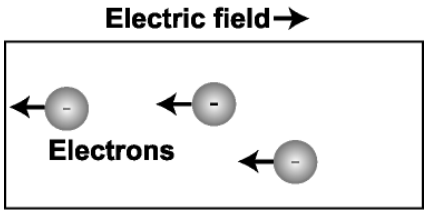
Theory of Transistors and Other Semiconductor Devices

1. SEMICONDUCTORS

1.1. Metals and insulators

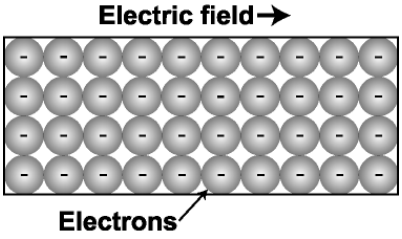
1.1.1. Conduction in metals

Metals are filled with electrons. Many of these, typically one or two per atom in the metal, are free to move about throughout the metal. When an electric field is applied, the electrons move in the direction *opposite* the field. Since they are negatively charged, this corresponds to a positive current in the direction *opposite* the motion, that is, in the direction of the electric field:



1.1.2. Insulators

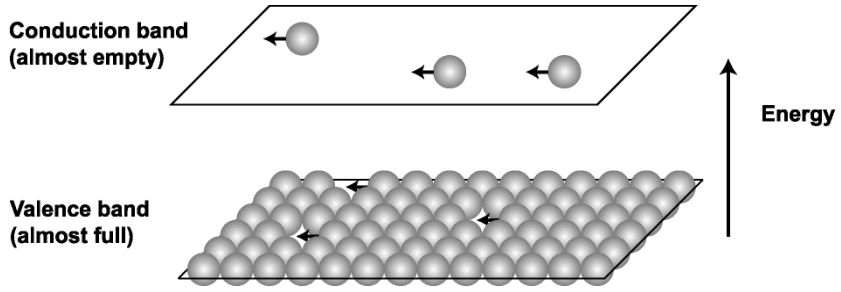
Insulators also have many electrons in them, but the electrons cannot move. Some of them are trapped in individual atoms and can't get away from them. Others are nominally free to move about but are locked in place by "gridlock." There are so many electrons that there is no place for an electron to move to, so it stays put even in an electric field:



1.2. Semiconductors

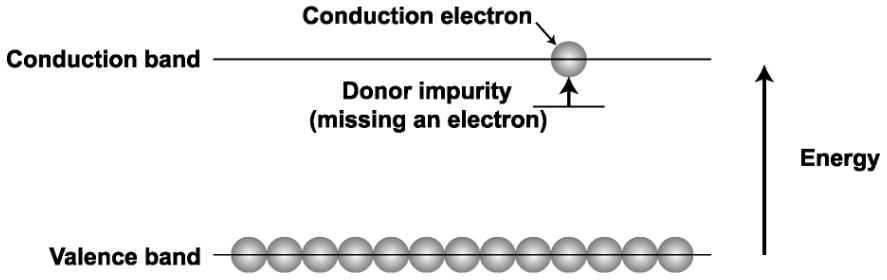
1.2.1. Intrinsic semiconductors

Not surprisingly, semiconductors are somewhat like metals and somewhat like insulators. As in atoms, the electrons in solids are in certain energy levels, or, more correctly, certain bands of energy levels, so this is called the band theory of solids. In metals, the uppermost band has a few electrons in it and these can move around, as in the first figure, above. In insulators the uppermost band is completely filled with electrons and they are gridlocked, as in the second figure , above. In semiconductors, the next-to the highest band is *almost* completely filled with electrons and they are *almost* gridlocked. However, the highest band is only a little higher than the next-to-highest, and a few electrons jump into this band and conduct a little electricity. In addition, the holes left in the band below give a little room for the gridlocked electrons to move. Actually, the holes look like positive charges (due to the positive ions left behind when the electrons move away) and these holes appear to move in the direction opposite the electrons. Therefore, the current carried by a semiconductor consists of the motion of the negatively-charged electrons in the conduction band (the uppermost band) and the positively-charged holes in the valence band (the next-to highest band).

