gaps are around 1 eV for semiconductors, the chemical potential sits smack in the middle of the band gap. Thus it cooperates in enforcing (19.17), making the semiconductor as nondegenerate as possible by staying away from the band edges.

Compact Expressions. Combining Eqs. (19.26) and (19.25) puts the law of mass action in the general compact form

$$np = n_i^2 \tag{19.28}$$

and allows rewriting Eqs. (19.17) for electron and hole densities as

$$n = n_i e^{-\beta(\mu_i - \mu)}, \quad p = n_i e^{-\beta(\mu - \mu_i)}. \tag{19.29}$$

19.3.4 Extrinsic Semiconductor

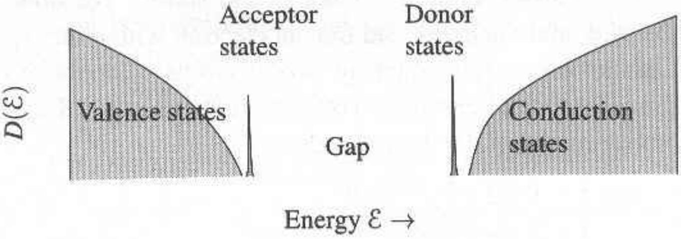

Figure 19.10. The effect of adding donors to a semiconductor is to create a population of bound states sitting just below the conduction band, while adding acceptors creates bound states just above the valence band. At room temperature, almost all the bound states break apart; each donor gives an electron to the conduction band, while each acceptor gives a hole to the valence band.

When certain sorts of impurities are used to dope a semiconductor, the physics changes rather dramatically. The most interesting impurities are noncompensated, lying one column to the right or one column to the left of a semiconductor in the periodic table. Examples of common dopants appear in Table 19.2. As discussed in Section 18.3, addition of these impurities creates a population of bound states. Atoms from column V added to a semiconductor of column IV create states just below the conduction band edge called donors, while atoms from column III added to a semiconductor of column IV create states just above the valence band edge called acceptors (Figure 19.10). However, at room temperature, these bound states are not occupied; they are almost completely ionized. Thus the practical effect of adding a donor is to add a single mobile electron, while the practical effect of adding an acceptor is to add a single mobile hole.

The energies in Table 19.2 do not make it obvious that impurity states should be ionized completely. The thermal energy at room temperature is around 25 meV, which is at best comparable to the binding energies. It is entropy more than energy that ionizes the impurities. For the purposes of a rough estimate, let the density of impurity sites per volume be $\mathcal{N}_d$, and denote by $\mathcal{N}_c$ the density of conduction states per volume into which an electron could choose to move. In a system of volume $\mathcal{V}$, the number of different ways to arrange electrons originally bound on the impurities among the conduction states is roughly $(\mathcal{N}_c/\mathcal{N}_d)^{\mathcal{V}\mathcal{N}_d}$, leading to entropy $k_B\mathcal{V}\mathcal{N}_d\ \ln \mathcal{N}_c/\mathcal{N}_d$. Therefore the temperature at which ionization occurs is not $k_BT \approx \mathcal{E}_b$, but $k_BT\ \ln \mathcal{N}_c/\mathcal{N}_d \approx \mathcal{E}_b$. The fewer impurities there are, the more mobile their electrons become. In practice, for doping levels of $\mathcal{N}_d \leq 10^{18}$ cm^{-3}, ionization is probably complete, but if doping rises higher the approximation must be checked, because $\mathcal{N}_c \approx 10^{22}$ cm^{-3}.

Verifying these claims requires a simple statistical calculation. Consider a crystal with a valence band, a conduction band lying at energy $\mathcal{E}_g = \mathcal{E}_c - \mathcal{E}_v$ higher, and donor states with maximum binding energy $\mathcal{E}_d$ just below the bottom of the conduction band. Because the impurity potential is weak, the probability of an electron being trapped in anything but the "ground state" of the effective hydrogen atom problem is negligible. In addition to occupying the conduction and valence bands, electrons can also occupy the donor bound states. The donor occupation number can be zero, and the donor can trap an electron with either spin up or spin down, but it cannot bind simultaneously two electrons of opposite spin. Therefore, in the grand canonical ensemble conventionally used for the Fermi gas, the occupation probability f_d of the donor levels is

$$f_d = \frac{0 \times 1 + 1 \times 2 \times e^{-\beta(\mathcal{E}_d - \mu)}}{1 + 2 \times e^{-\beta(\mathcal{E}_d - \mu)}} \tag{19.30}$$

$$= \frac{1}{1 + \frac{1}{2}e^{\beta(\mathcal{E}_d - \mu)}} \ll 1. \tag{19.31}$$

Equation (19.38) will show that μ lies typi-. cally in the middle of the gap, so that at room temperature $\mathcal{E}_d - \mu$ is much larger than k_BT, and f_d is nearly zero.

Similarly, if acceptor impurities are placed at an energy $\mathcal{E}_a$ above the valence band, the probability that a hole, spin up or spin down, will be localized on them is

$$f_a = \frac{1}{\frac{1}{4}e^{\beta(\mu - \mathcal{E}_a)} + 1} \ll 1. \tag{19.32}$$

The factor of $1/4$ appears in the denominator because the valence maximum is fourfold degenerate, including spin degeneracy. Again, Eq. (19.38) shows that typically this occupation number is much less than 1.

The way that entropy ionizes impurities is hidden in the value of the chemical potential, and the chemical potential is determined simply by the total number of mobile electrons. Suppose that a density of $\mathcal{N}_d$ donors per volume is added to the semiconducting crystal, which otherwise contains n electrons per volume in the valence and conduction bands. The total density of electrons is then

$$n + \mathcal{N}_d = \int_{\mathcal{E}_c} d\mathcal{E}\ D(\mathcal{E})\frac{1}{1 + e^{\beta(\mathcal{E} - \mu)}} + \int^{\mathcal{E}_v} d\mathcal{E}\ D(\mathcal{E})\frac{1}{1 + e^{\beta(\mathcal{E} - \mu)}} + \mathcal{N}_d f_d. \tag{19.33}$$

Because the integral of $D(\mathcal{E})$ over the valence band gives n, and assuming f_d is negligible,

$$\mathcal{N}_d = \int_{\mathcal{E}_c} d\mathcal{E}\ D(\mathcal{E})\frac{1}{1 + e^{\beta(\mathcal{E} - \mu)}} - \int^{\mathcal{E}_v} d\mathcal{E}\ D(\mathcal{E})\frac{1}{1 + e^{-\beta(\mathcal{E} - \mu)}} \tag{19.34}$$

$$\Rightarrow \mathcal{N}_d = n - p = n_i e^{-\beta(\mu_i - \mu)} - n_i e^{-\beta(\mu - \mu_i)}. \tag{19.35}$$

When both donors and $\mathcal{N}_a$ acceptors per volume are present, then similarly

$$n - p = \mathcal{N}_d - \mathcal{N}_a. \tag{19.36}$$

Using the law of mass action Eq. (19.28), one now easily solves for n and p, and finds

$$n = \frac{1}{2}[\mathcal{N}_d - \mathcal{N}_a] + \frac{1}{2}\left[(\mathcal{N}_d - \mathcal{N}_a)^2 + 4n_i^2\right]^{1/2} \tag{19.37a}$$

$$p = \frac{1}{2}[\mathcal{N}_a - \mathcal{N}_d] + \frac{1}{2}\left[(\mathcal{N}_d - \mathcal{N}_a)^2 + 4n_i^2\right]^{1/2}. \tag{19.37b}$$

To check that everything is consistent, one needs to make sure that the chemical potential is in fact in the middle of the gap, making f_d and f_a small. From Eq. (19.29)

$$n - p = 2n_i \sinh \beta(\mu - \mu_i) \Rightarrow \mu = \mu_i + k_B T \sinh^{-1}\left([\mathcal{N}_d - \mathcal{N}_a]/2n_i\right). \tag{19.38}$$

Thus dopants must exceed by many orders of magnitude the intrinsic carrier density before the chemical potential departs far enough from the center of the gap to endanger the conditions for nondegeneracy in (19.16).

Equation (19.37) simplifies when $\mathcal{N}_d \gg \mathcal{N}_a$, and it becomes

$$n \approx \mathcal{N}_d \qquad \text{The number of mobile electrons is essentially the number of donors.} \tag{19.39a}$$

$$p \approx \frac{n_i^2}{\mathcal{N}_d}. \qquad \text{Holes are the minority carrier.} \tag{19.39b}$$

There is a similar result when the number of acceptors exceeds the number of donors; in this case,

$$p \approx \mathcal{N}_a \qquad \text{The number of mobile holes is essentially the number of acceptors.} \tag{19.40a}$$

$$n \approx \frac{n_i^2}{\mathcal{N}_a}. \qquad \text{Electrons are the minority carrier.} \tag{19.40b}$$

19.4 Diodes and Transistors

The first semiconductor device was the *point-contact rectifier* or *Schottky diode*, in which a metal whisker was placed against a semiconducting crystal. The contact of metal with semiconductor remains an important element in electronic design, and it is worth understanding the conditions under which this junction rectifies current.

Ideal Schottky Diode. Suppose that an ideal contact between semiconductor and metal is possible, in which the atoms of the metal join seamlessly with those of the semiconductor. Such a joint is actually extremely difficult to create in practice for numerous reasons. Immediately after cleaving, semiconductor surfaces acquire oxide layers; any sort of mechanical polishing produces surfaces that are far from atomically flat; and even conventional molecular beam epitaxy often fails to lay metal smoothly down upon a semiconductor surface, producing instead blobs and islands. Nevertheless, suppose that a smooth contact has been achieved. After examining this ideal case, the consequences of a defective interface will also be mentioned.

Figure 19.11 shows the equilibrium behavior of an n-doped semiconductor brought into contact with a metal. The work function of the semiconductor is taken to be less than the work function of the metal; if the reverse is the case, the junction may have little or no rectifying power, and the contact is called *ohmic*, as in Problem 2. Because of the higher chemical potential, electrons rush from the semiconductor to the metal, lowering the voltage of the metal until electrostatic forces prevent further motion of charge. The resulting potential profile is depicted in the lower parts of Figure 19.11. The representation of the junction in Figure 19.11 explains why the electrostatic potential is said to cause *band bending*.

