Chapter 4: Diodes and Transistors

A. Semiconductors and Diodes

I will not try to explain the workings of diodes and transistors in any detail; just give you a brief introduction to their behavior. Both diodes and transistors are based on the behavior of semiconductors. These are materials where the outer electrons are localized in bonds and not free to move around at low temperatures. As a result, the materials are insulators if they are very pure, or intrinsic. Silicon and Germanium are semiconductors, as is Gallium Arsenide. Even at room temperature, the thermal energy (KE $\approx k_B T$, and kT $\approx$ 0.026eV at room temperature or 300K) is much smaller than the energy necessary to free an electron from its bonding orbital. However these materials may be doped with impurities that have an "extra" electron, i.e. one more than is needed for the bonding structure. These impurities are called donors. A typical donor atom for silicon is phosphorus. It requires about 1.12eV of energy to break an electron out of a Si-Si bond, but the extra electron in phosphorus needs only 0.044eV to break free and go into the conduction band. Therefore, at room temperature many of the extra electrons from the P atoms are "free" and able to conduct electricity. (As you increase the temperature, a greater fraction will break free.) Silicon doped with phosphorus is called n-type material, because the primary or majority charge carriers are negative electrons. The typical concentrations of donors might be 10^{16} to 10^{19}per cm^3, while pure silicon has about $5x10^{22}$ atoms per cm^3. As a result, only about 1 in 10^4 or 10^6 atoms is a donor. (They donate electrons to the conduction band.)

You can also dope the silicon with an atom, e.g. boron, that has three outer electrons, one short of the four needed to form bonds with all the four adjacent silicon atoms. However, the boron atom can acquire an electron from another silicon atom so that the boron can form bonds with the four neighboring silicon atoms. The electron the boron atom has acquired is at a slightly higher energy level, it takes about 0.045eV of extra energy to move the electron from a Si atom to the B atom, i.e. a silicon atom that is not adjacent to the boron atom. At low temperatures the thermal energy is too small for this to be very probable, but at room temperature a noticeable fraction of the B atoms will have acquired an extra electron. This means that a Si atom is missing an electron and a Si-Si bond is not complete. This incomplete bond is called a hole. (The usual terminology is that an electron from the valance band goes to the boron atom, leaving a hole in the valance band.) The interesting thing is that this hole can move around the crystal and acts like a positive charge moving through the crystal. This type of material is called p-type, because the primary charge carriers are the positive holes. Again only a small fraction of the atoms are acceptors.

The interesting effects occur when you dope a slab of silicon so that one side is n-type and the other is p-type. They form a pn junction where they meet. At this junction, holes diffuse from the p-type material to the n-type where the hole soon are 'filled' with an electron. Similarly the electrons from the n-type material diffuse across to the p-type material where they soon 'fall' into a hole and are trapped. This usually occurs within a region very close to the junction, perhaps a couple microns or less, depending on the doping concentrations. This junction region is actually depleted of charge carriers since the holes from the p-type material have

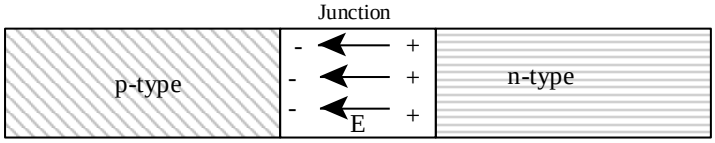


Fig. 4.1 A pn Junction

combined with electrons and visa versa, effectively making the free electrons and holes disappear. As a result this region is almost an insulator. (I've greatly exaggerated the size of the junction in fig. 4.1.)

If electrons have migrated to the p-type material, they will make it more negative and leave the n-type material positive. Holes migrating from the p-type to the n-type material leave the p-type material more negative and make the n-type more positive. This migration leaves the n-type material slightly positive and the p-type slightly negative, which produces a potential difference, or contact potential difference, between the two materials. This is usually 500 to 700mV in silicon. (It depends on temperature and dopant concentrations.) Thus there is an electric field in the depletion region pointing from the n-type to the p-type. You can't measure this if you connect a voltmeter because there are other contact potentials between the wires of the meter and the p & n-type materials. If they are all at the same temperature, these contact potentials cancel.

