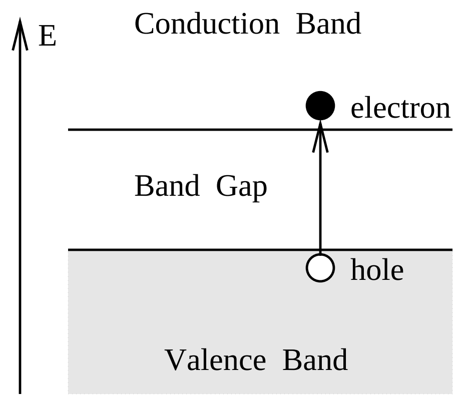
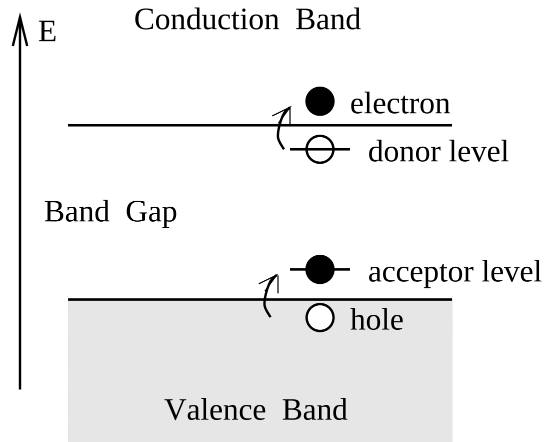
Semiconductors

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.

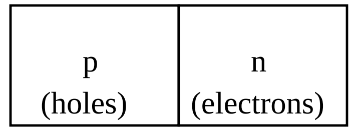
If a semiconductor has primarily donor impurities, we call it an n–type semiconductor because it has primarily negatively charged carriers. If a semiconductor has primarily acceptor impurities, we call it a p–type semiconductor because it has primarily positively charged carriers.

Semiconductor Devices

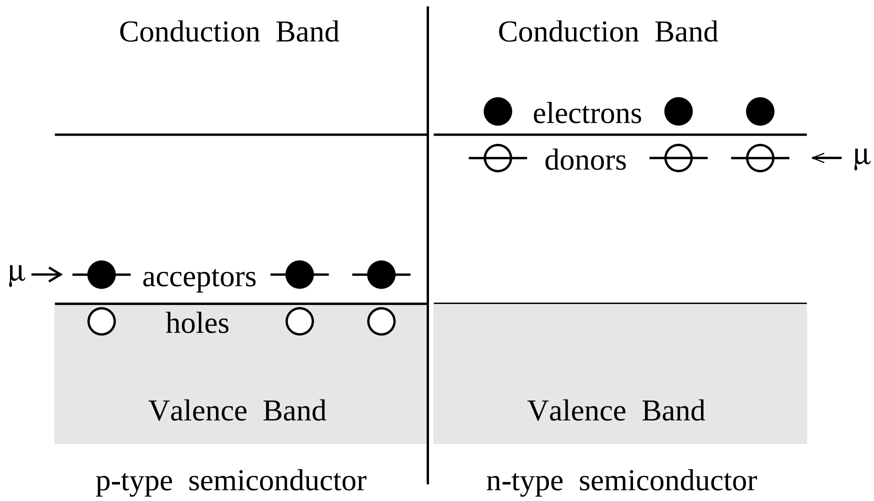
Of all the discoveries and inventions by physicists in the 20th century, the one with the most impact on technology and the economy is probably the transistor. A transistor is a current amplifier or regulator. The transistor was invented in 1948 by John Bardeen, William Shockley, and Walter Brattain 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