When an external voltage V_A is applied to raise the metal relative to the semiconductor, the situation changes qualitatively as in Figure 19.12(A). The barrier for

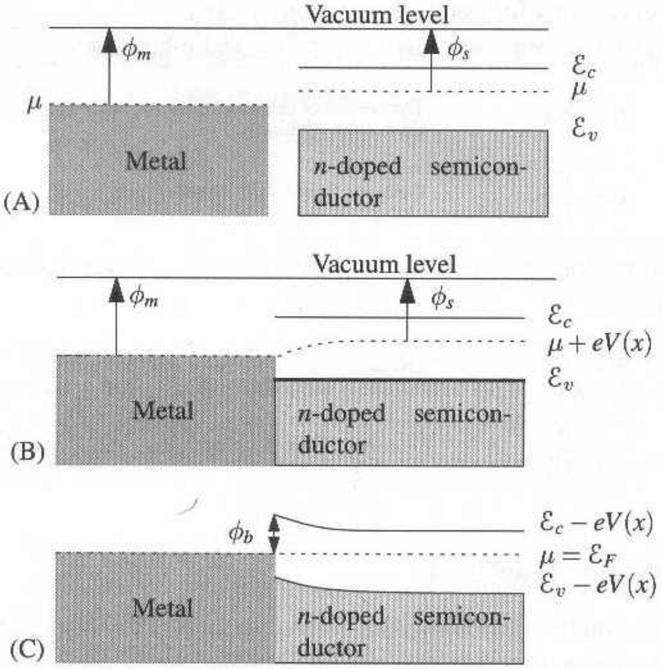

Figure 19.11. (A) In the first instant that a metal and semiconductor are brought together, their chemical potentials do not coincide. (B) Very quickly, however, charge moves from the solid with higher chemical potential to the one with lower chemical potential—in this case, from the n-doped semiconductor to the metal—until the rise in voltage of the semiconductor compensates for the difference in chemical potential. (C) The customary representation of the potentials experienced by the electrons and holes shows the chemical potential μ as constant, and it adds the electrostatic potential $-eV$ to the conduction and valence band levels. The bands have been bent by the potentials which form across the junction. The chemical potential is often referred to as the Fermi energy $\mathcal{E}_F$.

electrons is lowered, and the electrons flow into the metal from the semiconductor as a current. However, if the voltage changes in the opposite fashion, as in Figure 19.12(B), the barrier seen by electrons in the metal does not change, and therefore current does not increase in the opposite direction.

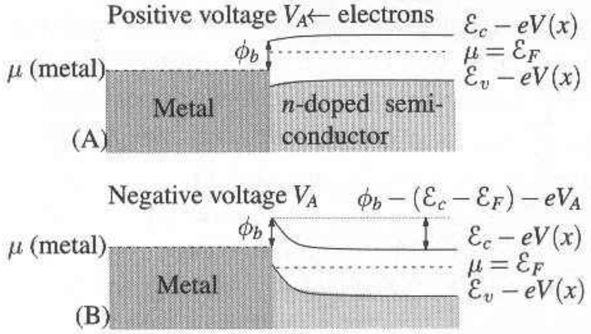

Figure 19.12. (A) When the voltage of the metal is raised relative to the semiconductor by eV_A, electrons flow to the metal. (B) However, when the voltage of the metal is lowered, the barrier perceived by the electrons does not change, and little current flows.

Quantitative Theory. The quantitative theory for rectification in the Schottky diode is almost identical to the theory of thermionic emission from metals. The electrons in the semiconductor with enough energy to travel to the metal are those in the conduction band whose velocity toward the metal is large enough that they can cross the barrier between semiconductor and metal. According to Figures 19.11(C), and 19.12, the height of this barrier in the presence of an applied voltage V_A is $\phi_b - (\mathcal{E}_c - \mu) - V_A$, so the condition is

$$\frac{\hbar^2 k_x^2}{2m_n^\star} > \phi_b - (\mathcal{E}_c - \mu) - eV_A. \tag{19.41}$$

Using the density of states effective mass in an approximation for the (anisotropic) kinetic energy.

The current density $j_{s\to m}$ due to this collection of electrons is

$$j_{s\to m} = \int [d\vec{k}]\, \theta\Big(\frac{\hbar^2 k_x^2}{2m_n^\star} - [\phi_b - (\mathcal{E}_c - \mu) - eV_A\Big)\, \frac{e\hbar k_x}{m_n^\star}\, e^{-\beta(\hbar^2 k^2/2m_n^\star + \mathcal{E}_c - \mu)} \tag{19.42}$$

Using the nondegenerate limit of the Fermi function and assuming the electrons to travel in the $-x$ direction, with the minus canceling the sign of the charge. $[d\vec{k}]$ defined in Eq. (6.15).

$$= \frac{2}{(2\pi)^3} \frac{2m_n^\star \pi k_B T}{\hbar^2} \frac{1}{\hbar} \int_{\phi_b - \mathcal{E}_c + \mu - eV_A}^{\infty} d\left(\frac{\hbar^2 k_x^2}{2m_n^\star}\right) e^{-\beta(\hbar^2 k_x^2/2m_n^\star + \mathcal{E}_c - \mu)} \tag{19.43}$$

Doing the integrals over k_x and k_y.

$$= \frac{m_n^\star}{m} \mathcal{A} T^2 \exp\left\{-\beta[\phi_b - eV_A]\right\}. \tag{19.44}$$

$\mathcal{A}$=120 A K^{-2} cm^{-2} was given in Eq. (19.9).

When $V_A = 0$, the reverse current $j_{m\to s}$ flowing from metal to semiconductor must equal the one calculated in Eq. (19.44), and because the barrier seen from the metal does not change with V_A, the current flowing in this reverse direction will be independent of applied voltage. So the total current in the junction is

$$j = \frac{m_n^\star}{m} \mathcal{A} T^2 \left[\exp\left\{-\beta[\phi_b - eV_A]\right\} - 1\right]. \tag{19.45}$$

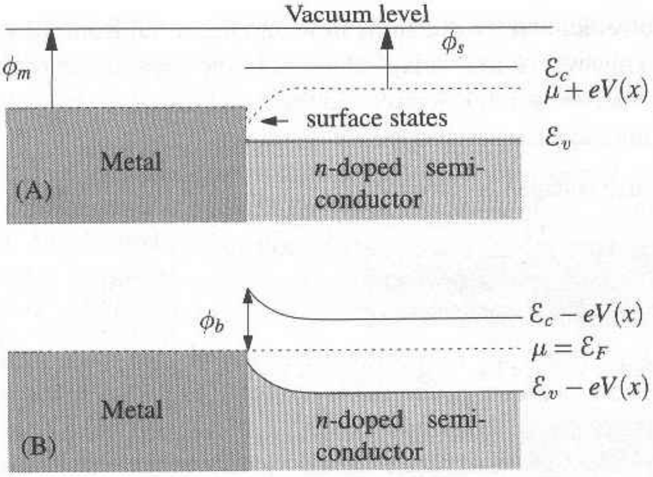

Figure 19.13. Effect of surface states on metal–semiconductor junction. (A) The electrochemical potential $\mu + eV$ at the surface of the semiconductor is fixed at a particular location in the gap independent of doping. (B) The bands are bent so that the barrier between metal and semiconductor is a constant ϕ_b that depends mainly upon properties of the semiconductor surface, and only slightly upon either metal or semiconductor work functions.

19.4.1 Surface States

Experimental measurements confirm Eq. (19.45), but with one troubling discrepancy. The constant ϕ_b does not equal $\phi_m - \phi_s - (\mathcal{E}_c - \mu)$ as it should according to Figure 19.11, and it is almost independent of the metal or doping level of the semiconductor involved in the contact. For example, n-type GaAs almost always appears to have $\phi_b \approx 2\mathcal{E}_g/3 = 0.95$eV, while p-type GaAs almost always appears to have $\phi_b \approx \mathcal{E}_g/3 = 0.47$eV. The explanation, proposed by Bardeen (1947), is that the surface of the semiconductor joins the metal in a rough fashion. At the interface there is a high density (10^{15} cm^{-2}) of *dangling bonds*—that is, atoms eagerly expecting to join onto neighbors to form a perfect diamond lattice, but frustrated by the presence of the surface. It is energetically favorable to steal charge from the nearby bulk and to place electrons on the dangling bonds, leaving a positively charged region several hundred angstroms thick below the semiconductor surface. In addition, there is a large density of propagating surface states with energies lying right within the gap and localized states due to defects. A schematic representation of the consequences for energy bands appears in Figure 19.13. When the metal and semiconductor come into contact, the chemical potentials equilibrate as charge moves from the metal into the surface states, creating a dipole layer at the interface. Because the charge density needed to create this layer is often small compared to the density of dangling bonds, the space charge distribution within the semiconductor is not much altered by the approach of the metal.

Semiconductor electronics avoids the problem of surface states by building junctions out of single crystals. Instead of preparing two separately doped crystals, polishing the surfaces, and gluing them together, the dopants are injected into a single sample at desired locations, either as an ion beam or by diffusion.

19.4.2 Semiconductor Junctions

Junction in Equilibrium. Consider a junction between an n-doped and a p-doped region. Electrons move from the n-type region to the p-type region until charge buildup cancels out the advantage of populating lower energy levels. Figures 19.14 and 19.15 help in visualizing why.

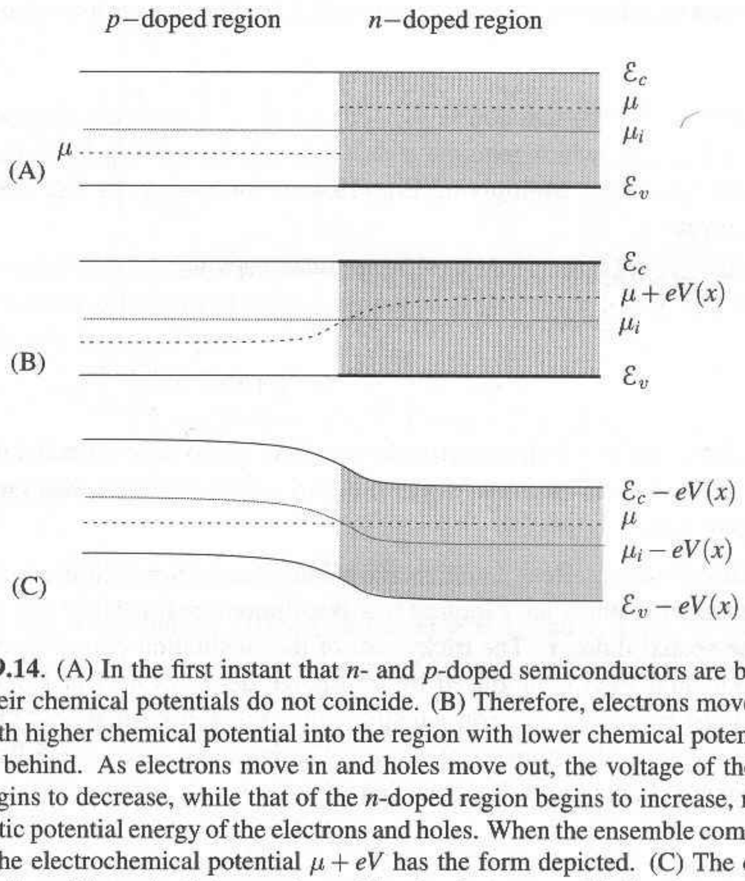

Figure 19.14. (A) In the first instant that n- and p-doped semiconductors are brought together, their chemical potentials do not coincide. (B) Therefore, electrons move from the region with higher chemical potential into the region with lower chemical potentials leaving holes behind. As electrons move in and holes move out, the voltage of the p-doped region begins to decrease, while that of the n-doped region begins to increase, raising the electrostatic potential energy of the electrons and holes. When the ensemble comes to equilibrium, the electrochemical potential $\mu+eV$ has the form depicted. (C) The customary representation of the potentials experienced by the electrons and holes shows the chemical potential μ as constant, and it adds the electrostatic potential $-eV$ to the conduction and valence band levels. The bands have been bent by the potentials that form across the junction.