What happens if you connect the two slabs to a battery and try to drive current through them? It depends on how you connect the battery and on the battery voltage. If the + terminal is connected to the n-type material and the – terminal is connected to the p-type, very little current flows. This configuration tends to pull the electrons in the n-type material toward the + terminal of the battery and away from the junction and pulls the holes in the p-type material toward the – terminal of the battery and away from the junction. This does not push more charge carriers into the depletion region at the junction; rather it pulls them away from it. As a result, there are very few charge carriers in this region to carry a current through the junction, so very little current flows in the circuit.

If you reverse the connections, i.e. connect the – terminal to the n-type and the + terminal to the p-type, the holes in the p-type material are pushed toward the depletion region and the electrons in the n-type material are also pushed toward the depletion region. However, the electric field in the depletion region tends to keep the holes from the p-type region from crossing over to the n-type and tends to keep the electrons from the n-type region from crossing over to the p-type. The battery has to apply enough voltage to the system to overcome this field, which is due to the 600 to 700mV potential difference. If it can, then holes will flow from the p-type to the n-type and visa-versa. This means a current can flow, but only in this direction. Such a device is called a diode. The positive current flows from the p to the n type material.

The symbol for a diode is shown at the right. The left one is the symbol you would see in a circuit diagram and the right one is what an actual diode might look like. The band is on the n-type end on both of them.

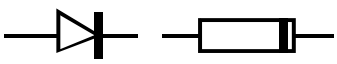


Fig. 4.2 Diodes, Symbol and Actual

If you connect a voltage source to a diode as in fig. 4.3 and measure the current as a function of voltage it will look like the plot in fig. 4.4. The current is an exponential function of the voltage and looks something like

$$I(V) = I_o\left(\exp\left\{\frac{\eta e V}{k_B T}\right\} - 1\right) \qquad 4.1$$

where e =1.6x10^{-19}C, V is the voltage and is positive if the p-type material is at a higher potential than the n-type, k_B is Boltzmann's constant, T is the absolute temperature in Kelvin, I_o is a 'constant' that depends on T but not on V, and η is a constant. When the diode is biased so that current flows, we say it is forward biased. Fig. 4.3 shows a forward biased diode. **I do not recommend connecting a diode in this fashion**. **You should put a resistor in series with the diode to limit the current flow**.

