

1 Semiconductor Devices

January 19, 2008

At the turn of the last century it was known that most metals had a profusion of electrons, low-mass carriers of negative charge, and that insulators had few free electrons. Semiconductors were somewhere in between metals and insulators. The sign charge of the carriers was known from the Hall effect, the bending by a magnetic field of an electric current away from the direction of the applied electric field. It was also known that some materials appeared to have the opposite sign of Hall resistance, indicating positively charged carriers, while others had almost no Hall resistance. These and some other features of semiconductors and metals were unexplained until the main features of quantum mechanics were developed in the first quarter of the last century.

The most important features of the modern quantum theory developed by Planck, Einstein, Bohr, Sommerfeld, de Broglie, Pauli, Schrödinger, Heisenberg, and Dirac are:

1. Electromagnetic waves, such as light, radio waves and X-rays, are emitted and absorbed as discrete particles whose energy E is related to the frequency ν by $E = h\nu$, where h is *Planck's constant*. We often write this as $E = \hbar\omega$, where $\hbar = h/2\pi$, and ω is the angular frequency $2\pi\nu$. The quanta of light and of other electromagnetic radiation are known as *photons*, while the quanta of sound are known as *phonons*.
2. Particles such as electrons, nuclei and atoms also have a wave-like character, and are described by a complex wave function ψ that satisfies the Schrödinger equation.
3. A particle cannot simultaneously be assigned a definite position and momentum. The probability that the particle is near any point \mathbf{r} is proportional to $|\psi(\mathbf{r})|^2$.
4. Electrons, neutrons and protons are known as *fermions*, and fermions satisfy the *Pauli exclusion principle*, that says there can be not more than one fermion occupying each state allowed by the Schrödinger equation. Electrons, neutrons and protons have two possible *spin states*, that can be separated in energy by a magnetic field.

5. Photons and phonons are *bosons*, and bosons have no such restriction on the number in each state. Compound particles containing an even number of fermions, such as a helium atom (two protons, two neutrons and two electrons), are also bosons.
6. An important experimental discovery by Rutherford was that most of the mass of an atom is concentrated in the *atomic nucleus*, which occupies a tiny fraction of its volume, generally less than a trillionth part (1 part in 10^{12}).
7. Another important element of the theory of metals, insulators and semiconductors, which was developed by Brillouin, is that in a periodic medium, such as is provided by the periodically arranged atoms in a crystal, there are gaps in the spectrum of allowed energy levels.

These developments immediately allowed theorists such as Felix Bloch and Rudolf Peierls to explain the hitherto mysterious properties of metals, insulators and semiconductors, and I give a brief outline of this theory in the next section.

There were two major discoveries in more recent years, between sixty and fifty years ago, which led to the fantastic reductions in size and cost and equally fantastic increases in speed, of modern electronics. Both of these have been recognized by the award of Nobel Prizes in Physics. The first was the construction of the semiconductor transistor by Bardeen, Brattain and Shockley at Bell Laboratories. Once the problems of large-scale manufacture had been overcome this allowed bulky, fragile, unreliable and power-consuming thermionic vacuum tubes to be replaced by small, robust, reliable and energy-saving semiconductor transistors.

The second was based on an older idea that only became practical when high purity silicon became readily available. That was to confine the conduction electrons to the interface between semiconducting silicon and insulating SiO_2 and to manipulate the electrons by electric fields applied across the insulating layer. This allowed the transistors and the connections between them to be shrunk greatly in size, and, most importantly, allowed large integrated silicon circuits to be printed with few wires connecting them.

1.1 Charge transport in ideal metals and insulators

In a perfect crystal, with regular spacing between the atoms, the allowed energy levels are given by wavelike solutions of the Schrödinger equation. If we consider a one-dimensional array of atoms at the points $x = na$, where a is the interatomic spacing, the wave functions have the form

$$\psi(x) = \cos(kx)f_k^\alpha(x) \quad \text{or} \quad \sin(kx)f_k^\alpha(x), \quad (1)$$

where f_k^α is a periodic function of x , so that it satisfies $f_k^\alpha(x + na) = f_k^\alpha(x)$ for any integer value of n . The energies of these states are smoothly varying functions of k in the range $-\pi/a < k \leq \pi/a$, and these energies form what is known as an *energy band*. However, there are many energy bands corresponding to different values of the label α , and the energy bands are separated from one another by an *energy gap*.