To obtain a quantitative theory, observe that in the presence of an electrical potential $V(x)$, the densities of electrons n and holes p in a nondegenerate semiconductor are given by

$$n(x)=n_ie^{\beta(\mu+eV(x)-\mu_i)} \tag{19.46a}$$

Generalize Eqs. (19.29) to include spatial variations; valid if spatial gradients are small enough that the semiconductor is locally in equilibrium.

$$p(x)=n_ie^{\beta(\mu_i-eV(x)-\mu)}. \tag{19.46b}$$

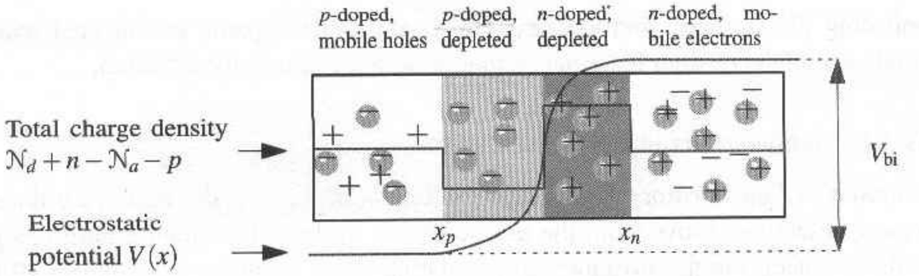
p-doped, mobile holes
p-doped, depleted
n-doped, depleted
n-doped, mobile electrons
Total charge density $\mathcal{N}_d + n - \mathcal{N}_a - p$
Electrostatic potential $V(x)$
x_p
x_n
V_{bi}

Figure 19.15. Illustration of the redistribution of mobile charges near a *p*–*n* junction. The mobile carriers abandon the region between x_n and x_p, leaving nonzero ionic charge density behind.

Far to the left of the junction ($x \to -\infty$), on the *p*-doped side, the total charge density must vanish, which requires $p = \mathcal{N}_a$. Similarly, $n = \mathcal{N}_d$ far to the right of the junction ($x \to \infty$). Multiplying Eq. (19.46a) for $x \to \infty$ by Eq. (19.46b) for $x \to -\infty$ gives

$$n(\infty)p(-\infty) = \mathcal{N}_d\mathcal{N}_a = n_i^2 e^{\beta(eV(\infty)-eV(-\infty))} \tag{19.47}$$

$$\Rightarrow eV_{\mathrm{bi}} \equiv e[V(\infty) - V(-\infty)] \tag{19.48}$$

$$= k_BT \ln \frac{\mathcal{N}_d\mathcal{N}_a}{n_i^2} = \mathcal{E}_g + k_BT \ln[\frac{\mathcal{N}_d\mathcal{N}_a}{\mathcal{N}_c\mathcal{N}_v}], \tag{19.49}$$

where V_{bi} is the *built-in voltage* across the junction, an intrinsic potential difference due to the fact that the electrons of the *n*-doped region combine with the holes of the *p*-doped region.

Charge Distribution. Real junctions have complicated three-dimensional forms, but the essential features are captured in a one-dimensional calculation, as a function of the spatial index x. The tricky part of the calculation comes from the fact that the potential $V(x)$ is produced by the charge densities $n(x)$ and $p(x)$, so the problem must be solved self-consistently, using Poisson's equation. The charge density is the sum of a number of terms. The impurity states are fully ionized, leaving behind charged ions that contribute

$$en_{\mathrm{ions}} = e[\mathcal{N}_a(x) - \mathcal{N}_d(x)]. \tag{19.50}$$

In addition, one has to consider the contributions from the electrons $n(x)$ in the conduction band and the holes $p(x)$ in the valence band, so Poisson's equation reads

$$\frac{\partial^2 V}{\partial x^2} = -4\pi e[-\mathcal{N}_d(x) - n(x) + \mathcal{N}_a(x) + p(x)]/\epsilon^0, \tag{19.51}$$

with ϵ^0 the dielectric constant.

For the junction depicted in Figure 19.15 there is an abrupt transition between an *n*-type semiconductor and a *p*-type semiconductor, so

$$\mathcal{N}_d(x) = \mathcal{N}_d\theta(-x) \quad \theta(x) \text{ is a Heaviside step function.} \tag{19.52a}$$

$$\mathcal{N}_a(x) = \mathcal{N}_a\theta(x). \tag{19.52b}$$

Junctions are not actually infinitely sharp, but they can certainly be less than 100 Å, which is going to be the scale of the *depletion regions* in which charge builds up.

Equation (19.46) shows that once the potential begins to deviate from its value at infinity, the number of carriers n or p drops below $\mathcal{N}_d$ or $\mathcal{N}_a$ like a stone. It is therefore reasonable to construct an approximation in which the charge density is zero everywhere up to $x_p < 0$, at which point the charge density abruptly changes to $-e\mathcal{N}_d$. At $x = 0$ the charge density rises to $e\mathcal{N}_a$, and finally at some $x_n > 0$, it falls abruptly back to zero. The potential produced by such a charge density is

$$V(x) = \begin{cases} V(-\infty) & \text{for } x < x_p \\ V(-\infty) + 2\pi e \frac{\mathcal{N}_d}{\epsilon^0}(x - x_p)^2 & \text{for } 0 > x > x_p \\ V(\infty) \quad - 2\pi e \frac{\mathcal{N}_a}{\epsilon^0}(x - x_n)^2 & \text{for } 0 < x < x_n \\ V(\infty) & \text{for } x > x_n. \end{cases} \tag{19.53}$$

Equation (19.53) is obviously a solution of Eq. (19.51), and the only thing left to check is that the solution and its first derivative are continuous at $x = 0$. Continuity of (19.53) at 0 demands that

$$V(-\infty) + 2\pi e \frac{\mathcal{N}_a}{\epsilon^0} x_p^2 = V(\infty) - 2\pi e \frac{\mathcal{N}_d}{\epsilon^0} x_n^2, \tag{19.54}$$

while continuity of the derivative requires that

$$\mathcal{N}_d x_n = -\mathcal{N}_a x_p. \tag{19.55}$$

Solving Eq. (19.54) and Eq. (19.55) for the lengths x_n and x_p gives

$$x_n = \sqrt{\frac{\epsilon^0 \mathcal{N}_a V_{\text{bi}}}{2\pi e \mathcal{N}_d[\mathcal{N}_a + \mathcal{N}_d]}} \tag{19.56a}$$

$$x_p = -\sqrt{\frac{\epsilon^0 \mathcal{N}_d V_{\text{bi}}}{2\pi e \mathcal{N}_a,\ [\mathcal{N}_a + \mathcal{N}_d]}}, \tag{19.56b}$$

using again the built-in voltage V_{bi} defined in Eq. (19.48). Placing typical numerical values into Eq. (19.56), dopant densities on the order of 10^{18} cm^{-3}, and potential differences eV_{bi} on the order of 0.1 eV gives depletion layers on the order of a few hundred angstroms. Because the depletion region has no mobile change, its resistance is considerably greater than that of the doped regions to either side.

When an external voltage V_A is applied to such a junction, the net effect depends greatly upon the direction in which it happens. If the potential of the left-hand side is raised relative to the right-hand side, electrons are attracted to the left, and holes are attracted to the right. As a consequence, x_p moves further to the left, and x_n moves further to the right. Conversely, if the potential is lowered to the left, electrons are repelled from the left, and the size of the depletion region decreases. Quantitatively, applying a voltage corresponds to a case in which the system departs slightly from equilibrium, so that the chemical potential μ is no longer constant, but instead changes by amount eV_A from one end of the sample to the other. Most of the voltage drop occurs in the depletion region. One does not need to determine the spatial profile to observe, however, that because μ is now different on the two sides of the sample, the potentials $V(\infty)$ and $V(-\infty)$ must also change accordingly so as to maintain charge neutrality, and the difference between them also changes by V_A. According to Eq. (19.56), the effect of applied voltage is to send $V_{\rm bi} \rightarrow V_{\rm bi} - V_A$ and thereby change the lengths of x_n and x_p by a factor of $\sqrt{1 - V_A/V_{\rm bi}}$.

The applied voltage V_A is taken positive if it raises the voltage of the p-doped region with respect to the n-doped region in Figure 19.15. As the size of the depletion region varies, the amount of current that flows through the junction changes dramatically, increasing exponentially as V_A increases. The reason for the exponential rise is that for an electron to flow through the depletion region, it must be a mobile carrier on the left side of Figure 19.15 with enough thermal energy to surmount the potential barrier $eV_{\rm bi}$; the number of such electrons is proportional to $\exp[-\beta eV_{\rm bi}]$ and changes in response to external voltages as $\exp[\beta eV_A]$. When the external voltage is zero, the number of electrons returning from the left must exactly equal the number jumping over the potential barrier from the right; electrons in the p-doped region are always attracted back to the n-doped region and have no barrier to cross. This electron current from left to right should not change much while external voltage rises from zero, so the total current J has the form

$$J \propto e^{\beta eV_A} - 1, \tag{19.57}$$

showing the exponential dependence upon external voltage that characterizes rectification.