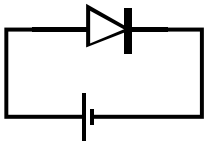


Fig. 4.3

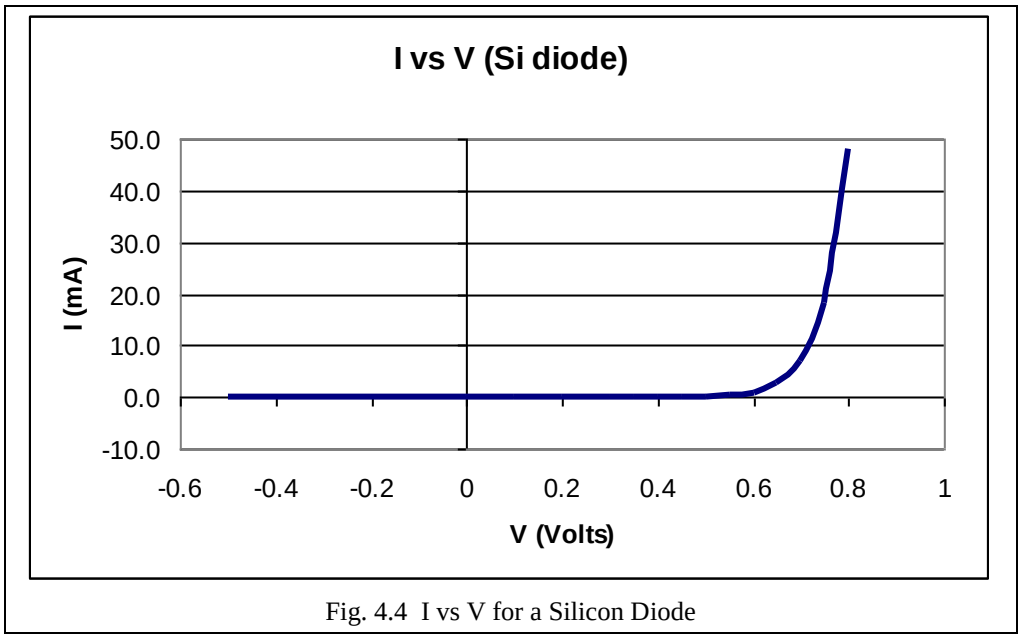


Fig. 4.4 I vs V for a Silicon Diode

In fig. 4.4, the current is small on this scale until the forward voltage, V_F, reaches about 0.6V and then starts to increase rapidly. As a result we often **approximate** the behavior of a silicon diode by saying that if the forward voltage is < 0.6V, the diode is off and the current is 0. If the voltage starts to rise above 0.6V the diode turns on and it will conduct as much current as necessary to keep the forward voltage from rising above 0.6V. This level of approximation is sufficient for most of the work in this course.

Consider the circuit at the right, fig. 4.5, for two cases. Note that the arrow through the battery means that the voltage can be varied.

1. $V_F < 0.6V$.
2. $V_F > 0.6V$.

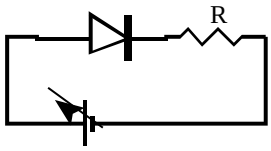


Fig. 4.5

In the **first case**, the diode is off and no current flows. Then the voltage across the resistor is 0, i.e. IR = 0 since I = 0. The entire voltage drop is across the diode, $V_F = V_{battery}$. In the **second case**, there is current flowing and $V_F = 0.6V$. This means that the rest of the battery's voltage must be dropped across the resistor, so $V_R = V_{battery} - 0.6V$. Therefore the current I is given by

$$I = \frac{V_{battery} - 0.6V}{R} \qquad 4.2$$

If R = 1k and $V_{battery}$ = 2V, V_R = 1.4V and I = 1.4mA. If $V_{battery}$ = – 2V, V_R = 0 and I = 0.

A typical use for diodes is rectification. In rectification you have an input voltage that varies in time, e.g. a sinusoidal voltage, but want an output that only has one polarity of voltage,

either positive or negative, but not both. Consider the circuit below where the source voltage varies between +3V and – 3V as shown in fig. 4.7. The output voltage, i.e. the voltage across the resistor is also shown in fig. 4.7. Note that if the input is < 0.6, $V_{out} = 0$ and if $V_{in} > 0.6$, $V_{out} = V_{in} - 0.6V$. If you reversed the direction of the diode, the negative part of the signal would come through.

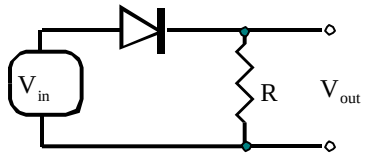


Fig. 4.6 Rectifier Circuit

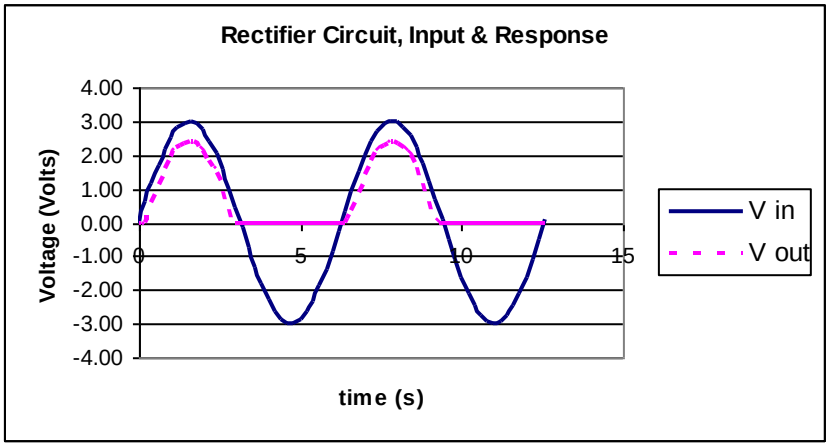


Fig. 4.7 Response of Rectifier Circuit

Diodes are often used in power supplies to convert the AC signal from a transformer to a DC signal. There are a couple of possible configurations, but the most common is a full wave bridge. This uses four diodes to take the absolute value of the input, minus two diode drops.

