LECTURE 12

Solid State Band Theory

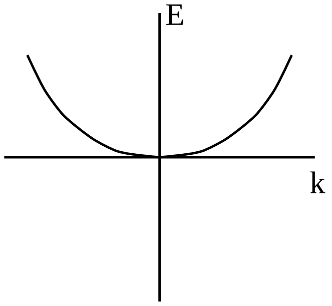
Recall from lecture 7 that when atoms made chemical bonds to form molecules, their atomic orbitals overlapped to create molecular orbitals whose energies were shifted relative to the energies of the original atomic orbitals. These energy shifts resulted from interactions between the atoms. In particular the atoms must be close enough to each other to transfer electrons between them. This is often called electron hopping. Now suppose we put lots of atoms together in a regular periodic array to form a crystalline lattice. There must be some interaction and electron transfer between the atoms in order to hold the lattice together. It's like a giant molecule with 6×10^{23} atoms. Just as in molecules, the bonding can be ionic, covalent, or metallic (delocalized). The electronic energy levels in the solid are shifted relative to the levels in the atom. The number of energy levels is conserved. So if there are N atoms in the solid, each with m energy levels, then there are Nm energy levels in the solid (each of which can hold 2 electrons, spin up and spin down).

Now let's consider how the energy levels are arranged. If the atoms had no interactions, e.g., if they were very far apart, then the atomic energy levels would not be shifted but would have a high degree of degeneracy, i.e., there would be N levels at each of the atomic energy eigenvalues. Let's look at one energy where there are N degenerate levels. As we turn on the interactions between the atoms, e.g., by bringing the atoms closer together, this degeneracy is lifted and the levels spread out in energy. The maximum spread will be of the order of the interaction energy. (This maximum spread in energy is called the bandwidth.) Since there will be N levels packed into this range, the separation between levels is quite small. As $N \to \infty$, the energy level separation goes to zero. We call this spread of levels a "band." The energy difference between the highest and lowest levels is the "bandwidth."

It's often useful to plot the energy E versus wavevector k of the band. Let's start by considering a free electron gas. In this case the system is isotropic, homogeneous, and has complete translational symmetry. So the momentum $\vec{k}$ is a good quantum number. For simplicity let's consider a one dimensional system. The energy is given by

$$E = \frac{\hbar^2 k^2}{2m} \tag{1}$$

This is a parabola.



Now let's consider a crystal with a periodic array of atoms. Now we only have discrete translational symmetry. Let a be the distance between 2 atoms in the lattice. a is called the lattice constant. For simplicity, let's consider a one dimensional lattice of length $L = Na$ where N is the number of atoms in the lattice. As we saw in lecture 1 when we considered the case of periodic boundary conditions, we found that the allowed wavevectors were

$$k = \frac{2\pi n}{L} = \frac{2\pi n}{Na} \tag{2}$$

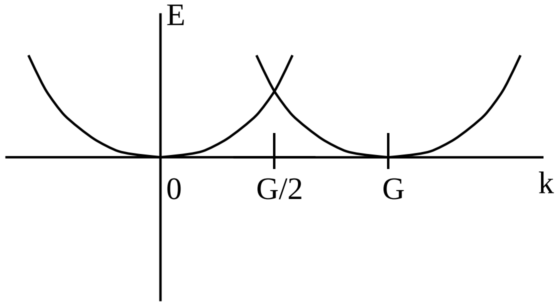
For example, if we have $N = 4$ sites, then $ka = 0,\ \pi/2,\ \pi,\ 3\pi/2,\ 2\pi$, etc. We would replace our graph of E vs. k by discrete states with these lattice vectors. Periodic boundary conditions in a one dimensional lattice amounts to having a ring. It's like a pearl necklace with each pearl being a lattice site. There is a one–to–one correspondence between the atoms in the lattice and the states in a band. Each atom contributes one electronic level to each band. When $n = Nm$ in (2),

$$k = G = \frac{2\pi m}{a} \quad \text{where} \quad m = 0, 1, 2, 3, ... \tag{3}$$