19.4.3 Boltzmann Equation for Semiconductors

Once an external voltage V_A is applied across a junction and current begins to flow, equilibrium equations such as (19.46) no longer directly apply. One must return to the Boltzmann equation, Section 17.2, and solve for the distribution function $g_{\vec{r}\vec{k}}$. The most convenient form of the Boltzmann equation for semiconductors is somewhat different from the most convenient form for metals because:

1. It is valuable to write the equations in a form that emphasizes the separate roles of electrons and holes.

2. It is useful to simplify the equations by averaging over wave vectors $\vec{k}$.

Using the Hamiltonian structure (17.1), rewrite Eq. (17.13) in the relaxation time approximation as

$$\frac{\partial g}{\partial t} = -\frac{\partial}{\partial \vec{r}} \cdot \dot{\vec{r}} g - \frac{\partial}{\partial \vec{k}} \cdot \dot{\vec{k}} g + \frac{f-g}{\tau}. \tag{19.58}$$

The density n of electrons at position $\vec{r}$ is defined by

$$n = \int [d\vec{k}]\; g_{\vec{r}\vec{k}}, \tag{19.59}$$

The integral is over the first Brillouin zone, and $[d\vec{k}]$ is defined in Eq. (6.15).

with g giving the occupation probability of states in the conduction band. Integrating $d\vec{k}$ over both sides of Eq. (19.58) gives

$$\frac{\partial n}{\partial t} = -\frac{\partial}{\partial \vec{r}} \cdot \langle \dot{\vec{r}} \rangle n + \frac{n^{(0)} - n}{\tau_n}, \tag{19.60}$$

The relaxation time τ should be independent of $\vec{k}$ to pass through the averaging process; otherwise, use a constant τ_n that gives the best approximation to the averaged collision term.

where $n^{(0)}$ is the equilibrium density of electrons in the conduction band, and $\langle \dot{\vec{r}} \rangle$ is the velocity $\vec{v}_{\vec{k}}$ averaged over the Brillouin zone,

$$\langle \dot{\vec{r}} \rangle = \frac{1}{n} \int [d\vec{k}]\; g_{\vec{r}\vec{k}} \vec{v}_{\vec{k}} \tag{19.61}$$

$$= \frac{1}{n} \int [d\vec{k}] \left[f - \tau \vec{v}_{\vec{k}} \cdot \left\{ e\vec{E} \frac{\partial f}{\partial \mu} + \frac{\partial f}{\partial \vec{r}} \right\} \right] \vec{v}_{\vec{k}} \tag{19.62}$$

Using Eq. (17.25), employing Eq. (17.23) to simplify some of the terms.

$$\approx \frac{1}{n} \int [d\vec{k}] \left[-\tau \vec{v}_{\vec{k}} \cdot \left\{ e\vec{E}\beta g + \frac{\partial g}{\partial \vec{r}} \right\} \right] \vec{v}_{\vec{k}} \tag{19.63}$$

$$= -\mu_n \vec{E} - \frac{\mathcal{D}_n}{n} \frac{\partial n}{\partial \vec{r}} \tag{19.64}$$

The first term vanishes by symmetry. $\partial f / \partial \mu = \beta f$ in the nondegenerate limit. Finally, replace f by g, because the two differ only by small quantities.

with the *mobility* μ_n,

$$\mu_n = \frac{e}{3} \beta \left\langle \tau v_{\vec{k}}^2 \right\rangle \tag{19.65}$$

Assuming that the conductivity tensor of Eq. (17.51) is diagonal. Otherwise, mobility and diffusion are tensors.

and the *diffusion constant* $\mathcal{D}_n$

$$\mathcal{D}_n = \frac{1}{3} \left\langle \tau v_{\vec{k}}^2 \right\rangle = \frac{k_B T \mu_n}{e}. \tag{19.66}$$

The factor of $1/3$ appears because only the component of $\vec{v}$ along $\vec{E}$ survives the average in Eq. (19.63).

Therefore currents of electrons and holes are

$$\vec{j}_n = e\mu_n n \vec{E} + e\mathcal{D}_n \vec{\nabla} n \tag{19.67a}$$

Multiply Eq. (19.64) by $-ne$.

$$\vec{j}_p = e\mu_p p \vec{E} - e\mathcal{D}_p \vec{\nabla} p, \tag{19.67b}$$

Working in an analogous fashion.

and the equations of motion for the electron and hole distributions are

$$\frac{\partial n}{\partial t} = \frac{1}{e} \vec{\nabla} \cdot \vec{j}_n + \frac{n^{(0)} - n}{\tau_n} \tag{19.68a}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{e} \vec{\nabla} \cdot \vec{j}_p + \frac{p^{(0)} - p}{\tau_p}, \tag{19.68b}$$

with the electric field determined from

$$\vec{\nabla} \cdot \vec{E} = \frac{4\pi e(p - n)}{\epsilon^0}. \tag{19.69}$$

Recombination and Generation. Several different physical processes may be encompassed by the relaxation times τ_n and τ_p. In addition to scattering off impurities, electrons and holes can collide and recombine; they can also be generated by very energetic collision events. Therefore, the collision term in semiconductors is thought of as a *recombination and generation current.* The relaxation times τ_p and τ_n are impossible to tabulate, because they depend sensitively upon sample purity and temperature, and can vary from 10^{-9} to 10^{-14} s. They can be measured in any given sample—for example, by exposing the crystal to a flash of light that excites electrons into the conduction band and by then measuring the decay of the conductivity.

19.4.4 Detailed Theory of Rectification

Solving Eqs. (19.67)–(19.69) poses numerous difficulties. The equations are nonlinear, because they involve products of n and p with the electric field $\vec{E}$. Exact analytical solution is out of the question, even in the simplified one-dimensional situation upon which attention is now focused. Numerical solution is also not entirely straightforward because of the wide range of scales over which the various quantities vary. For example, the characteristic scale of depletion layers is from 10^{-6} to 10^{-4} cm, while the characteristic scale for variation of n and p outside the depletions layers turns out to be on the order of 10^{-2} cm. In addition, the magnitudes of the charge distributions vary over many orders of magnitude.

Ideal Diode Equation. The best approach to these difficulties is a conventional solution, the *ideal diode equation.* As in the equilibrium case, the diode is divided into three regions, indicated in Figure 19.16:

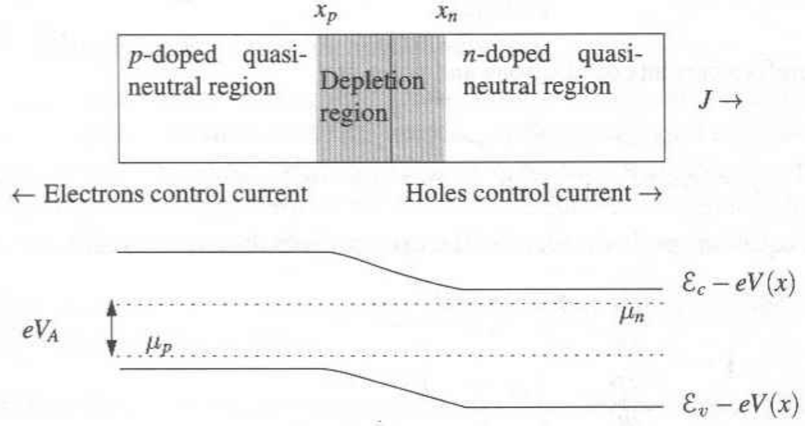

Figure 19.16. Sketch of p–n junction in forward bias, with voltage of the p side raised by voltage V_A above n side of the junction. Because the junction is out of equilibrium, the chemical potentials μ_n and μ_p in the n and p regions are not equal. The depletion region is compressed by the voltage difference, and current increases exponentially with V_A.

A quasi-neutral n-doped region where the electric field is extremely small and the density of mobile electrons is close to $\mathcal{N}_d$.

A depletion region where the electric field rises rapidly to large values and where consequently the concentrations of charge carriers rapidly fall below their values in homogeneously doped samples. In the midst of this depletion region, the doping changes from n to p type.

A quasi-neutral p-doped region where the electric field drops back to zero and the density of mobile holes is close to $\mathcal{N}_a$.

The vast majority of mobile charge carriers in the quasi-neutral n-doped region are electrons. However, there is also a small population of holes. Similarly there is a small population of electrons in the p-doped region. These two populations are called *minority carriers*, and the operation of the diode can be understood by carefully analyzing their behavior, because whichever of them is least mobile constitutes the main bottleneck restraining charge flow through the diode.

Further progress rests upon two simplifications:

1. The boundary between the depletion region and the quasi-neutral regions is sharp. In the quasi-neutral regions, the electric field is very small, and the drift currents $e\mu_n nE$ (p-region) and $e\mu_p pE$ of the minority carriers are negligible. That is, in regions where carriers are unlikely, they obey the purely linear equations

   $$j_n = e\mathcal{D}_n\vec{n}' \qquad \text{Electrons, specializing to one dimension; the prime means spatial derivative.} \qquad (19.70a)$$

   $$j_p = -e\mathcal{D}_p\vec{p}' \qquad \text{Holes.} \qquad (19.70b)$$

   This approximation is excellent.

2. Recombination and generation of charge carriers is neglected in the depletion region. This assumption is made for mathematical convenience only. The faster the charges sweep through the depletion region, the more appropriate it will be, but for low current flow it leads to appreciable deviation from experiment.

Solution in Depletion Region. The value of the second assumption lies in the fact that it makes possible an analytical relation for n and p in the depletion region, allowing the behavior of the diode to be obtained in closed form. Because recombination and generation are neglected, the currents of electrons and holes are separately conserved, and both j_n and j_p are constant in space. Using Eqs. (19.67), one can quickly find a solution for n and p which according to Problem 4 is

$$n(x) = \mathcal{N}_d e^{\beta e[V(x)-V(x_n)]}\left[1+\frac{j_n}{e\mathcal{N}_d\mathcal{D}_n}\int_{x_n}^{x} dx'\; e^{-\beta e[V(x')-V(x_n)]}\right] \quad (19.71a)$$

$$p(x) = \mathcal{N}_a e^{-\beta e[V(x)-V(x_p)]}\left[1-\frac{j_p}{e\mathcal{N}_a\mathcal{D}_p}\int_{x_p}^{x} dx'\; e^{\beta e[V(x')-V(x_p)]}\right] .(19.71b)$$

Under equilibrium conditions, where no current flows, the second terms on the right-hand side of Eqs. (19.71) vanish. Ignoring this second term is very convenient, because use of the first term requires only knowledge of the total change in potential across the depletion region, while the second would require knowledge of details of the profile of $V(x)$. Fortunately, for most cases of interest the second term remains negligible relative to the first, even out of equilibrium. Looking ahead to Eq. (19.77), one can estimate its size in the presence of applied voltage V_A to be

$$\frac{n_i^2}{\mathcal{N}_a\mathcal{N}_d}\frac{x_p-x_n}{L_N}e^{\beta eV_A}\approx 10^{-10}e^{\beta eV_A}. \tag{19.72}$$

Junction widths are typically 10^2 times smaller than diffusion lengths, and $n_i^2/\mathcal{N}_a\mathcal{N}_a \sim 10^{-8}$.