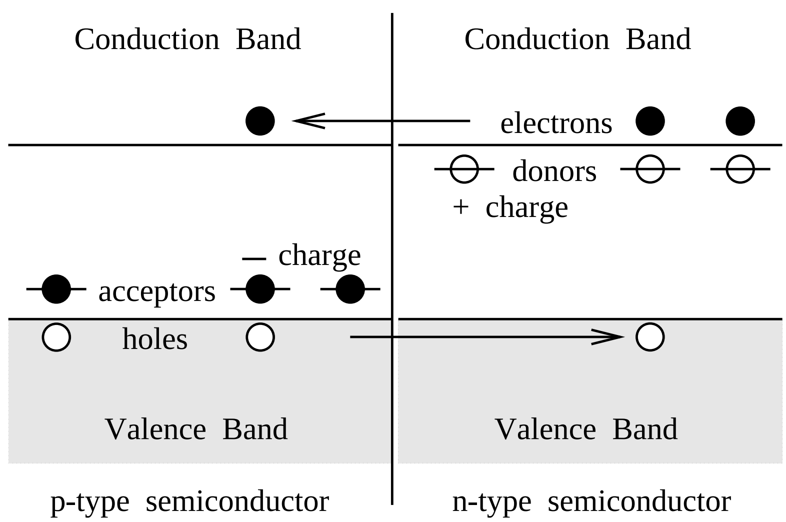
The basic element of solid state electronics is the pn junction, which is made by doping a semiconductor (say germanium) with donor and acceptor impurities in such a way that it is strongly n–type in one region and strongly p–type in another. The boundary layer is quite narrow, probably a few hundreds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



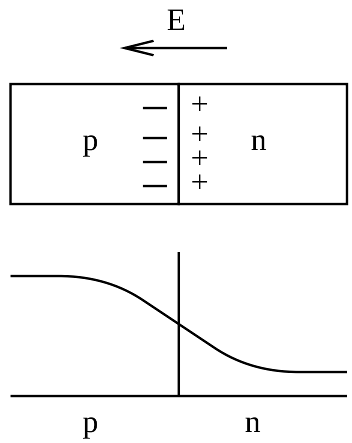
Let's suppose that initially the barrier between the p and n doped semiconductors is infinitely high. The chemical potential μ will be higher in the n–type semiconductor than in the p–type semiconductor.



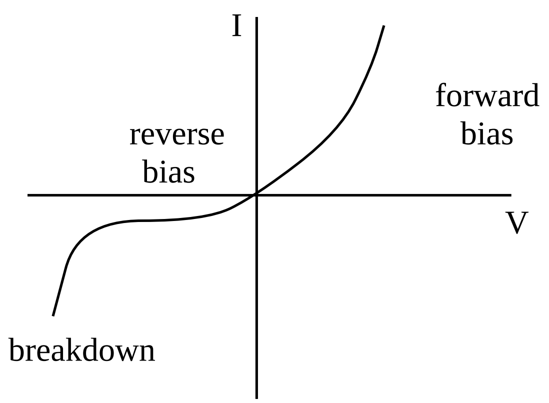
Now imagine that we remove the barrier. Electrons will flow over to the p–side, and holes to the n–side until the chemical potentials are the same.



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.

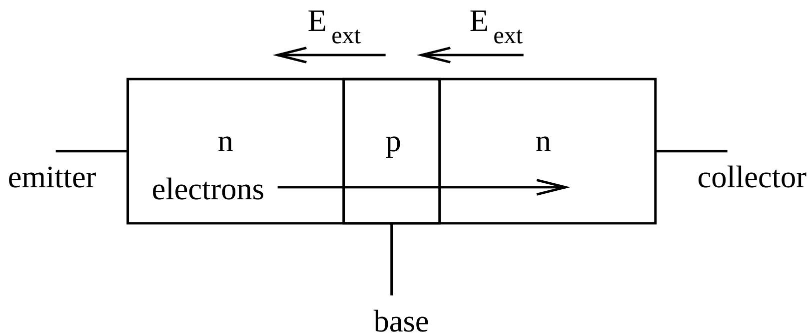


Electrons would rather go downhill than uphill. If we apply a voltage across the junction that increases the size of the drop, we encourage electrons to flow from the p–side to the n–side. This is called reverse bias. But they already want to do this, so it doesn't make much difference in the current. 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.