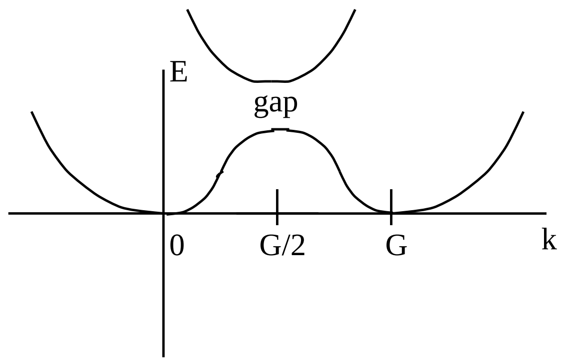
These values of the wavevector k are special. They are called reciprocal lattice vectors. Let's denote reciprocal lattice vectors by G. Notice that

$$\exp(iGa) = \exp(i\frac{2\pi m}{a}a) = \exp(i2\pi m) = 1 \tag{4}$$

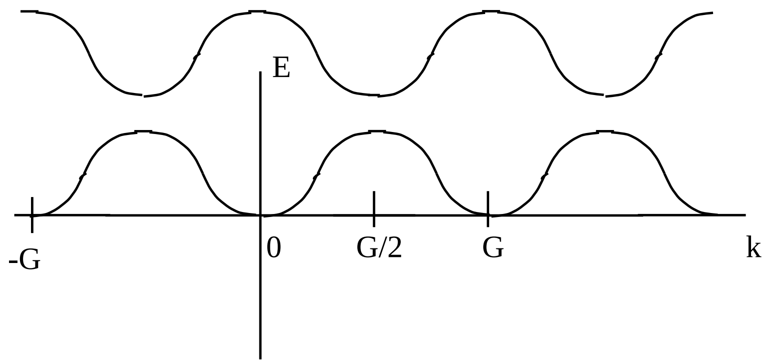
Just as the lattice looks the same if you translate by a lattice vector $\vec{a}$ or any of its multiples, so in $k-$space, things look the same if you translate by a reciprocal lattice vector G. So I could draw



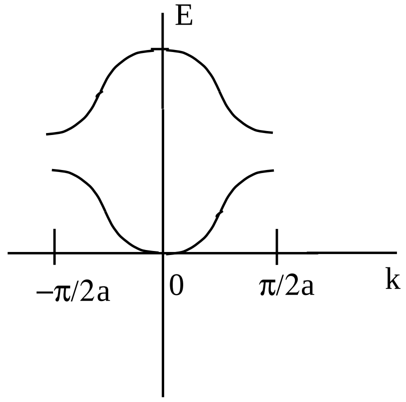
Notice that the curves cross at $G/2$. This level degeneracy is split by the periodic potential of the lattice of ions, resulting in band gaps:



The magnitude of the gap is of the same order of magnitude as the periodic potential ($V(G)$). So we have 2 bands which are separated by a gap:



This is the repeated zone scheme. Since translating by G doesn't add anything new, we usually just draw:



The range from $-G/2 = -\pi/2a$ to $G/2 = \pi/2a$ is called the first Brillouin zone.

So far we have been considering a lattice that is a periodic array of atoms. But rather than just one atom on each lattice site, we could have something more complicated, like 2 atoms or a molecule or several atoms. This unit which is repeated periodically is called a "unit cell." A unit cell can be just one atom or several atoms. If the unit cell is a molecule which retains some measure of its individual identity in the solid, then we have a molecular solid. Each unit cell contributes exactly one independent value of k to each energy band. Taking into account the 2 spin orientations of the electron, there are $2N$ independent levels in each energy band, where N is the number of unit cells in the crystal.