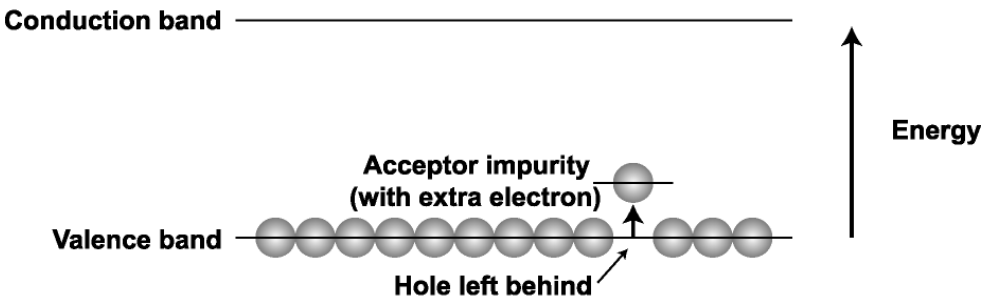


1.2.2. Doped semiconductors

The electrical conduction in semiconductors can be improved by doping the semiconductor with a small amount of some impurity. In a so-called n-type semiconductor, atoms are added which contribute a few extra electrons to the conduction band:



In a so-called p-type semiconductor, atoms are added that "suck up" a few electrons from the valence band, leaving behind holes in the valence band.



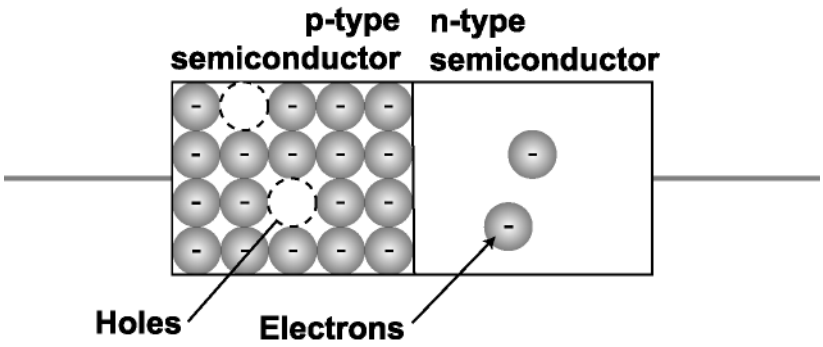
Thus, the dopant (the impurity added to the semiconductor) in a p-type semiconductor adds positive charge carriers (holes in the valence band), while the dopant in an n-type semiconductor adds negative charge carriers (electrons in the conduction band).

2. SEMICONDUCTOR JUNCTIONS AND DIODES

2.1. p-n junction

2.1.1. No electric field

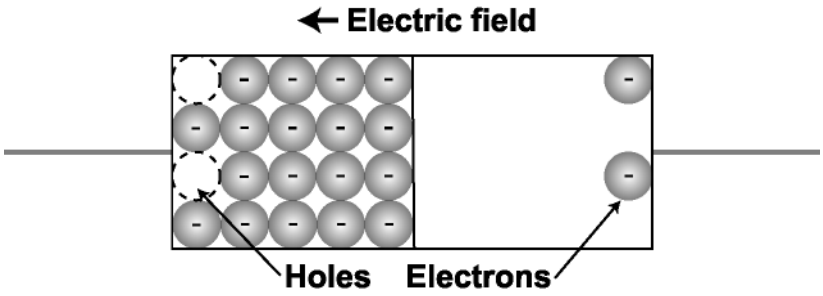
A p-n junction consists of a p-type semiconductor next to an n-type semiconductor:



When wires are connected to the ends, as shown, the device becomes a semiconductor diode.