The number of electrons of a given spin direction that is needed to fill all the available states in any band is equal to the number of atoms in the system. Since there are two possible spin directions for the electron, there are two electrons per atom needed to fill each band. The electrons in a filled band are prevented from moving by the exclusion principle, so such a system with filled bands is an insulator at low temperatures. At high temperatures electrons can be thermally excited from the filled band to the empty band above it, so that the thermal conductivity increases with temperature. An array of helium atoms, with two electrons per atom, is an example of such an insulator, although solid helium only exists at quite high pressures even when the temperature is close to zero. An array of hydrogen molecules, where the atoms are bound in pairs to form molecules, with two electrons per molecule, is also insulating. If the hydrogen atoms were equally spaced, each band for the two spin directions would be half filled, and the electrons would be free to move in response to a weak electric field. It is believed that metallic hydrogen may exist at high pressures such as occur in the interior of Jupiter.

In three dimensions the story is a little more complicated, as the energies of the bands may overlap, in such a way that there is no gap between them. For that reason the atoms with three and four electrons per atom, lithium and beryllium, are conductors. In a diamond crystal there is a second band that contains six electrons per atom, which is known as the *valence band*, since it is responsible for the chemical bonding of one atom to its four neighbors. The empty band immediately above it is known as the *conduction band*, since electrons thermally or optically excited to it can carry an electric current

while the holes left behind in the valence band also allow the valence band to carry current. In the case of diamond the gap between these two bands is very large, so the diamond is transparent to light. In the next row of the periodic table the first three elements, sodium, calcium and aluminum, are metals with no band gap corresponding to calcium, which has twelve electrons per atom. In silicon, immediately below carbon in the periodic table, each atom bonds to its four neighbors to form a diamond lattice, and the valence band is filled. The gap to the conduction band is relatively small, only 1.1 eV (electron volts), which is comparable with the quantum of energy for optical radiation, so silicon is not transparent, but looks metallic. This is much larger than the typical quantum of thermal energy at room temperature, which is about 0.025 eV, so there are very few electrons in pure silicon excited from the valence band to the conduction band, and pure silicon is more or less an insulator at room temperature.

There are many other semiconductors which have similar crystal structure to silicon, but silicon has been the dominant material for most applications, to some extent because its widespread use has led to efficient and cheap production of high quality materials. An early competitor was germanium, also in Group IV of the periodic table, which has the same crystal structure and a slightly smaller band gap. There are materials in which alternating atoms in the crystal are replaced by Group III and Group V elements, to form semiconductors such as GaAs (gallium arsenide) or InSb (indium antimonide). Similar materials can also be formed by combining group II and Group VI elements, for example CdS (cadmium sulfide).

High purity germanium is used to detect the passage of high energy particles in experimental particle physics, because the passage of such a particle excites many electrons from the valence band, giving a large temporary increase in the electrical conductivity of the germanium. Gallium arsenide is used to make the high performance amplifiers that are an essential feature of cell phones.

1.2 Doped semiconductors and the *pn* junction

Normally produced semiconductors have quite variable electrical properties, because the number of charge carriers is sensitive to rather small amounts of impurities. The solution to this problem is deliberately to *dope* them with particular types of impurities. One important type of impurity is a group V element such as phosphorus, which will displace a silicon atom from its