Appendix: Another Way To Understand Reciprocal Lattice Vectors

Let $\vec{a}$ be a lattice vector that points from one lattice site to another. Translational symmetry implies that if we move by a distance a, the lattice will look the same. The electron wavefunctions obey this too. In other words we would expect the electron wavefunction $u_{\vec{k}}(\vec{r})$ to have the symmetry of the lattice:

$$u_{\vec{k}}(\vec{r}+\vec{a}) = u_{\vec{k}}(\vec{r}) \tag{5}$$

One can make things a bit more general by noting that we can multiply a wavefunction by a phase factor $\exp(i\vec{k}\cdot\vec{r})$ without affecting the physics. For example the electron probability $|\psi|^2$ is unchanged by the phase factor. Matrix elements $<\psi|\hat{A}|\psi>$ are also unaffected because $\exp(i\vec{k}\cdot\vec{r})$ gets multiplied by its complex conjugate $\exp(-i\vec{k}\cdot\vec{r})$. In fact Bloch's theorem states that in a periodic lattice the electron wavefunction $\psi_{\vec{k}}(\vec{r})$ must be of the form

$$\psi_{\vec{k}}(\vec{r}) = e^{i\vec{k}\cdot\vec{r}}u_{\vec{k}}(\vec{r}) \tag{6}$$

where $\vec{k}$ can be any wavevector (not just an allowed lattice wavevector) and $u_{\vec{k}}(\vec{r})$ has the periodicity of the lattice: $u_{\vec{k}}(\vec{r}+\vec{a}) = u_{\vec{k}}(\vec{r})$.

For simplicity, let's consider a one dimensional lattice of length $L = Na$ where N is the number of atoms in the lattice. As we saw in lecture 1 when we considered the case of periodic boundary conditions, we found that the allowed wavevectors were

$$k = \frac{2\pi n}{L} = \frac{2\pi n}{Na} \tag{7}$$

For example, if we have $N = 4$ sites, then $ka = 0,\ \pi/2,\ \pi,\ 3\pi/2,\ 2\pi$, etc. We would replace our graph of E vs. k by discrete states with these lattice vectors. Now the values of k where

$$k = \frac{2\pi n}{a} \tag{8}$$

are special. They are called reciprocal lattice vectors. To see why they are special, let G be a reciprocal lattice vector. Then

$$\exp(iGa) = \exp(i\frac{2\pi n}{a}a) = \exp(i2\pi n) = 1 \tag{9}$$

This just reflects the fact that any function which is characterized by a wavevector G has a periodicity that is in phase with the lattice, e.g., $\cos(Gr)$. If we multiply a Bloch wavefunction by $\exp(iGr)$, then we just get another Bloch wavefunction:

$$e^{iGr}\psi_k(r) = e^{i(k+G)r}u_k(r) = e^{ikr}u_{k+G}(r) \tag{10}$$

where $u_{k+G}(r) = \exp(iGr)u_k(r)$. You can show that $u_{k+G}(r) = u_{k+G}(r+a)$.

Just as the lattice looks the same if you translate by a lattice vector $\vec{a}$ or any of its multiples, so in $k-$space, things look the same if you translate by a reciprocal lattice vector G.

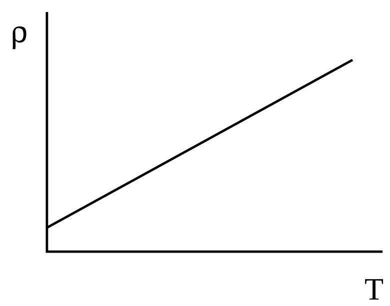
Metals and Insulators

Now that we have bands of energy levels, we can put electrons in these energy levels. If we have just the right number of electrons to completely fill one (or more) bands but not start a new band, then we have an insulator. (I am assuming that the empty band is separated by a gap from the filled band.) If we apply a small electric field, no current will flow because there are no easily accessible empty states for the electrons to jump into. At $T = 0$ the electrical resistance is infinite. Since each unit cell contributes one energy level which can hold 2 electrons to each band, if each unit cell contributes 2 valence electrons to a band, then the band is full and we have an insulator. Diamond is an example of an insulator. It has a band gap of 5.4 eV. There is another way to think about an insulator. If we look at the bonds between the carbon atoms in diamond, we see that they are covalent bonds. It is hard to get an electron to flow and carry current because it would have to break a covalent bond and that takes a large amount of energy.