2.1.2. Reverse bias

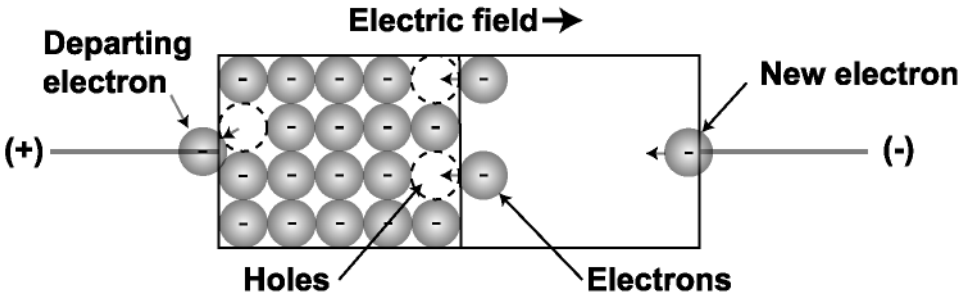
When an electric field is applied in the so-called reverse direction, the positive charges on the one side are pulled away from the junction and the negative charges on the other side are pulled the other way away from the junction.



The charges move apart until the electric field across the junction gets too strong, and the charges stop moving. A voltage applied in this way is called a bias.

2.1.3. Forward bias

When an electric field is applied in the forward direction, the opposite happens. The positive charge carriers (holes) move toward the junction from their side and the negative charge carriers (electrons) move toward the junction from their side. But a hole is nothing more than an atom missing one of its electrons, so when a hole reaches the junction at the same time as an electron, the electron falls into the hole and both disappear! This is called recombination.

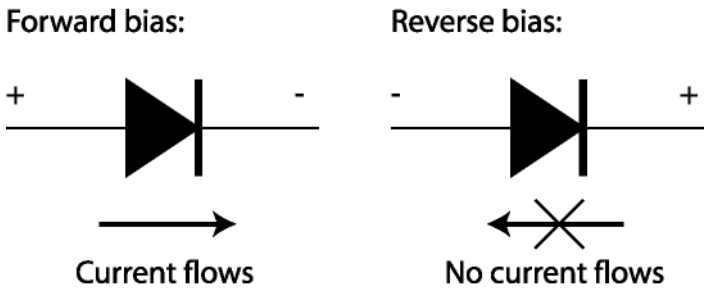


When recombination occurs, there is room for another electron and hole to move into the same place and the process continues as long as the forward bias is maintained. Of course, as the charge carriers move toward the junction and disappear, they are replaced at the ends by charges moving into and out of the wires.

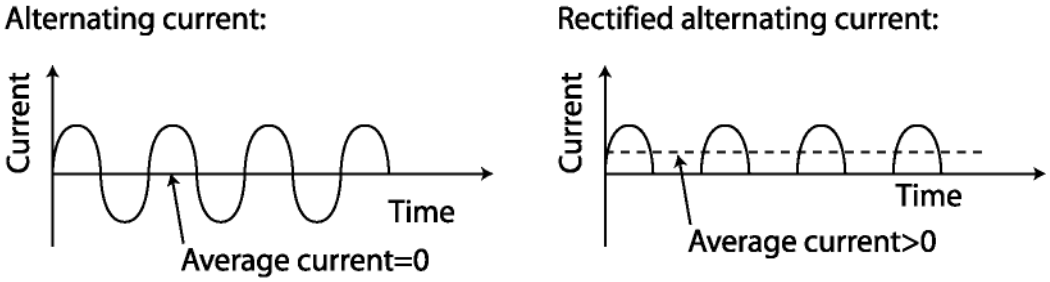
2.2. Diode rectifier

2.2.1. Rectification

Diodes have the property that they pass current in only one direction:



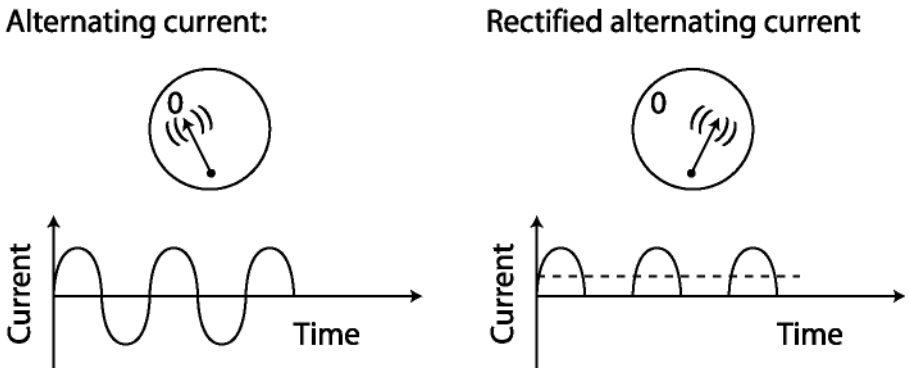
That is, they can be used to *rectify* an alternating current:



The result is not "pure" direct current, but "pulsed" direct current. However, this is good enough since there are many ways to smooth out the result.

2.2.2. Detection

When an ac signal oscillates too fast, it is impossible for a slow device like a voltmeter to detect the voltage. The meter just hovers around zero. But if the voltage is rectified, then the meter can respond to the average of the rectified voltage. This is called "detection" of the signal:

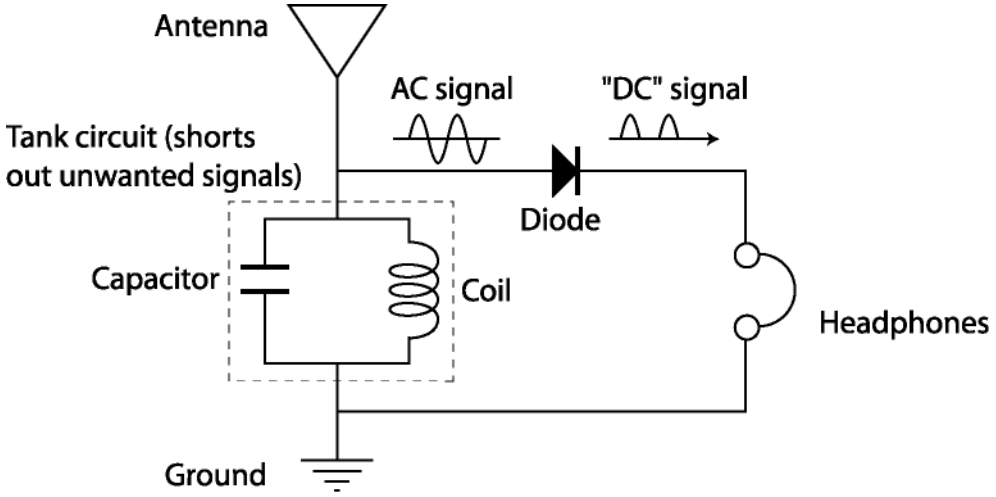


2.3. Detection of a radio signal

2.3.1. Tuning circuit

The radio-frequency signal from an antenna is an alternating current. In a radio, the speaker, or headphones, cannot keep up with the rf signal, but after it is detected the speaker can respond.