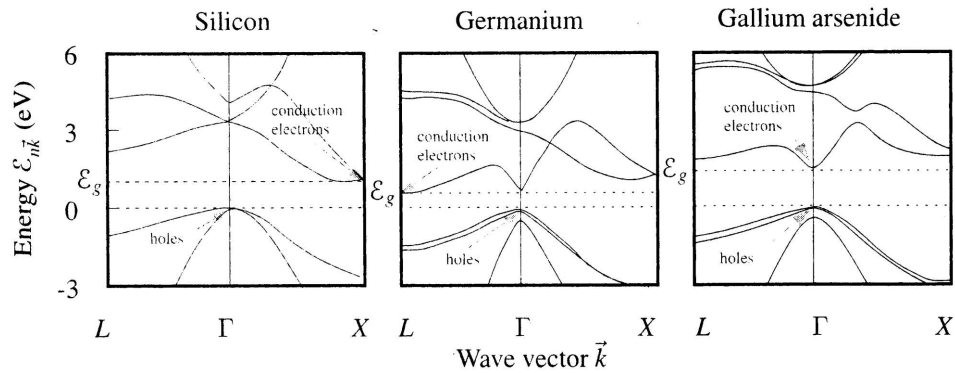


Figure 1: Energy bands in three typical semiconductors. The point Γ represents $\mathbf{k} = 0$, while the curves going out from Γ to the left and right represent plots for two distinct directions of \mathbf{k} . In all three cases the valence band has its highest energy at $\mathbf{k} = 0$, but only for GaAs is the minimum energy of the conduction band at Γ . For Si and Ge the minimum of the conduction band is far from Γ , and it is in different positions for the two semiconductors. This figure is copied from Fig.19.8 of Marder, *Condensed Matter Physics*.

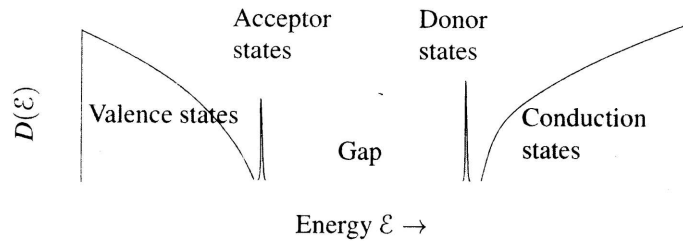


Figure 2: Density of states near the energy gap as a function of energy for a typical semiconductor, doped with both acceptors and donors. This figure is copied from Fig.19.10 of Marder, *Condensed Matter Physics*.

place in the lattice. Phosphorus has one more electron than silicon, and this electron cannot fit into the full valence band. This extra electron may remain bound to the phosphorus atom, with an energy just below the lowest energy of the valence band, or, at room temperature, it has a high probability of being excited to the conduction band, serving as a mobile negatively charged, or *n*-type, carrier, and a semiconductor doped in this way is known as an *n*-type semiconductor. Since these impurities donate electrons to the conduction band, such impurities are known as *donors*.

The other possibility is to dope the silicon with group III impurities such as aluminum or boron. These have one less electron than is needed to fill the valence band, and little energy is needed for the impurity to capture an extra electron, leaving a *hole* in the valence band. Such holes behave like positive charge carriers, since motion of a hole to the left implies motion of the electrons to the right. These holes are therefore called *p*-type carriers, and the semiconductors doped with atoms with one less valence electron are known as *p*-type semiconductors. Since these impurities accept electrons from the valence band, such impurities are known as *acceptors*.

A third type of semiconductor, with a very small proportion of either negative or positive carriers, is known as an *intrinsic* semiconductor. This could be an undoped semiconductor, but this has an electrical conductivity that is very sensitive to the residual impurities, and for many purposes it is more satisfactory to make a *compensated* semiconductor, which is doped with both donors and acceptors, in such a way that neither the donor nor the acceptor impurities are appreciably ionized.

It is a very important feature of a doped semiconductor that the binding of the electron to the positive ion, or of the hole to the negative ion, should be low enough that there is a high probability of dissociation of the bound state at room temperature. This probability is proportional to $\exp(-E_0/k_B T)$, where E_0 is the difference between the lowest energy in the conduction band and the energy of the electron in the bound state, and k_B is the Boltzmann constant. The electron bound to the positive phosphorus ion is in a state which is closely analogous to the ground state of the hydrogen atom, but the binding energy is much lower than the 13 eV binding energy of the hydrogen atom, primarily because the dielectric constant of a semiconductor ϵ is much larger than unity, and this reduces the binding energy by a factor $1/\epsilon^2$. Semiconductors typically have dielectric constants of 10 or more, and so E_0 is of the order of 0.1 eV, which is low enough to give a high probability of excitation of the electron into the conduction band.