Let's go back to the band picture. If each unit cell contributes 1 valence electron to a band, then the band will be half full, the Fermi energy will lie in the band, and the system will be metallic. A solid with a partially filled band is called a metal. In a metal electrons can flow and carry current because electrons in filled states below the Fermi energy can easily jump to empty states above the Fermi energy. The energy difference between the filled and empty states can easily be supplied by the applied electric field and thermal excitation.

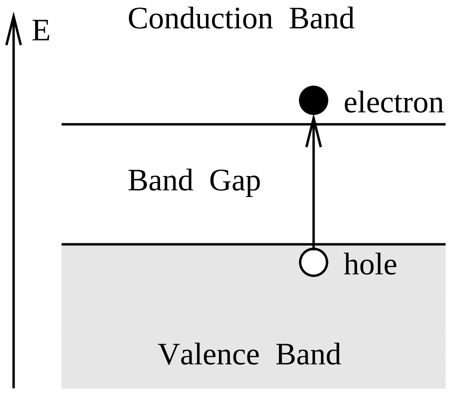
Experimentally the way to tell the difference between a metal and an insulator is by measuring electrical resistance. The difference between a good conductor and a good insulator is striking. The electrical resistivity of a pure metal may be as low as 10^{-10} ohm–cm at a temperature of 1 K (ignoring the possibility of superconductivity). The resistance of a good insulator may be as high as 10^{22} ohm–cm. This range of 10^{32} may be the widest of any common property of solids.

In a metal, the resistivity increases linearly with increasing temperature because the electrons scatter from phonons (lattice vibrations), and the number of phonons increases with temperature.



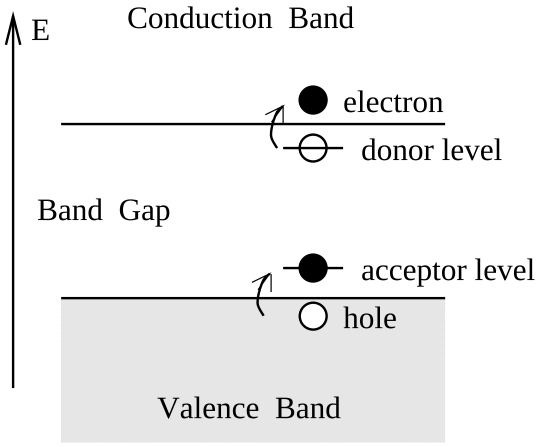
Semiconductors

Insulators whose band gaps are not too large are called semiconductors. In a semiconductor, a typical band gap is about 1 eV. Silicon has a band gap of 1.17 eV (indirect gap) and germanium has a band gap of 0.744 eV (indirect gap). There are also III-V semiconductors which are binary alloys consisting of one element from the third column of the periodic table and one element from the fifth column of the periodic table. For example, GaAs has a (direct) band gap of 1.52 eV. The band below the band gap is called the valence band and the band above the band gap is called the conduction band. As the temperature increases, the conductivity increases (and the resistivity decreases) because electrons are thermally excited from the valence band into the conduction band. The electrons in the conduction band are able to flow and carry current because there are easily accessible empty states that an electron can jump into. The electrons that make transitions into the conduction band from the valence band leave behind holes in the valence band. These holes act like positively charged carriers that are able to contribute to the electrical current.



Photons can also be used to excite electrons from the valence band into the conduction band. When electrons make transitions from the conduction band into the valence band and recombine with holes, photons can be given off. Semiconductor lasers take advantage of this.

Semiconductors whose primary source of carriers comes from the direct excitation of electrons from the valence band to the conduction band are called *intrinsic semiconductors*. Most of the electrical current carriers in *extrinsic semiconductors* come from impurities. These impurities produce states in the band gap which can supply electrons to the conduction band or holes to the valence band. Most electronic devices use extrinsic semiconductors that have been subjected to selective doping.