Therefore, Eq. (19.71) can be replaced by the *law of the junction*:

$$n(x)=\mathcal{N}_d e^{\beta e[V(x)-V(x_n)]} \tag{19.73a}$$

$$p(x)=\mathcal{N}_a e^{-\beta e[V(x)-V(x_p)]} \tag{19.73b}$$

$$\Rightarrow n(x_p)=\mathcal{N}_d e^{\beta e[V_A-V_{\text{bi}}]}=\frac{n_i^2}{\mathcal{N}_a}e^{\beta eV_A} \tag{19.73c}$$

See Eq. (19.49), and use approximation Eq. (19.40). Note that the density of minority carriers on the left side of the junction is being set by the density $\mathcal{N}_d$ of donors on the right side.

$$p(x_n)=\mathcal{N}_a e^{\beta e[V_A-V_{\text{bi}}]}=\frac{n_i^2}{\mathcal{N}_d}e^{\beta eV_A}. \tag{19.73d}$$

Solution in Quasi-Neutral Region. Equations (19.73) constitute a complete solution for the charge carriers in the depletion region. They cannot be used alone to find the current flowing through the diode, because Eqs. (19.73) produce a complete cancellation of diffusion and drift currents, and putting back in the tiny corrections of Eqs. (19.71) to obtain nonzero current means adding back in terms proportional to j_n and j_p which are still unknown. In this sense, Eqs. (19.73) are compatible with a huge range of currents through the diode. However, by using Eqs. (19.73) to impose a boundary condition upon the solutions of Eqs. (19.68), the currents are rapidly determined.

Using the expressions for current (19.70) in Eqs. (19.68) gives in steady state

$$0=\mathcal{D}_p\frac{d^2p}{dx^2}-\frac{p-p^{(0)}}{\tau_p} \tag{19.74a}$$

Applies only to minority carriers, and only in quasi-neutral regions. $p^{(0)}$ and $n^{(0)}$ are the equilibrium minority carrier densities, given by Eqs. (19.39) or (19.40).

$$0=\mathcal{D}_n\frac{d^2n}{dx^2}-\frac{n-n^{(0)}}{\tau_n}, \tag{19.74b}$$

which have solution

$$p-p^{(0)}=[p(x_n)-p^{(0)}]e^{-(x-x_n)/L_p} \tag{19.75a}$$

Applies where p is the minority carrier, to the right of x_n.

$$n-n^{(0)}=[n(x_p)-n^{(0)}]e^{(x-x_p)/L_n} \tag{19.75b}$$

Applies where n is the minority carrier, to the left of x_p.

where

$$L_n=\sqrt{\mathcal{D}_n\tau_n}\quad\text{and}\quad L_p=\sqrt{\mathcal{D}_p\tau_p} \tag{19.76}$$

are the *diffusion lengths* of electrons in the p-doped region, and of holes in the n-doped region, respectively. The currents due to these minority carriers are, from Eqs. (19.67),

$$\vec{j}_n = e\frac{\mathcal{D}_n}{L_n}[n(x_p) - n^{(0)}] \qquad \text{Evaluate Eq. (19.75a) at } x_p. \tag{19.77a}$$

$$= e\frac{\mathcal{D}_n}{L_n}\frac{n_i^2}{\mathcal{N}_a}[e^{\beta eV_A} - 1] \qquad \text{Use Eqs. (19.73c) and (19.40b).} \tag{19.77b}$$

$$\vec{j}_p = e\frac{\mathcal{D}_p}{L_p}[p(x_n) - p^{(0)}], \tag{19.77c}$$

$$= e\frac{\mathcal{D}_p}{L_p}\frac{n_i^2}{\mathcal{N}_d}[e^{\beta eV_A} - 1], \qquad \text{Use Eqs. (19.73d) and (19.39b).} \tag{19.77d}$$

producing a total current per volume given by the *ideal diode* or *Shockley equation*,

$$j = en_i^2[e^{\beta eV_A} - 1]\left[\frac{\mathcal{D}_n}{L_n\mathcal{N}_a} + \frac{\mathcal{D}_p}{L_d\mathcal{N}_d}\right]. \qquad \text{Doping must be heavy enough that Eqs. (19.39) and (19.40) hold.} \tag{19.78}$$

One of the most important features of Eq. (19.78) is that because $\mathcal{N}_a$ and $\mathcal{N}_d$ appear in denominators, current flow is set by the side of the diode that is most lightly doped. The heavily doped side acts like a short circuit. This fact is particularly important for the design of the transistor.

19.4.5 Transistor

By the 1920s numerous scientists realized that because electronics was based upon the diode and the triode, and because semiconductor diodes could be created (although unreliably), it would be valuable to create a semiconductor analog of the triode. Twenty-five years elapsed between the first ideas, and the first practical implementation, called the *transistor* by Bardeen and Brattain (1948). The first working transistor involved contact between thin metal whiskers and semiconductors, rather like the Schottky diodes. It was unable to carry large currents and never developed into a commercial device, but the research project in which the point-contact transistor was created uncovered much of the basic physics of semiconductor junctions, particularly the fact that transport in diodes is dominated by minority carriers. The *bipolar junction transistor* followed not longer after and served as the foundation for the first developments of semiconductor electronics.

The basic idea of the bipolar junction transistor is to take advantage of the large disparity between electron and hole currents in a diode where one side is much more heavily doped than the other. Consider, for example, a p^+n junction, where the superscript $+$ indicates heavy doping, on the order of 10^{18} cm^{-3}, so that the assumption the semiconductor is nondegenerate breaks down. For steady current flow under forward bias, a tiny electron current flows into the n region, and a large hole current flows in to the p^+ region. In a diode, the hole current would be drawn to the n region and out of the semiconductor, but in the transistor the hole current is diverted by making the n region much narrower than the diffusion length L_p of the minority carriers and placing it in contact with a second pn junction, which is under reverse bias. The reverse bias means that in the depletion region electric fields propel holes toward the p region and repel electrons. Whenever a hole diffusing about in the n region wanders into this second depletion region, it is trapped and sent off to the collector. The net effect is to split the current traveling into the emitter into its constituent components, with almost all the holes going out the collector and almost all the electrons coming in from the base. The large ratio between these two currents, along with the fact that they are linearly related according to Eqs. (19.77), means that the transistor can function as a linear amplifier. On the other hand, if the current to the base is reversed, the current out the collector does not follow it linearly but drops to very low values. Thus the transistor also rectifies current and can be used as a binary switch.

The mathematical analysis of the binary junction transistor involves no ideas or assumptions not already present in the case of the ideal diode. The only difficulty is that there is now a large number of different regions, so the notation becomes confusing. Once again the basic idea is to assume steady-state conditions and

1. Separate the device into quasi-neutral and depletion regions.

2. Ignore recombination–generation in depletion regions.

Also as before, the strategy is to focus upon the minority carriers in each region. The fields that need to be found are $n_E(x)$, the electron concentration in the *emitter*, $p_B(x)$, the hole concentration in the *base*, and $n_C(x)$, the electron concentration in the *collector*, regions labeled in Figure 19.17.

The concentrations of the minority carriers at the edges of each depletion region are determined by precisely the considerations that produced the ideal diode equation. So, in analogy with Eqs. (19.73c) and (19.73d),

$$n_E(x_a) = \frac{n_i^2}{\mathcal{N}_E} e^{\beta e V_{EB}} \tag{19.79a}$$

$\mathcal{N}_E$ is the acceptor concentration in the emitter region, $V_{EB} > 0$ (for active bias) is the voltage of emitter over base.

$$p_B(x_b) = \frac{n_i^2}{\mathcal{N}_B} e^{\beta e V_{EB}} \tag{19.79b}$$

$\mathcal{N}_B$ is the donor concentration in the base region.

$$p_B(x_c) = \frac{n_i^2}{\mathcal{N}_B} e^{\beta e V_{CB}} \tag{19.79c}$$

$V_{CB} < 0$ (for active bias) is the voltage of the collector relative to the base; when V_{CB} is negative, collector voltage is below base.

$$n_C(x_d) = \frac{n_i^2}{\mathcal{N}_C} e^{\beta e V_{CB}}. \tag{19.79d}$$

$\mathcal{N}_C$ is the acceptor concentration in the collector region.

These boundary equations are coupled to the diffusion equations in the three quasi-neutral regions, which are unchanged from Eqs. (19.70). The currents of electrons and holes in the emitter and collector can then be calculated from

$$j_{En} = \quad e\mathcal{D}_E n_E'(x_a) \tag{19.80a}$$

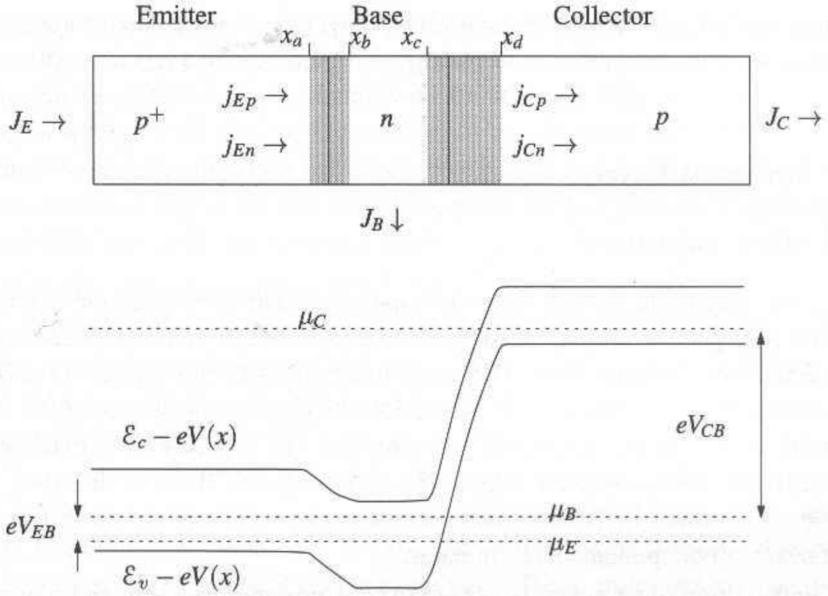

Figure 19.17. The binary junction transistor can be made from two back-to-back p–n junctions. When actively biased, the voltage of the emitter is raised by V_{EB} over the voltage of the base, and the voltage of the base is raised by $|V_{CB}| = -V_{CB}$ over the voltage of the collector. The chemical potential of electrons in the three regions is indicated by dotted lines. The left-hand depletion region shrinks relative to equilibrium while the right-hand one grows. The currents J are positive.