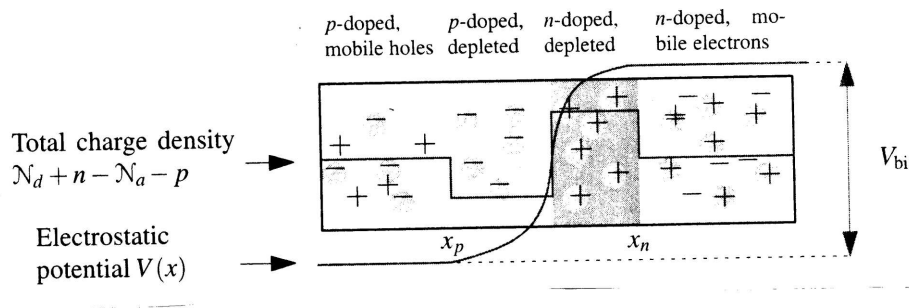


Figure 3: Typical arrangement for a pn junction, showing the p region to the left and the n region to the right, with a shaded depletion region between them, where the energy needed to ionize the impurities is too great, so there are few free charge carriers. This figure is copied from Fig.19.15 of Marder, *Condensed Matter Physics*.

The simplest electronic component that can be made from a semiconductor is a rectifier, a single pn junction, which consists of a p type region in contact with an n type region. At the edge of the p -region there are extra negative ions and a reduced number of holes, while at the edge of the n -region there are extra positive ions and a reduced number of free electrons, and these produce an electric field pointing from the p -region to the n -region. If a more positive potential is applied to the p type region than to the n type region, then the energy gap in the intrinsic region is increased, and little current will flow. If a more negative potential is applied to the p type region than to the n type region, then the energy gap in the intrinsic region is decreased, the number of carriers in this region is increased, and more current will flow. This current consists both of electrons going from the cathode (n type region) to the anode (p type region) and of holes going from the anode to the cathode. More complicated electronic components, such as amplifiers, can be made by combining p and n regions in various ways.

1.3 Inversion layers and interface electronics

An inversion layer may be formed by taking a block of p type silicon, bounded by a thin insulating layer of glassy SiO_2 and putting a metal electrode in contact with the oxide layer; this electrode is known as a *gate*. When a suf-

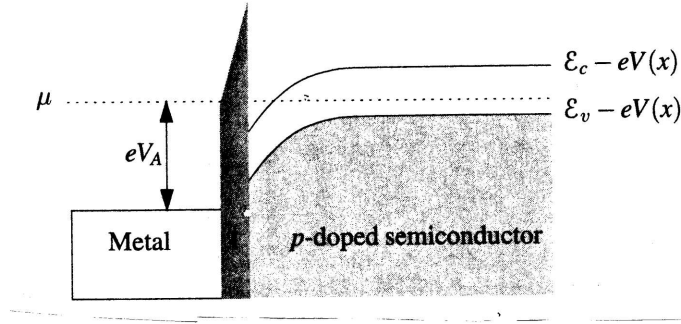


Figure 4: Configuration for a MOS device, the basis of most modern electronics. A potential applied to the metal gate attracts electrons to the surface region, where they form a negatively charged layer separated from the p -type semiconductor by an insulating depletion region. Modulation of the gate voltage can make big changes in the conductivity of the negative inversion layer, and so act as an amplifier. This figure is copied from Fig.19.19 of Marder, *Condensed Matter Physics*.

ficiently large negative potential is applied to the electrode, so that there is a strong electric field going into the semiconductor, it may be energetically favorable for electrons to be pulled out from the neighboring region of the valence band, to form what is known as an *inversion layer* of negative electrons close to the interface. Beyond this there is a *depletion region* in which there are no mobile carriers, just positive acceptor ions, and these screen the interior from the applied electric field. Finally there is the bulk region of mobile holes.

Connections may be made to strongly n type regions in contact with the inversion layer, and metal wires can go into these n type regions. An external current can be passed through the device. By modulating the gate voltage, the density of carriers in the inversions layer can be modulated, and this results in a modulated current. In this way the device can serve as an amplifier of the gate voltage. Such a device is known as a MOSFET (Metal-Oxide-Semiconductor Field Effect Transistor).