Donors are impurities which contribute levels that are just below the conduction band edge. They donate electrons to the conduction band which can contribute to electrical conduction. Donors have more valence electrons than the host. For example, arsenide (valence=5) is a donor impurity doped into the host semiconductor germanium (valence=4). Acceptors are impurities which have less valence electrons than the host, e.g., gallium (valence=3) doped into germanium (valence=4). Acceptors contribute impurity energy levels just above the valence band edge. They accept electrons from the valence band, which leaves holes in the valence band that can contribute to electrical conduction.

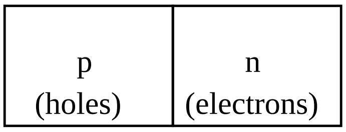
If a semiconductor has primarily donor impurities, we call it an n–type semiconductor because it has primarily negatively charged carriers. If a semiconductor has primarily acceptor impurities, we call it a p–type semiconductor because it has primarily positively charged carriers.

Semiconductor Devices

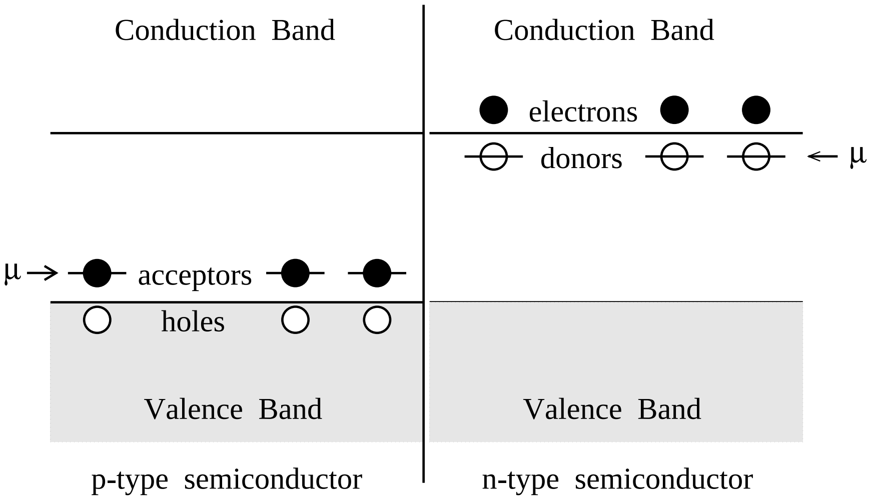
Of all the discoveries and inventions by physicists in the 20th century, the one with the most impact on technology and the economy is probably the transistor. A transistor is a current amplifier or regulator. The transistor was invented in 1948 by John Bardeen, William Shockley, and Walter Brattain at Bell Laboratories. For this they received the Nobel prize in 1956. I seem to remember that the number of transistors made each day is roughly equal to the number of calories consumed by all the people on the earth each day. That means about 1800 transistors are produced each day for every man, woman and child. In other words there are about 40 million transistors for each person on earth. Nowadays a typical chip has about 1 million transistors. (These are order of magnitude estimates.) The cpu of the G4 Mac computer has 56 million transistors.

pn Junction: Diode

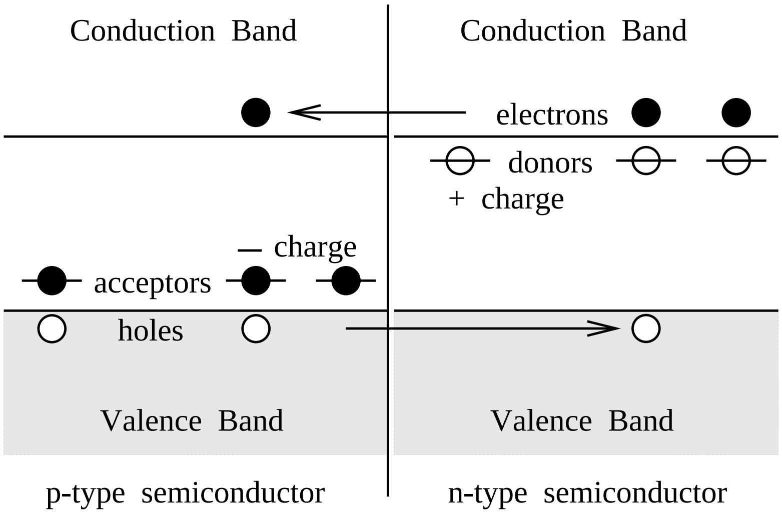
The basic element of solid state electronics is the pn junction, which is made by doping a semiconductor (say germanium) with donor and acceptor impurities in such a way that it is strongly n–type in one region and strongly p–type in another. The boundary layer is quite narrow, probably a few hundreds or thousands of angstroms, and for simplicity we replace it by an abrupt barrier.