The two AC signal lines are connected to A and B. The + DC output is taken at C and the – is taken at D. When A is positive and B is negative, the current flows from A to C through the top left diode, then through the load resistor to D and then to B through the lower right diode. For this to occur $V_A > V_B + 1.2V$ to turn on both diodes. Try and figure out what happens if $V_A < V_B - 1.2V$.

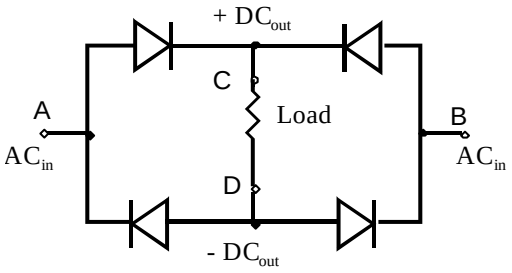
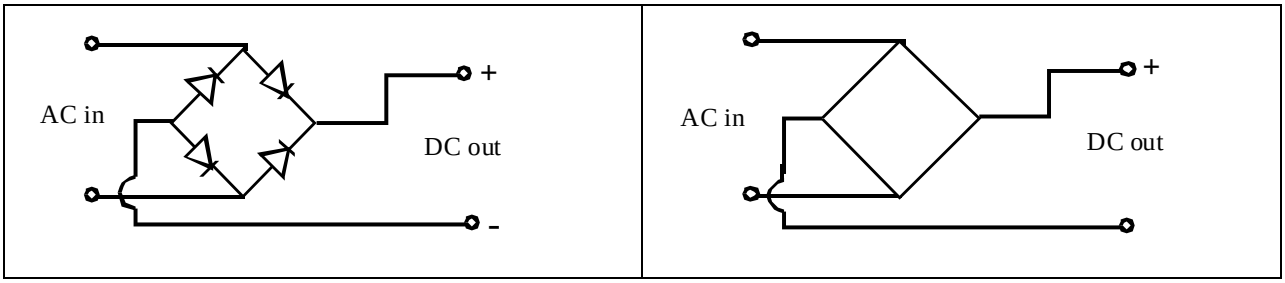


Fig. 4.8 Full Wave Bridge

Note that the current always flows through the load from top to bottom, making C the positive side of the load and D the negative. This is because the diodes only conduct current flowing into C and out of D. This is often drawn as shown below left. Sometimes the diodes are not shown, just a diamond shaped box, with the diodes understood, below at the right. (The diodes are hard to draw, but it is easy to draw a diamond shaped box.)



The figure at the right shows the waveform resulting from a full wave bridge. You should notice that the output peak is always below the |input| because of the voltage necessary to turn on two diodes. If the voltage isn't large enough to turn on two diodes (i.e. > 1.2V), then the output is 0V.

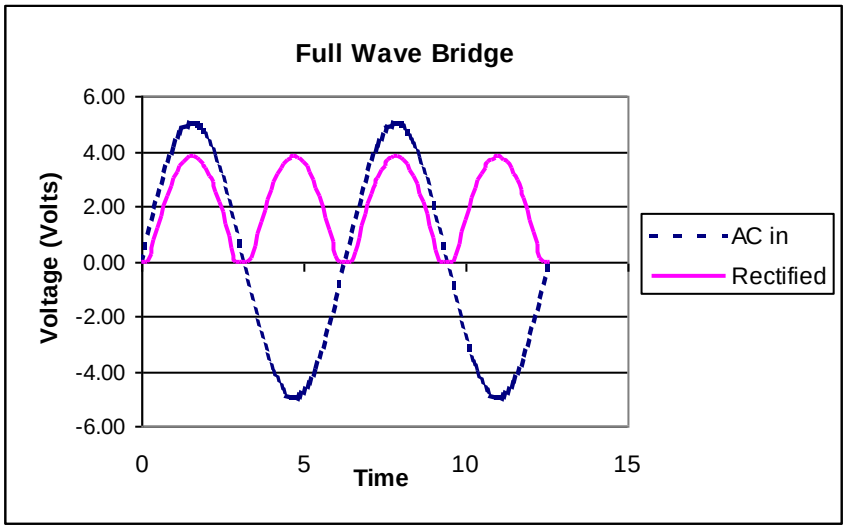


Fig. 4.9 Full Wave Rectification

If you are making a power supply, the AC_{in} is usually the 60Hz voltage from the power mains, but stepped down by a transformer to a lower voltage, say 10-20V_{rms}. In fig. 4.9 above the peak voltage in is 5V, so the RMS is 3.5V.