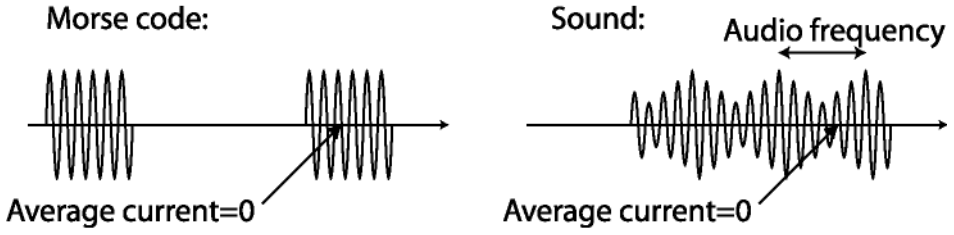
The signal from the antenna in a radio must first be tuned in to separate it from other radio signals. This is done with a coil and a capacitor in a "tank circuit" (don't ask me why it is called a tank circuit). The tank circuit shorts out signals at the "wrong" frequencies, but not at the desired frequency. This signal is then detected (rectified) by the diode so the headphones can respond.



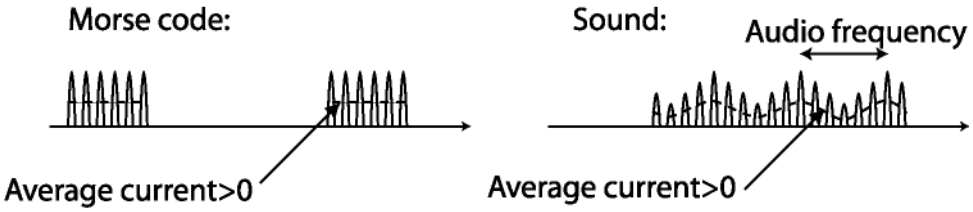
In this radio, which has no batteries or other source of energy, all the power to drive the headphones comes from the antenna. The power you hear comes out of the air!

2.3.2. Diode detector

When the radio is used to transmit Morse code, the signal from the transmitter consists of pulses of the "carrier" frequency of the radio station. When the radio is used to transmit sounds, the signal from the transmitter consists of pulses at the carrier frequency modulated at the audio frequency:



After detection (and smoothing), the signal consists of pulses at the (much lower) audio frequency, to which the headphones can respond:

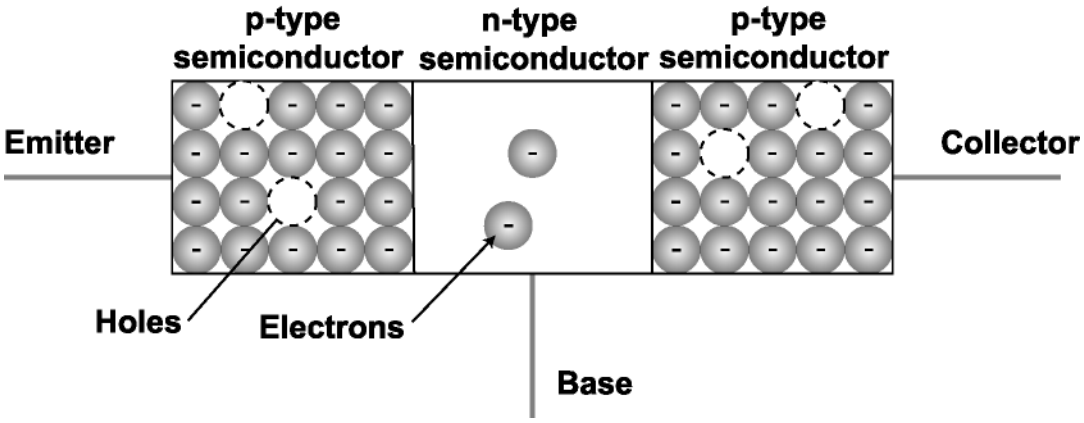


3. TRANSISTORS AND TRANSISTOR CIRCUITS

3.1. Double junction

3.1.1. Emitter, collector and base

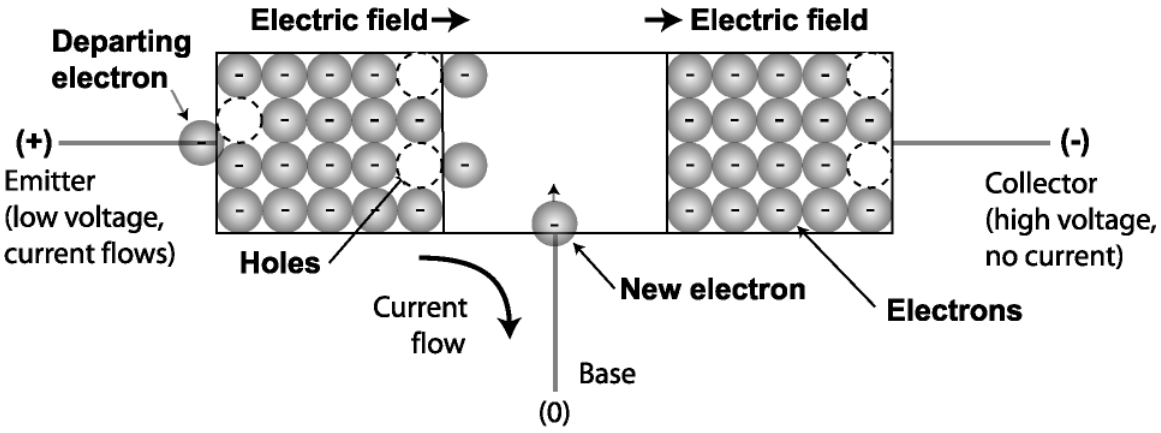
In its essence, a transistor consists of two diodes arranged back to back:



Transistors can be either n-p-n or p-n-p. The configuration shown above is called an n-p-n transistor. The terminal on the left is called the emitter, the terminal on the right is called the collector, and the region in the middle is called the base.

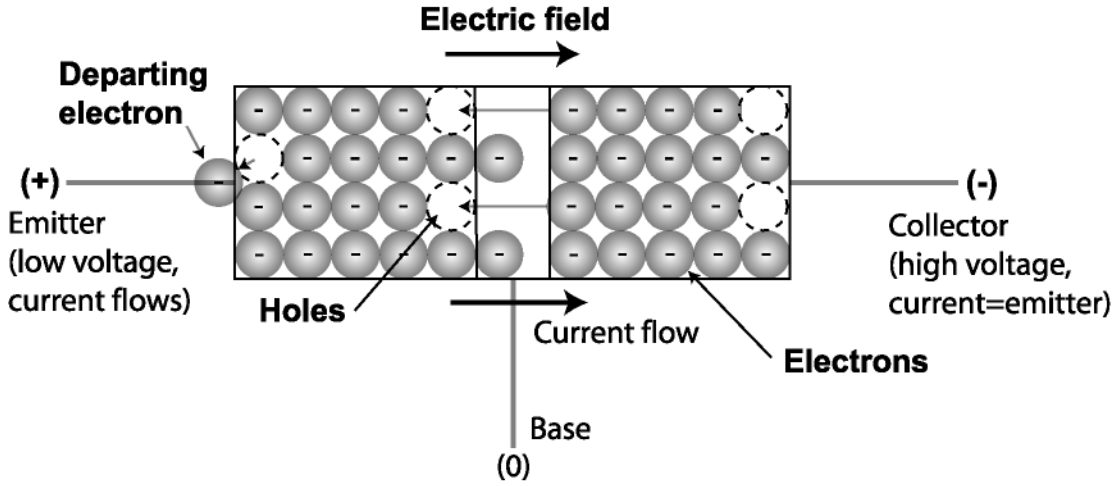
3.1.2. Biasing collector and emitter

In operation, the collector is reverse biased, relative to the base, so no current flows, even if the voltage is large. On the other hand, the emitter can be forward biased, in which case a current flows even for a relatively small voltage.