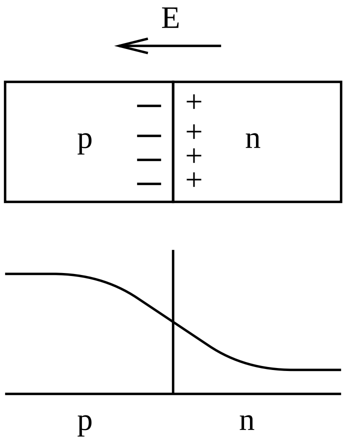
Let's suppose that initially the barrier between the p and n doped semiconductors is infinitely high. The chemical potential μ will be higher in the n–type semiconductor than in the p–type semiconductor.



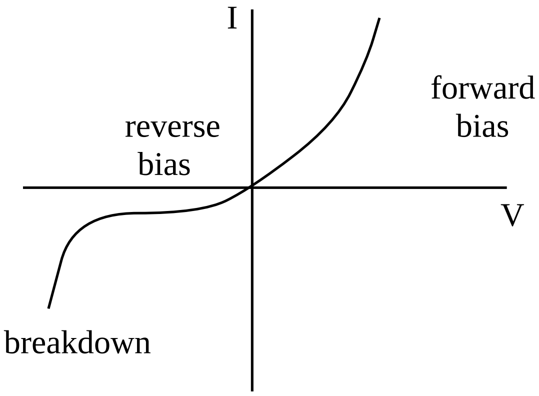
Now imagine that we remove the barrier. Electrons will flow over to the p–side, and holes to the n–side until the chemical potentials are the same.



As soon as a small charge transfer by diffusion has taken place, there is left behind on the p–side an excess of – ionized acceptor atoms and on the n–side an excess of + ionized donor atoms. This double layer of charge creates an electric field directed from n to p that inhibits further diffusion and maintains the separation of the two carrier types. We can draw the potential seen by the electrons. The potential drop will be at the interface because the ionized donors and acceptors attract each other.

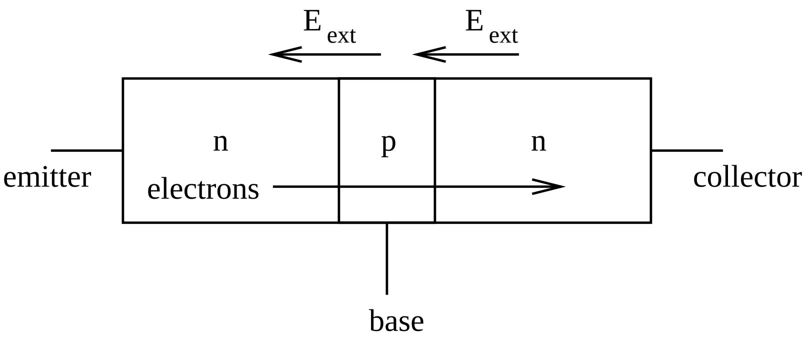


Electrons would rather go downhill than uphill. If we apply a voltage across the junction that increases the size of the drop, we encourage electrons to flow from the p–side to the n–side. This is called reverse bias. But they already want to do this, so it doesn't make much difference in the current. If we really crank up the reverse bias, we get what is called "breakdown" and electrons avalanche from the p–side to the n–side. If we apply voltage in the other direction, the electrons are less reluctant to go from the n–side to the p–side. (the conductivity $\sigma \sim \exp(-V/kT)$ where V is the barrier height.) This is called forward bias. This asymmetry in the preference of the direction of the current is how a diode works. A diode allows current to go one way but not the other way. Increasing current flows as the forward bias increases but not much current flows when reverse bias is applied.