The output voltage in fig. 4.9 for the circuit in fig. 4.8 goes to 0V when the input signal crosses 0V. If you are making a DC power supply, you want the output voltage to be constant. You can make it more constant if you put a capacitor in parallel with the load resistance. If the input AC is 60Hz, you want the time RC time constant for the load resistance and the capacitance to be long compared to one half the period of 60Hz, or somewhat longer than 8.3ms. If you expect a load resistance of 10Ω or so, you might use a C of 2,200μF. This would ensure that the voltage would only vary by about 30%. If you use 4,700 μF it would vary by about 15%. You can almost eliminate the variation by regulating the output voltage, but to understand regulation you need to look at op amps.

B: Special Diodes

Sometimes we want to generate a fixed voltage from an input that is varying like the rectified and filtered signal discussed above. Say you have and input voltage that varies between 9 and 11Volts, but you want an output of 3.0Volts that is relatively stable. A crude way of doing this is shown at the right. The 470Ω resistor limits and the five diodes each drop 0.6V producing an output voltage of 3.0V. As long as I don't try to draw too much current out of the circuit at V_{out}, say less than 5mA, the output voltage remains fairly stable at 3.0V. The voltage across a diode does depend slightly on the current through it. Doubling the current increases the voltage across a silicon diode by almost 40mV.

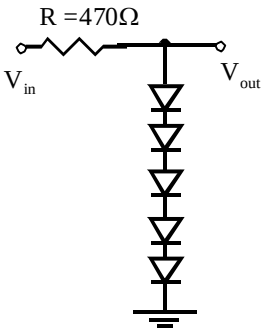


fig. 4.10

As a result you want R small enough so that the current through the diodes does not vary too much when you draw current out of V_{out}. Also the drop across a silicon diode changes with temperature. Both of these defects make this a crude way to produce a fixed voltage.

Another type of diode is designed to be used as a reference. If you reverse bias a diode, very little current will flow. However, if the reverse voltage gets too large, the diode will "break down" and conduct current. If you allow too much current to flow, it will destroy the diode. The silicon signal diodes we use in the lab have breakdown voltages on the order of 50 to 100V.

Power diodes used for rectification in power supplies often have breakdown voltages of 50V to 1000V.

Special diodes, called Zener diodes, are made to breakdown at a specific voltage, e.g. 5.1V. The symbol for a Zener diode is show at the right. In the forward direction they act just like a regular diode, but if they are reverse biased, they won't conduct until they reach their designed breakdown voltage. A reference made with a Zener diode is shown at the right. This is a slightly better way to make a voltage reference, but the output will still have some dependence on the current through the diode and on the temperature. However, it will be a better, and simpler, than the five regular diodes in series. One can make better references with transistors and op amps.

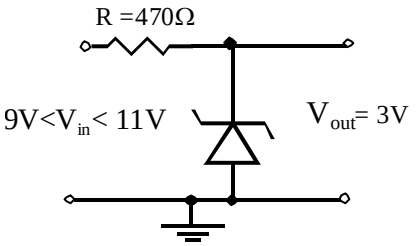


fig. 4.11

C: Bipolar Transistors

The behavior of transistors is somewhat more complicated, so I'll use a very simple model. There are three major types of transistors, bipolar junction transistors (BJT) , junction field effect transistors (JFET), and metal-oxide semiconductor field effect transistors (MOSFET). (Within these major types there are also subtypes which we will only consider briefly.) All these can be considered three terminal devices where a signal applied at one terminal (the gate for FET's and the base for BJT's) controls the flow of current between the other two terminals. One is called the collector (for BJT's) or the Drain (for FET's) and the other is called an emitter (for BJT's) or a source (for FET's). For BJT's a small current at the base controls a larger current flowing from the collector to the emitter. A typical setup for a BJT might have a battery connected between the collector and the emitter, with the + terminal connected to the collector and the – to the emitter. (For safety, you often have some resistance in series with the battery to limit the current in case you connect something incorrectly.)