$$j_{Ep} = -e\mathcal{D}_B p'_B(x_b) \tag{19.80b}$$

$$j_{Cp} = -e\mathcal{D}_B p'_B(x_c) \tag{19.80c}$$

$$j_{Cn} = \ \ e\mathcal{D}_C n'_C(x_d). \tag{19.80d}$$

Solving the diffusion equations analogous to (19.74) in the three quasi-neutral regions subject to the boundary conditions (19.79) results in total currents J_E and J_C to the emitter and from the collector

$$J_E = J_{\text{FO}}(e^{\beta eV_{\text{EB}}} - 1) - \alpha_{\text{R}} J_{\text{RO}}(e^{\beta eV_{\text{CB}}} - 1) \tag{19.81a}$$

$$J_C = \alpha_{\text{F}} J_{\text{FO}}(e^{\beta eV_{\text{EB}}} - 1) - J_{\text{RO}}(e^{\beta eV_{\text{CB}}} - 1) \tag{19.81b}$$

with

$$J_{\text{FO}} = eA\left(\frac{\mathcal{D}_E}{L_E}\frac{n_i^2}{\mathcal{N}_E} + \frac{\mathcal{D}_B}{L_B}\frac{n_i^2}{\mathcal{N}_B}\coth(\frac{x_c - x_b}{L_B})\right) \tag{19.81c}$$

A is the area perpendicular to current flow in the transistor. L_B is the diffusion length in the base; see Eq. (19.76).

$$J_{\text{RO}} = eA\left(\frac{\mathcal{D}_C}{L_C}\frac{n_i^2}{\mathcal{N}_C} + \frac{\mathcal{D}_B}{L_B}\frac{n_i^2}{\mathcal{N}_B}\coth(\frac{x_c - x_b}{L_B})\right) \tag{19.81d}$$

$$\alpha_{\text{F}} J_{\text{FO}} = \alpha_{\text{R}} J_{\text{RO}} = eA\frac{\mathcal{D}_B}{L_B}\frac{n_i^2}{\mathcal{N}_B}\text{cosech}(\frac{x_c - x_b}{L_B}). \tag{19.81e}$$

Equations (19.81) are the *Ebers–Moll* equations; they form one of the bases for practical circuit design, and their detailed derivation is the subject of Problem 5.

Note that the diffusion length of the base L_B must be comparable to or greater than $x_c - x_b$, or else control of the collector current by the base is lost.

19.5 Inversion Layers

19.5.1 Heterostructures

The earliest electronic devices depended upon the contact between metal and vacuum, the next generation depended upon contact between metal and semiconductor, and the next industry depended upon junctions between regions of different doping, as well as junctions between semiconductors and insulators. A new generation of semiconductor devices is now evolving that depends upon junctions between different semiconductor alloys. The advantage of these is that they make possible the creation of *heterostructures* where the band gap varies in ways that would never occur spontaneously in nature.

A widely employed example is GaAlAs. Aluminum replaces gallium substitutionally in the alloy, lying right above it in column IIIA of the periodic table. The lattice constant of GaAs is 5.63 Å, that of AlAs, 5.62 Å, both adopting the zincblende structure, so there is no appreciable lattice distortion incurred by placing, say, a layer of $Ga_{.7}Al_{.3}As$ upon GaAs. However, the band gap of $Ga_{.7}Al_{.3}As$ is 1.82 eV, compared to 1.42 eV for GaAs. The technique of molecular beam epitaxy, described in Section 4.3, makes it possible to alternate layers of one alloy with another with atomic scale precision.

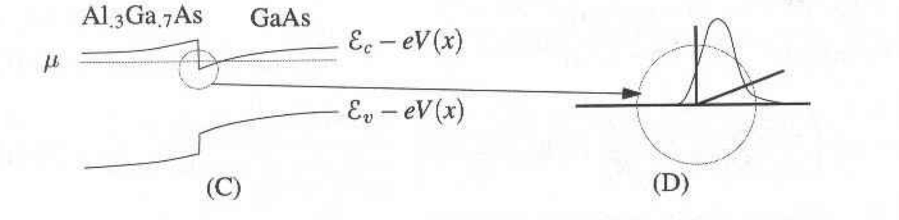

Figure 19.18. (A) Schematic picture of junction between two semiconductors with different band gaps, illustrated with numbers appropriate to $Ga_{.7}Al_{.3}As$. Calculating the band offsets is difficult and is discussed by Yu et al. (1992). (B) Same as (A), but drawing different quantities to illustrate band bending. (C) In case of heavy enough doping, the chemical potential can rise above the conduction band edge in a small notch-like region. (D) Enlargement of the conduction band region that would remain occupied even at temperature $T = 0$, with a sketch of a bound-state wave function trapped in the potential.

The formulas describing the profiles of charge around heterostructure junctions are not dramatically different from those of Section 19.4.2, and the main physical results can be deduced from diagrams in the spirit of Figs. 19.11 and 19.14, as displayed in Figure 19.18. The electron bands are discontinuous in the vicinity of the junction, which permits some interesting possibilities. A notch in the bands, such as shown in Figure 19.18(C), creates a small region that is occupied even at zero temperature, called an *inversion layer.*

Metal–Oxide–Silicon Junctions.

A similar notched potential can be created in a layered structure with a thin insulating coating separating metal and semiconductor, as illustrated in Figure 19.19. When the semiconductor is silicon and the insulator is silicon oxide, the junction is known by the acronym *MOS.* This combination can be used to create very compact, fast transistors, with low power dissipation, and has therefore become the most important technology in the creation of integrated circuits. The acronym CMOS refers to *complementary metal–oxide–silicon,* which means that both p- and n-type structures are built on the same chip. These structures are discussed in texts on semiconductor devices, such as Sze (1981) and Sze (1998).

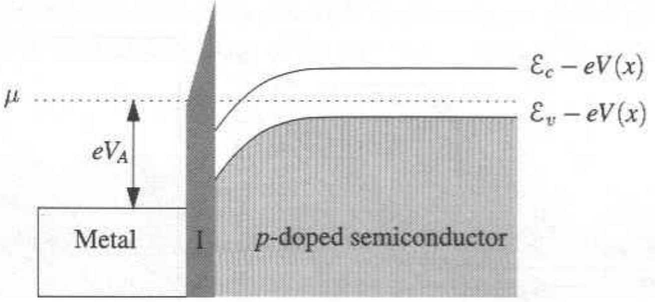

Figure 19.19. Metal–insulator–semiconductor (MIS), and, more particularly, metal–oxide–silicon (MOS) junctions provide an alternative to heterojunctions in forming inversion layers. By raising the voltage of the metal by V_A above the silicon, electrons are pulled over to the interface with the insulator, and the Fermi level μ can be pulled above the conduction band edge.

Two-Dimensional Electron Gas. Some of the most interesting physical discoveries in heterostructures have been built upon the *two-dimensional electron gas* (2DEG), the principle behind which was illustrated in Figures 19.18 and 19.19. By doping both sides of a heterojunction sufficiently, the chemical potential can be made to rise until it intersects a small corner of the conduction band, as shown in Figures 19.18(C) and 19.18(D). Even at the very lowest temperatures, electronic states must be populated in the vicinity of the corner. One way to view Figure 19.18(D) is that it sets up a one-dimensional problem of elementary quantum mechanics, which is to find the eigenstates of a particle in a triangular potential. As shown in Section 18.3.4, a one-dimensional attractive potential always has at least one bound state, no matter how shallow and small it may be. The potential barriers in the vicinity of the heterojunction are on the order of 0.1 eV. At room temperature, electrons would escape the restraining potential, and in fact the region to the right of the junction in Figure 19.18 would constitute an n-doped semiconductor in the degenerate limit. However, at temperatures of a few kelvin or less where experiments are characteristically performed, only the ground state has measurable occupation. This restriction to low temperatures is clearly a disadvantage. To overcome this restriction, it is not sufficient to find materials so that the energy scale of Figure 19.18 is multiplied by 100. The great mobility of electrons at low temperatures and the great purity achievable in semiconductors are equally important.

Figure 19.18 may lead to a mental picture in which electrons are trapped in one-dimensional potentials. The trapping is only in the z direction, as shown in Figure 19.20. Along x and y the electrons are free to move; the atomic sharpness of the heterojunction, the extreme purity of the samples, and the subkelvin temperatures all conspire to give electrons exceptionally high mobilities in the remaining two dimensions. For a GaAs–$Al_{0.29}Ga_{0.71}As$ interface, the electron mobility reaches 10^5 cm^2 V^{-1} s^{-1}, while the relaxation time τ can reach $4 \cdot 10^{-12}$ s. This relaxation time is two orders of magnitude larger than the characteristic values emerging from Eq. (16.7).

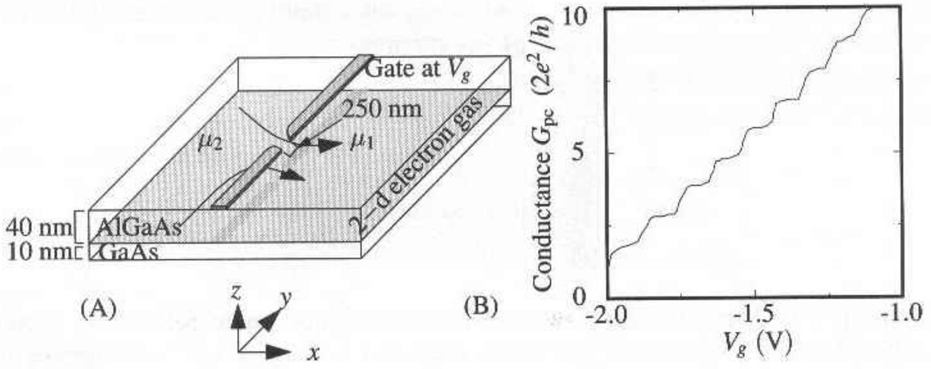

Figure 19.20. (A) Geometry of quantum point contact. Electrons can only pass through the region between the gates, which is shaped more like a blunted arrow than a long narrow channel. By raising and lowering the gate voltage V_g, the effective width of the constriction can be controlled. (B) Quantized conductance across the constriction, using Eq. (19.90) to process the raw data, observed by van Wees et al. (1988), p. 849.

The two-dimensional electron gas is the setting for many remarkable experiments, including the quantum Hall effect to be discussed in Section 25.5. In the context of electronic devices, it constitutes the starting point for building more elaborate structures.

19.5.2 Quantum Point Contact

To create a *quantum point contact*, two metal layers are deposited on top of a two-dimensional electron gas, as shown in Figure 19.21. By applying a negative voltage of around −0.5 V to the strips of metal, the Fermi level underneath them is driven downwards, and the electron gas completely depleted. The only path the electrons can follow is through the narrow channel left behind. As discussed by Beenakker (1997) and van Houten and Beenakker (1996), conductance through a channel of this type is quantized in units of $2e^2/h$.