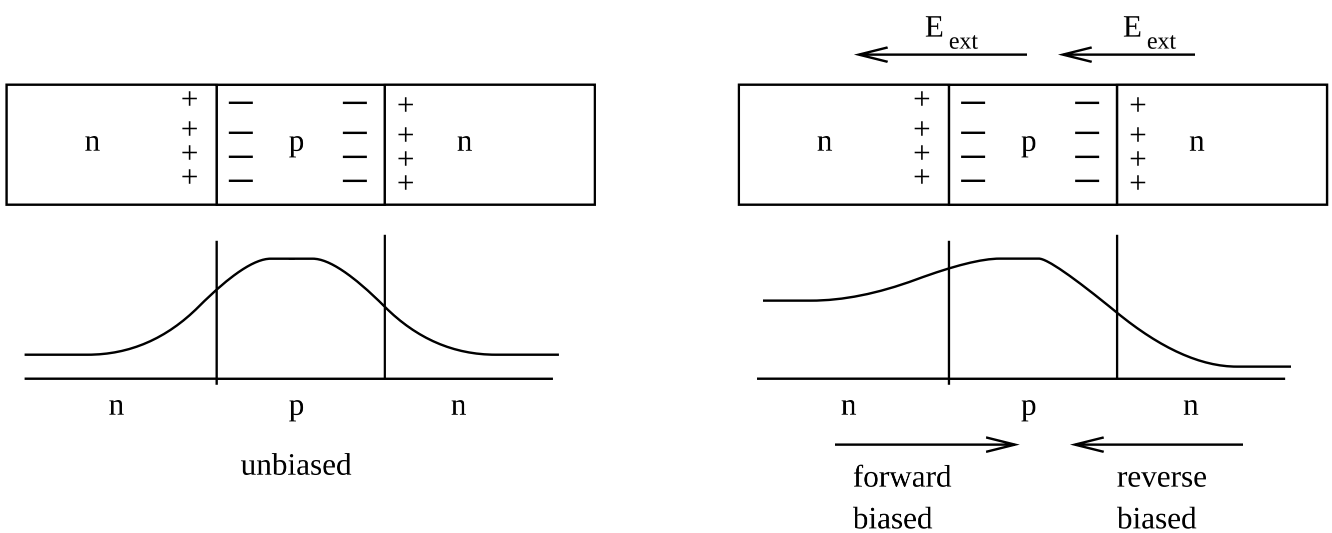


Bipolar Transistors

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.

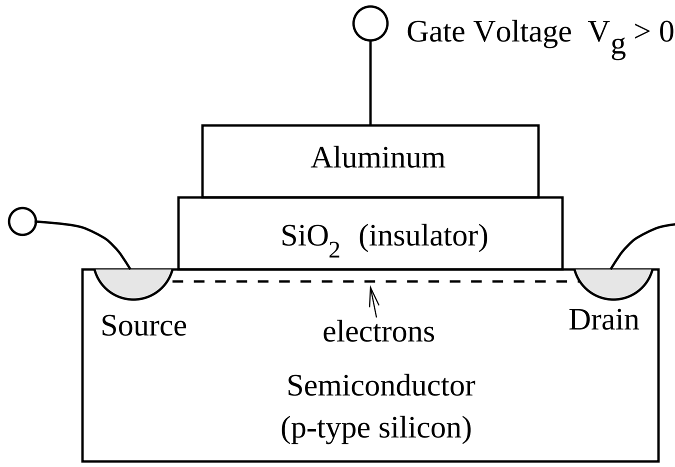


A bipolar transistor is a current amplifier. In normal operation the emitter to base junction is forward biased, and the collector to base junction is reverse biased. Consider the electrons coming into the base from the emitter due to the forward biased emitter–base junction. For a thin enough base section, 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 ($\lesssim 100$). 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

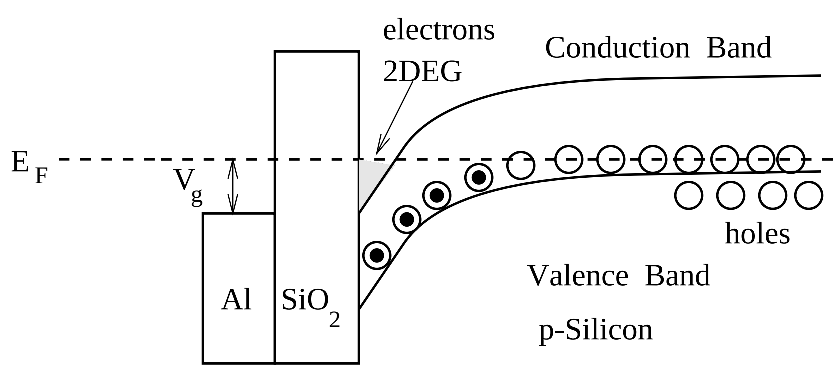
As an example of a semiconducting device, let's look at a MOSFET which stands for Metal Oxide Semiconducting Field Effect Transistor.



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. We will talk more about this later.