3.1.3. Narrow base region; transistor

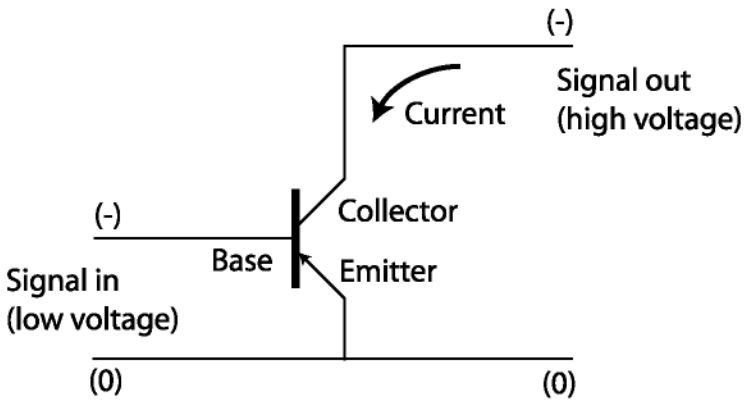
What makes the transistor interesting is that the base region can be made very thin. When this is done, the negative charges that flow into the base region from the emitter are attracted to the collector before they can flow out the base connection. In this way, the voltage on the emitter, compared with the base, controls the current in the collector. This allows one connection to control the current another connection.



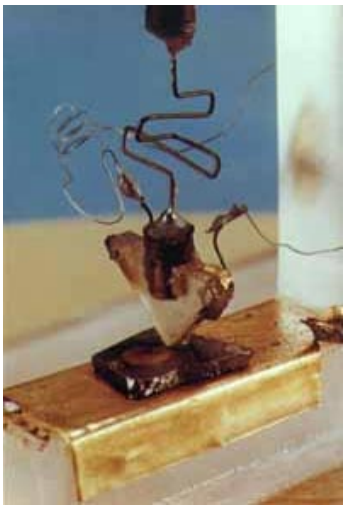
3.2. Amplification

3.2.1. Grounded emitter amplifier

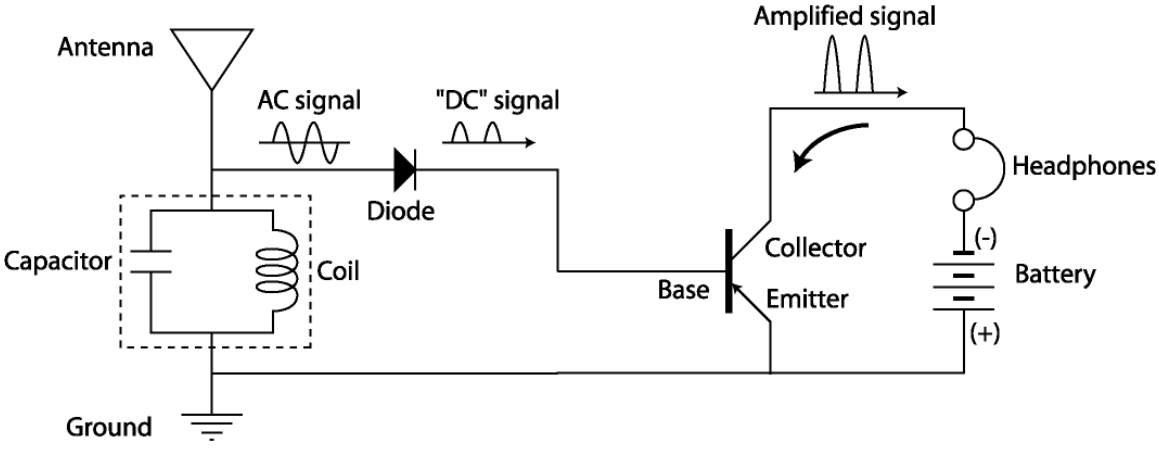
As it would be drawn by an electrical engineer, the simplest amplifier circuit using a transistor is



The funny symbol for a transistor comes from the fact that the first transistor didn't look like the block structure shown above. Rather, it was fabricated by taking a flat piece of germanium (the base) and attaching the emitter and base connections to the surface right near each other. As the metals from the wires diffused slightly into the surface they formed the p-doped regions separated by a very small distance.