Demonstrating this claim requires a fairly careful consideration of what electrical conductance really means. The quantum point contact is just a static quantum mechanical potential, through which wave functions travel or from which they reflect. Wave propagation conserves energy. Yet any wire with resistivity greater than zero must dissipate energy. How are these two views compatible? Landauer (1957) gave a conceptual resolution. He pointed out that experiments measuring conductivity contain the ingredients shown in Figure 19.21. Saying that there is a voltage difference between two points in a circuit really means that there are two reservoirs of electrons independently in thermal equilibrium, and with different chemical potentials, and that they have been connected by the channel whose conductance is to be measured. Any electron transmitted across the channel must give up energy, on average, once it arrives at the second reservoir, because the second reservoir is at lower potential than the first, and the arriving electron comes to equilibrium with its fellow electrons. All dissipation occurs in the reservoirs, not in the channel, but the dissipation is inevitable because of the way that voltage differences are defined.

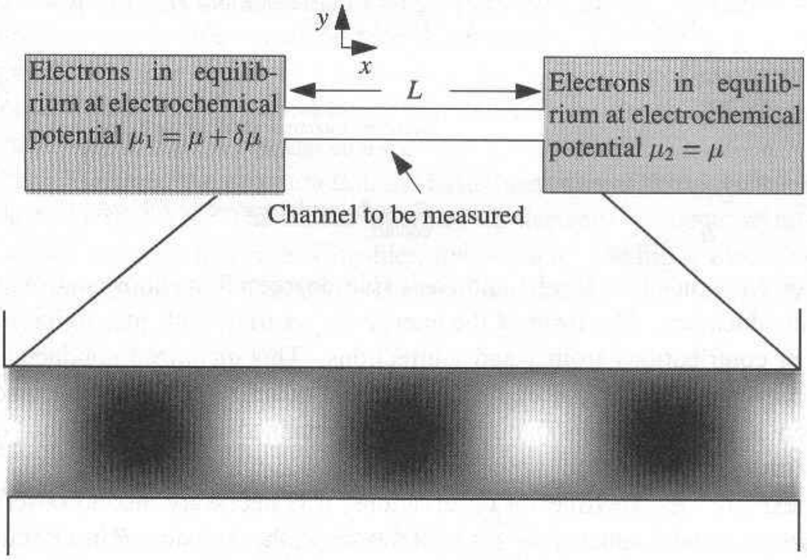
y
x
Electrons in equilibrium at electrochemical potential $\mu_1 = \mu + \delta\mu$
L
Electrons in equilibrium at electrochemical potential $\mu_2 = \mu$
Channel to be measured

Figure 19.21. Two reservoirs at electrochemical potential $\mu_1 = \mu + \delta\mu$ and $\mu_2 = \mu$ are connected by a channel so narrow that quantization of waves in the y direction becomes important.

The channel depicted in Figure 19.21 is so narrow that quantization in the y direction becomes important; when this experiment is performed, channel widths are on the order of tens to hundreds of nanometers. The energy levels in the channel therefore have the form

$$\mathcal{E}_{lk_x} = \mathcal{E}^y_l + \frac{\hbar^2 k_x^2}{2m}. \tag{19.82}$$

l is a discrete index. In fact, one does not even have to assume the free-electron form along x, just that energies in the two directions are additive.

For $\delta\mu > 0$, reservoir 1 will populate quantum levels in the channel that cause electrons to move down the channel along x, with no corresponding current coming back because those levels are empty in reservoir 2. The net current flowing through the channel is therefore

$$J = \frac{1}{L}\sum_{l\vec{k}} -ev_{lk_x}[f_2(\mathcal{E}_{lk_x}) - f_1(\mathcal{E}_{lk_x})] \tag{19.83}$$

v_{lk_x} is the velocity of an electron along the channel, and f_1 and f_2 are the Fermi functions of the two reservoirs. Summing over $\vec{k}$ counts all the particles in the channel, so multiply by v/L to get the flux.

$$= -e\sum_l \int dk_x D_{k_x} \frac{\partial \mathcal{E}_{lk_x}}{\partial \hbar k_x}[\theta(\mu + \delta\mu - \mathcal{E}_{lk_x}) - \theta(\mu - \mathcal{E}_{lk_x})] \tag{19.84}$$

Specialize to low temperatures, and change the sum over k_x to an integral using the one-dimensional density of states D_{k_x}.

$$= -e\frac{2}{2\pi\hbar}\sum_l \int_{\mathcal{E}^y_l}^{\infty} d\mathcal{E}\ [\theta(\mu + \delta\mu - \mathcal{E}) - \theta(\mu - \mathcal{E})] \tag{19.85}$$

$$= -e\frac{2}{2\pi\hbar}\delta\mu \sum_l \theta(\mu - \mathcal{E}^y_l) \tag{19.86}$$

Ignore the rare values of μ where $\mu + \delta\mu > \mathcal{E}^y_l$ and $\mu < \mathcal{E}^y_l$.

$$= \frac{2Ne^2}{h}V \tag{19.87}$$

Where $N = \sum_l \theta(\mu - \mathcal{E}^y_l)$ is the number of occupied quantum states along y, and $V = -\mu/e$ is the voltage.

$$\Rightarrow G_{\text{pc}} = \frac{2Ne^2}{h}. \tag{19.88}$$

G_{pc} is the conductance of the quantum point contact.

Thus each quantum level l and each spin degree of freedom contribute e^2/h to the conductance. The form of the energy $\mathcal{E}_{lk_x}$ is irrelevant, just so long as it is a sum of contributions from x and y directions. This quantized conductance has been observed, as shown in Figure 19.20. The quantization of conductance was first predicted by Imry (1987), but the subject remained controversial until matters were settled by experiment.

To explain the experimental observations, it is necessary also to observe that the quantum point contact is always in series with other resistors R in a circuit. The complete relation between current J and voltage V is

$$V = J\left(R + \frac{1}{G_{\text{pc}}}\right) \tag{19.89}$$

Conductance is the inverse of resistance.

$$\Rightarrow G_{\text{pc}} = \frac{J}{V - JR}. \tag{19.90}$$

The resistance R can be treated as a single free parameter to make the steps in G_{pc} of equal height.

19.5.3 Quantum Dot

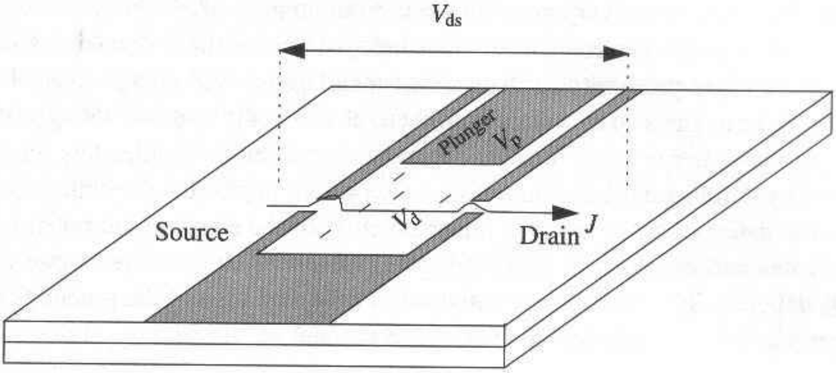

Figure 19.22. The quantum dot is a puddle of charge trapped between two quantum point contacts. Because of the micron-scale dimensions of the trapping region, the number of trapped electrons can be very small. One side of the trap is a plunger whose voltage can be raised and lowered to alter the electrostatic properties of the dot. The current J flowing from source to drain is the main experimental observable.

The *quantum dot*, as shown in Figure 19.22, is a structure one level more complex than the quantum point contact. It mainly consists of two quantum point contacts in series, but there is an additional interesting twist. The region between the point contacts is rather small, an area on the order of $0.5\mu\ \mathrm{m}\times 0.5\ \mu\mathrm{m}$. In rough analogy with the gate region of a transistor, there is also a metallic contact called the *plunger* whose voltage can be raised and lowered in order to affect the number of electrons in the central region.

Although named the quantum dot, the basic operation of this device is in large part curiously classical. The kinetic energies of electrons, so decisive in metals, are relatively small in this case. Consider, for example, placing N electrons into a quantum dot with area d^2. The single-electron quantum states would have energies approximately

$$\frac{\hbar^2 k^2}{2m} = 1.5\cdot 10^{-6}\frac{\mathrm{eV}}{d^2/[\mu\mathrm{m}]^2}. \tag{19.91}$$

This energy should be compared with the Coulomb repulsion of two electrons at distance d, which is

$$\frac{e^2}{d} = 1.4\cdot 10^{-3}\frac{\mathrm{eV}}{d/[\mu\mathrm{m}]}. \tag{19.92}$$

The difference in scale between the two energies is overstated by Eq. (19.92), because the Coulomb repulsion is diminished by screening effects to be discussed next, but it is still correct to start with Coulomb repulsion as the main physical effect and then add kinetic energy later as a small perturbation. The energy of electrons in the quantum dot can be treated as a purely classical problem of adding particles to a box, because no matter what the shape of their wave functions, only the Coulomb integral has much importance.

Screening and Capacitance. To begin, suppose that the quantum point contacts to the left and right of the dot are impenetrable barriers, and investigate how the energy of the dot would vary as a function of the number of electrons placed in it. The electrons in the dot cannot be taken independent of the rest of the universe. They are in close proximity to the various metal gates, and charge must flow in and out of these gates so as to maintain them at externally applied voltages whenever electrons enter or leave the dot. Classical electrostatics handles this screening problem by defining a capacitance matrix $C_{\alpha\beta}$ which posits that the charge $Q_{\alpha\beta}$ on any of the gates, or in the dot, is a linear function of the electrostatic potentials V_α of the gates and the dot. To make things simple, suppose that the charge on the dot $Q_{\rm d}$ depends only upon the potential within the dot, $V_{\rm d}$, and the potential of the plunger, $V_{\rm p}$. Write the charge on the dot and plunger as

$$Q_{\rm d} = C_{\rm d}V_{\rm d} - C_{\rm dp}V_{\rm p}, \tag{19.93}$$

The minus sign in front of $C_{\rm dp}$ is conventional, and it ensures that $C_{\rm dp}$ will be positive.

$$Q_{\rm p} = -C_{\rm pd}V_{\rm d} + C_{\rm p}V_{\rm p}. \tag{19.94}$$

Because the possibility of electron motion through the junctions is being neglected for the moment, the only feature of the outside world with which electrons in the dot interact is the plunger. Therefore the charge on the dot must be a function of $V_{\rm d} - V_{\rm p}$, which means that

$$C_{\rm d} = C_{\rm dp} = C_{\rm pd}. \tag{19.95}$$

The capacitance matrix must be symmetric, because it is given by second derivatives of the energy U in Eq. (19.96) with respect to potential, and therefore $C_{\rm pd} = C_{\rm dp}$.