The great advantage of using a device of this sort, operating with an inversion layer is that complicated *integrated circuits* can be constructed on a single thin layer of silicon. Current will only be carried in the inversion layer where the gate voltage is sufficiently high, so switches can be made which

operate by adjustment of the gate voltage above them. The shrinking of circuit elements and the increasing complexity of integrated chips over the years have been obtained by shrinking the thickness of the silicon wafers and of the insulating layer on top of them, and thus allowing more compact circuit elements to be made without interference between neighboring elements.

1.4 Applications of semiconductor devices

Bell Laboratories was the birthplace of the semiconductor electronic revolution, but the Bell company continued to make telephones and switchboards in much the way that it had made them in the past. Large-scale manufacture of high-performance semiconductors caused Texas Instruments to flourish, and a little-known Japanese company called Sony had the idea of selling cheap and reliable transistor radios to the general public.

In the fifties, when I was a student, computers were rare, bulky, expensive, and liable to crash in the middle of a calculation. Cambridge University, unlike most British universities, had its own computer, which was home-made and designed. The thermionic valves filled a large room, perhaps 3 m by 5 m. Its speed was about 250 Hz, and the size of its RAM perhaps 10 kB. When I wanted to do a calculation that would take about 15 minutes (225,000 operations) I had to present a detailed case to a committee of seven distinguished professors. Around that time there was a debate on the maximum number of computers that might be needed in Britain, and estimates of around twelve were being made. My current, obsolescent, computer has a volume of 2 liters, down by a factor of 20,000, its speed is 667 MHz, up by a factor of 2.5 million, and its RAM size 256 MB, about a factor 25,000 larger. I do not know the difference in cost, but I would guess that it is a factor less than 10^{-4} . Such changes in size, speed, capacity, and cost do not only allow much more ambitious computer projects, but they also allow for the construction of cheap processors to do simple tasks like adjusting the timing of sparks in a car engine.

A different early application of silicon technology was again developed in Bell Laboratories, by Chaffin, Pearson and Fuller in 1954. They used a *pn* junction as a device to convert the energy of light from the sun to electrical power. Each photon falling on the intrinsic region between the *p* and *n* regions of a large *pn* junction is absorbed in the silicon, exciting one electron and one hole. The strong electric field between the *n* and *p* layers draws the electrons to the *p* layer and the holes to the *n* layer. The external connections

to these layers will act just like the terminals of a cell in a battery, providing a source of electrical power at about 1 V. The p layer acts as the negative terminal and the n layer as the positive terminal. To get a useful source of power these cells have to be connected in series, to get a high enough voltage, and in parallel to get a high enough output current. The original silicon solar cells were very expensive, but were sold to the government for high cost purposes, such as providing power for the instruments on a satellite. Now they are lower in price, so that they are useful in remote areas with no mains electricity, and can be made competitive in urban areas with the help of a government subsidy, particularly in countries like Japan, Germany and Israel with few internal conventional sources of power.

In the other direction, injection of a current into a semiconductor can excite electron-hole pairs, and then an electron and a hole can recombine and emit a photon, so that the electric power used up is partially converted into light, at much higher efficiency than power is converted into light in an incandescent light bulb. As a result of this higher efficiency there is less heat generated, and so they can be made more compact.

It has already been mentioned that high purity semiconductors can be used to detect the passage of the high energy charged particles created by the collisions of a beam of elementary particles with a target in experimental particle physics. The increase in conductivity of such a detector gives a measure of the energy loss of the charged particle as it passes through the individual detector, and the sum of responses from a large array of such detectors gives a measure of the total energy released in the collision.

Semiconductors can be used to construct lasers, as will be discussed in more detail later. When particle-hole pairs have been excited in a semiconductor by injection of current, many of the holes and electrons will lose energy until they are in the highest energy states of the valence band and the lowest energy state of the conduction band respectively. If these two extreme states, with wave functions of the form of Eq. (1), have the same value of \mathbf{k} , particles and holes can recombine emitting a single photon, whose quantum energy is equal to the band gap. Because each electron-hole pair can emit a photon of the same energy, the emission of one photon can stimulate the emission of many more; this typical of laser action. Silicon and germanium are not suitable for this, because the valence bands have their maxima at $\mathbf{k} = 0$, but the conduction bands have their minima at different values of \mathbf{k} .