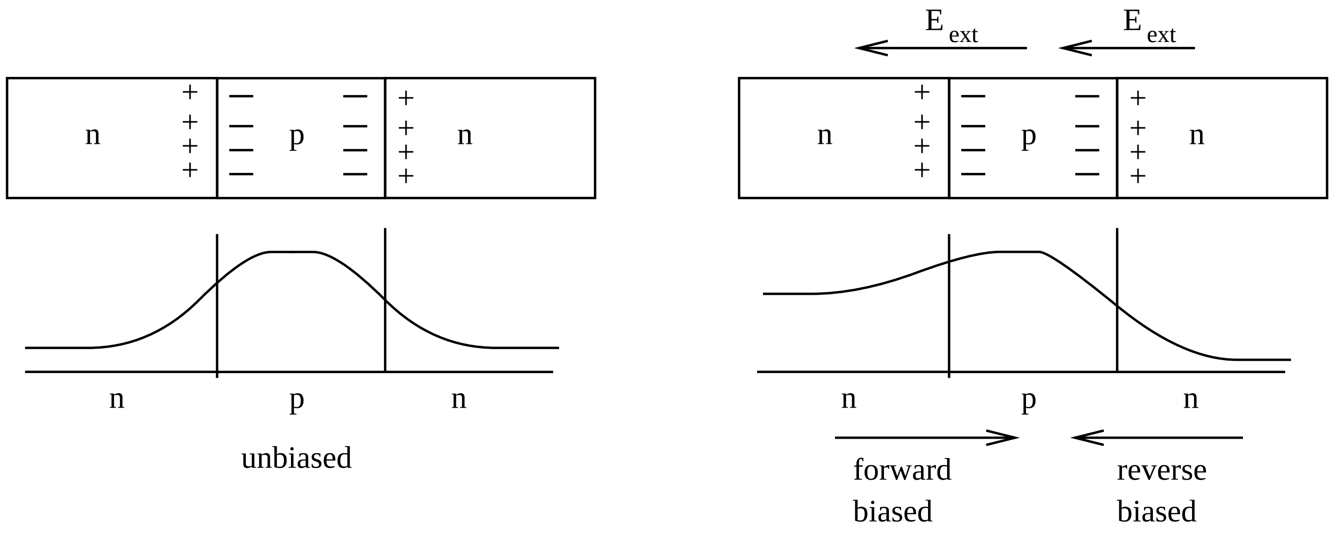


Bipolar Transistors

Since we now understand how pn junctions work, we can understand schematically how bipolar transistors work. An n–p–n type transistor consists of 2 pn junctions. A small p–type region is sandwiched between two n–type regions and connections are made to all three regions. The terminals are labelled emitter, base, and collector.