The plunger is not similarly isolated. It is connected to a large reservoir of electrons at potential $V_{\rm p}$ that enables it to remain at potential $V_{\rm p}$ no matter what happens on the dot. The electrostatic energy of the system is therefore

$$U_{\rm electrostatic} = \frac{1}{2}\,[Q_{\rm d}V_{\rm d} + Q_{\rm p}V_{\rm p}] + [Q_{\rm reservoir} - Q_{\rm p}]V_{\rm p}. \tag{19.96}$$

$$= \frac{Q_{\rm d}^2}{2C_{\rm d}} + V_{\rm p}Q_{\rm d} + \dots. \tag{19.97}$$

The remaining terms depend only upon $V_{\rm p}$, are independent of $Q_{\rm d}$, and so can be dropped. Make use of Eqs. (19.93), (19.94), and (19.95).

The number of electrons preferred on the dot in equilibrium is given by minimizing Eq. (19.97) with respect to $Q_{\rm d}$ and is

$$N \equiv \frac{Q_{\rm d}}{-e} = \frac{C_{\rm d}V_{\rm p}}{e}. \tag{19.98}$$

If $C_{\rm d}$ were a capacitance on the order of 1 farad, this equilibrium number would be immense. The point of the quantum dot is to generate capacitances $C_{\rm d}$ so small that the equilibrium occupation is of order unity—that is, capacitances on the order of aF$= 10^{-18}$ F (anofarad). In terms of this unit, Eq. (19.98) can be rewritten as

$$N = 0.625\frac{C_{\rm d}}{100\text{ aF}}\frac{V_{\rm p}}{10^{-3}\text{ V}}, \tag{19.99}$$

showing that voltages on the order of millivolts applied to the plunger should produce changes of order unity in the number of electrons sitting on the dot. For such small numbers of electrons, one must take into account the fact that N is an integer. What Eqs. (19.98) and (19.99) in fact predict is that the number of electrons N increases in steps, with a transition occurring whenever states with N and $N+1$ electrons have the same energy, at a voltage

$$V_{\mathrm{p}} = [N+1/2]\frac{e}{C_{\mathrm{d}}}. \qquad (19.100)$$

Set Eq. (19.97) equal to itself evaluated at $-Q_{\mathrm{d}}/e = N$ and $N+1$, using Eq. (19.98), and solve for V_{p}.

Having established the energetics of the problem in a simple classical fashion, quantum features begin to creep into the interpretation of the results. The theory based upon Eq. (19.98) is the theory of the *Coulomb blockade*, and it makes three main predictions.

1. If a very small voltage is applied across the quantum dot, from source to drain, the current from source to drain should show sharp narrow peaks as a function of the plunger voltage V_{p}, with the peaks spaced in voltage by a distance e/C_{d}.
2. For fixed plunger voltage V_{p}, current from source to drain should be relatively tiny until the voltage from source to drain exceeds a critical threshold, either positive or negative. The gap in voltage between the negative and positive thresholds is e/C_{d}.
3. The characteristic energy scale on which temperature fluctuations should destroy these effects is $k_B T \sim e^2/2C_{\mathrm{d}}$, which works out to be a few kelvin.

The logic behind these predictions has to do with imagining physically how electrons will manage to traverse the quantum dot in the presence of a voltage between source and drain. In order to do so, an electron must tunnel across the first quantum point contact, dwell for some time on the dot, and then tunnel across the second quantum point contact. If according to Eq. (19.98) the electrostatic energy of the dot goes up when an extra electron hops on, tunneling will be made difficult. There are three ways around. First, whenever the plunger voltage sits at one of the values indicated by Eq. (19.100), the energies of having N and $N+1$ electrons on the dot are degenerate. There is no energy penalty preventing an electron from flowing in and out of the dot, so for these special plunger voltages the dot has a high conductivity, leading to prediction 1. Second, for an arbitrary plunger voltage, the voltage between source and drain can be made large enough that it supplies the energy needed to hop on and off the dot. Hence a prediction that current through the dot will rapidly increase after a critical threshold no matter what the plunger voltage. Third, thermal fluctuations may be large enough to supply the missing energy, leading to the final prediction of a temperature scale on which the quantum effects disappear.

All three of these predictions are beautifully verified by experiment. Figure 19.23 shows both (A) the periodic current peaks as a function of plunger voltage, and (B) the very nonlinear relation between current and source–drain voltage.

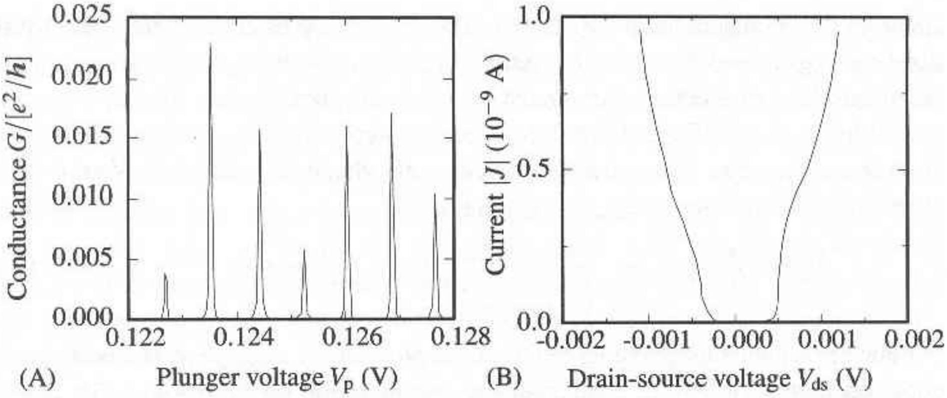

Figure 19.23. (A) Conductance as a function of plunger voltage displaying sharp, equally spaced peaks. (B) Current across a quantum dot plotter versus the voltage V_{ds} across the dot. [Source: Meirav and Foxman (1996), p. 257.]

Problems

1. **Chemical potential in intrinsic semiconductor:** Consider a crystal of silicon, with a very low level of doping, but a slight excess of acceptors over donors.

(a) At sufficiently low temperatures, the chemical potential moves far away from the center of the gap. At what temperature does this happen, and why?

(b) Next, consider the same situation, but with a slight excess of donors over acceptors, and answer the same questions.

2. **Ohmic junction:** Diagrams of electronic circuits frequently show wires connected to portions of semiconductors. These connections are supposed to be ohmic, to conduct current with equal ease in either direction, and in linear proportion to applied voltage. Because metal–semiconductor junctions have intrinsic rectifying properties, ohmic response cannot be taken for granted.

(a) To have a sense of ways to obtain such junctions, copy the three parts of Figure 19.11, but assuming that the work function of the semiconductor is greater than the work function of the metal.

(b) Argue from the graphical construction that the rectifying powers of the junction are plausibly diminished, by considering as in Figure 19.12 how the band bending is affected by an applied voltage.

3. **Thermopower of semiconductors:**

(a) Consider an n-doped semiconductor, where transport is dominated by electrons in the conduction band. Measure all energies $\mathcal{E}$ relative to the bottom of the conduction band, so $\mathcal{E}_c = 0$. Assume that $\mathcal{E} = m^\star v^2/2$, that $D(\mathcal{E}) \propto \sqrt{\mathcal{E}}$, that $\tau_\mathcal{E} = a\mathcal{E}^{-s}$, and that the matrices of Eq. (17.62) are all diagonal. Equation

(17.63) is still valid, but for semiconductors the approximation in Eq. (17.66) cannot be used. Show that the thermopower α is given by

$$\alpha = -\frac{k_B}{e}\left[\frac{5}{2} - s - \frac{\mu}{k_B T}\right] \cdot \text{ The identity } \Gamma(1+x) = x\Gamma(x) \text{ is helpful.} \tag{19.101}$$

(b) Now consider a p-doped semiconductor, where transport is dominated by holes. How does Eq. (19.101) change in this case?

4. **Carriers in depletion region:** Using Eqs. (19.67), verify Eqs. (19.71). Notice that because (19.71) is obtained from expressions for the current, it cannot be used directly to predict it.

5. **Ebers–Moll equations:**

(a) Write down the three equations analogous to Eqs. (19.74) for minority carriers in the three quasi-neutral regions of the transistor.

(b) Write down the solutions of these equations; in the collector and emitter regions, the solutions are immediately determined up to an overall constant, while in the base region, there are two unknown constants to calculate.

(c) Find the unknown constants by imposing boundary conditions (19.79), and show that the currents described by (19.80) can be put in the form (19.81).

References

J. A. Appelbaum and D. R. Hamann (1976), The electronic structure of solid surfaces, *Reviews of Modern Physics*, **48**, 479–496.

J. Bardeen (1947), Surface states and rectification at a metal semi-conductor contact, *Physical Review*, **71**, 717–727.

J. Bardeen and W. H. Brattain (1948), The transistor, a semi-conductor triode, *Physical Review*, **74**, 230–231.

J. C. Bean (1986), The growth of novel silicon materials, *Physics Today*, **39**(10), 36–42.

C. W. J. Beenakker (1997), Random-matrix theory of quantum transport, *Reviews of Modern Physics*, **69**, 731–808.

F. Braun (1874), On current flow in metallic sulfides, *Annalen der Physik und Chemie*, **229**, 556–563. In German.

L. L. Chang and L. Esaki (1992), Semiconductor quantum heterostructures, *Physics Today*, **45**(10), 36–43.

L. F. Eastman (1986), Compound-semiconductor transistors, *Physics Today*, **39**(10), 77–83.

P. M. Fahey, P. B. Griffin, and J. D. Plummer (1989), Point defects and dopant diffusion in silicon, *Reviews of Modern Physics*, **61**, 289–384.

F. J. Feigl (1986), VLSI technology and dielectric film science, *Physics Today*, **39**(10), 47–54.

D. K. Ferry and H. L. Grubin (1995), Modeling of quantum transport in semiconductor devices, *Solid State Physics: Advances in Research and Applications*, **49**, 283–448.

A. Fowler (1997), On some modern uses of the electron in logic and memory, *Physics Today*, **50**(10), 50–54.

A. B. Fowler (1993), A semicentury of semiconductors, *Physics Today*, **46**(10), 59–62.

J. M. Gibson (1997), Reading and writing with electron beams, *Physics Today*, **50**(10), 56–61.

N. C. Greenham and R. H. Friend (1995), Semiconductor device physics of conjugated polymers, *Solid State Physics: Advances in Research and Applications*, **49**, 1–149.