This amplifier circuit can be used to improve the crystal radio:

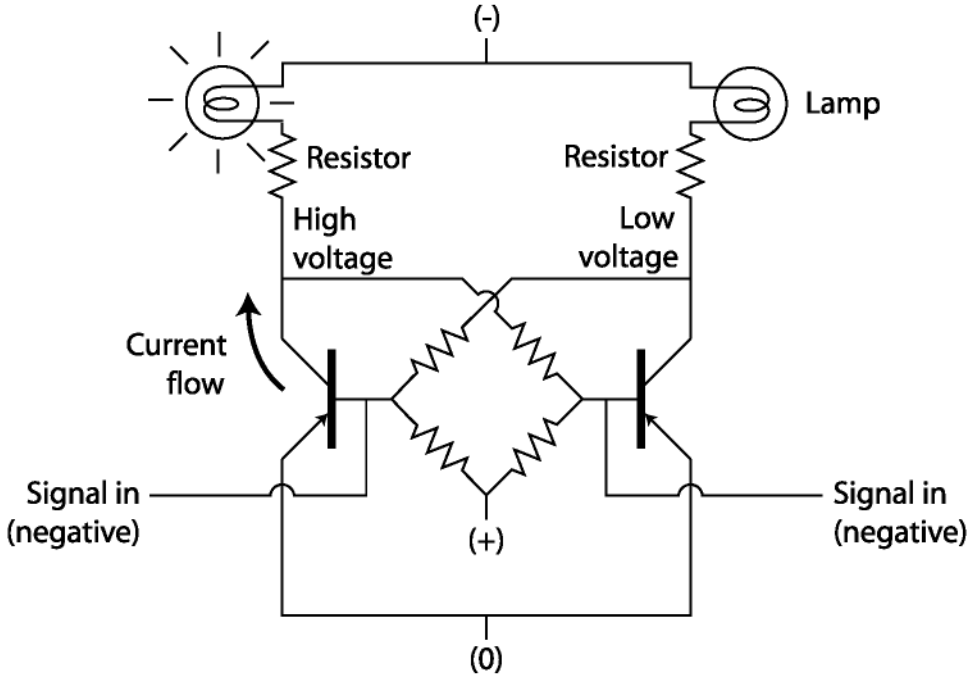


In this circuit the power for the headphones comes from the battery, which has much more power than can be collected by the antenna. By the way, the funny symbol for the battery represents the stack of copper and zinc disks and moistened cardboard that Volta use to make the first battery in 1800.

3.3. Flip-flop and computer circuits

3.3.1. Flip flop circuit

The basic computer circuit is called the flip-flop. It can exist in either of two states. As shown, the output from the left-hand transistor (the collector voltage) is used as the input to the right side (the base current), and vice versa. When the current is flowing in the left-hand transistor, the voltage at the collector is high (but negative) due to the voltage drop in the resistor. When this high voltage is applied to the base of the right-hand transistor the collector current in that transistor is small. In fact, it can be essentially shut off. In this case the voltage at the collector on the right-hand side is low (that is, more negative), since there is no voltage drop across the resistor. This low voltage is applied to the base of the left-hand transistor, which causes the left-hand transistor to turn on, as we assumed in the first place. Thus, the circuit is stable in this configuration.



But the circuit shown above is symmetric, so either the right-hand light or the left-hand light can be "on." In fact, the state of the system can be flipped. If a signal is applied in the right place, the circuit will flip. Specifically, if a negative voltage is applied to the base of the transistor that is not conducting, then it will begin to conduct and turn the other transistor off. Thus, the circuit "remembers" the last input it received.

3.3.2. Computer memories

Computers use millions (now billions) of these circuits to remember the information that is input. However, everything in the memory must be represented by 0's and 1's, since flip-flops only have two states, corresponding to 0 or 1. This is why computers use the binary number system.