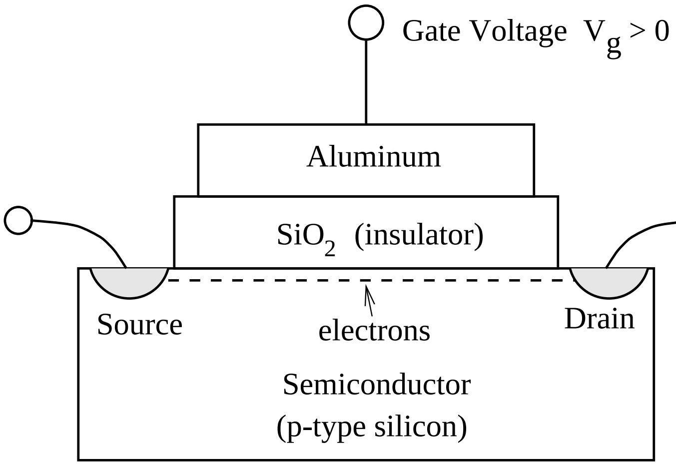


A bipolar transistor is a current amplifier. In normal operation the emitter to base junction is forward biased, and the collector to base junction is reverse biased. Consider the electrons coming into the base from the emitter due to the forward biased emitter–base junction. For a thin enough base section, these carriers sweep through the base layer, cross the base–collector junction, and contribute to the collector current. The essential action is the emitting of carriers from the emitter region and the collection of practically all of these carriers by the collector. Let's denote this current I_{ec}. A small hole current from the base region also flows across the emitter junction. We will denote this hole current I_{be}. This adds to the electron current from the emitter to the base. By proper design of the impurity concentrations and base layer width, the ratio I_{ec}/I_{be} can be made very large ($\gtrsim 100$). If the input current is taken to be the small hole current I_{be}, and the output current is taken to be the large emitter–collector current I_{ec}, a significant current gain is thus achieved.



MOSFET

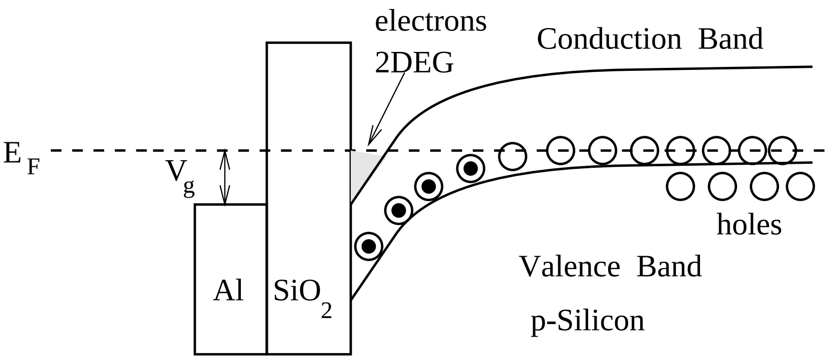
As an example of a semiconducting device, let's look at a MOSFET which stands for Metal Oxide Semiconducting Field Effect Transistor.



MOSFET

The positive gate voltage attracts electrons to the interface. By adjusting the magnitude of the gate voltage V_g, we can adjust the charge density at the interface between the semiconductor and the insulator. This is like a capacitor where $Q = CV_g$. The current flows between the source and the drain. Since current is $I = dQ/dt$, we can adjust the amount of current by adjusting the gate voltage.

The potential seen by the electrons is lower near the interface between the semiconductor and the oxide layer than deep inside the semiconductor. We can describe this lower potential by "band bending."



In this figure it is assumed that the semiconductor is p–type, i.e., some electrons of the valence band have become bound to acceptor impurities leaving empty states or holes. The lowest energy holes are at the top of the valence band. This means that the Fermi energy E_F is close to the top of the valence band. The electrons attracted to the interface first fill up these hole states leaving a net negative charge near the interface. However, if the gate voltage V_g is large enough, the bottom of the conduction band will become lower than E_F. This is called the inversion layer since the bottom of the conduction band is below the top of the valence band, inverting the order. Electrons will occupy states in the part of the conduction band below the Fermi level. These electrons at the interface form a 2 dimensional electron gas (2DEG). These are the electrons which carry current from the source to the drain.

(When one puts this 2DEG in a large magnetic field perpendicular to the plane of the interface, one gets the quantum Hall effect. The discovery of the integer quantum Hall effect by Klaus von Klitzing won him the Nobel prize in 1985. The 1998 Nobel prize was for the fractional quantum Hall effect which was discovered by Daniel Tsui and Horst Stormer and explained by Robert Laughlin. In the fractional quantum Hall effect the 2DEG becomes a quantum fluid with fractionally charged excitations which have charges like $e/3$. The smaller the fraction, the larger the applied field.)