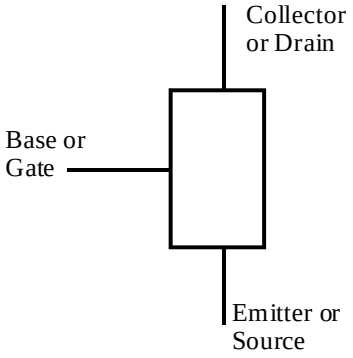


Fig. 4.12 Model for a Transistor

This is shown in fig. 4.13. The base-emitter junction (BE junction) is a diode with the base a p-type material and the emitter n-type. This type of BJT is called an npn transistor. To get current to flow in the transistor the base voltage must be high enough to forward bias the BE junction, or $V_{BE} > 0.6V$. The small current flowing from the base to emitter, I_B, allows a much larger current to flow from the collector to the emitter, I_C. I_C is approximately proportional to I_B and the relationship is written as

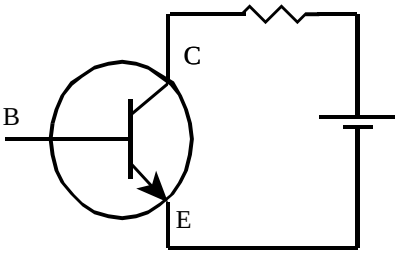


Fig. 4.13 A BJT Connected to a Battery

$$I_C = \beta I_B = h_{FE} I_B \qquad 4.3$$

h_{FE} and β are the same. Normally $10 < \beta < 300$, depending on the transistor sub-type. For this npn transistor to work properly, you need $V_C > V_E$ and typically you want it at least 1 to 2 V greater than V_E. If $\beta = 100$, then a base current of 1mA would allow 100mA to flow from the collector to the emitter. This type of configuration is referred to as a common emitter configuration and it is often used as an amplifier or switch. The circuit in figure 4.14 is a slightly rearranged version of fig. 4.13. Normally one puts a resistor in the base lead to prevent too much current from being drawn through the base lead and to modify the gain of the circuit. If Vin < 0.6V, then no current flows and the collector is at V_+, the positive power supply voltage. (If there is no current through R_C, there is no drop across it.) As V_{in} becomes greater than 0.6V, the BE diode turns on and V_B stays at 0.6V. The difference between V_{in} and V_B is dropped across R_B, so

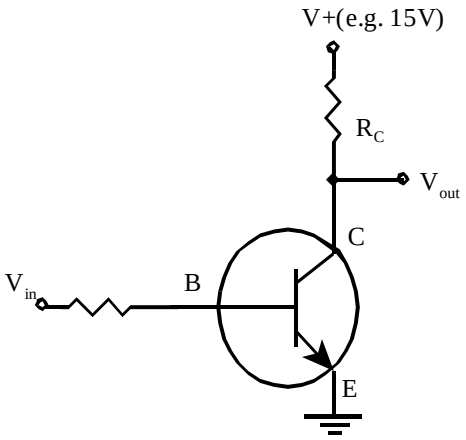


Fig. 4.14 Common Emitter Amplifier

$$I_B R_B = V_{in} - 0.6V \qquad \text{or} \qquad I_B = \frac{V_{in} - 0.6V}{R_B} \tag{4.4}$$

Since $I_C = \beta I_B$,

$$I_C = \beta \frac{V_{in} - 0.6V}{R_B} \tag{4.5}$$

Now $V_C = V_+ - I_C R_C$, so

$$V_C = V_+ - \beta (V_{in} - 0.6V) \frac{R_C}{R_B} \tag{4.6}$$

It is interesting to make a plot of V_C vs V_{in}. It will be a straight line with a slope of

$$\text{slope} = \frac{\partial V_C}{\partial V_{in}} = -\beta \frac{R_C}{R_B} \tag{4.7}$$

This slope is often called the gain of the circuit. If $\beta = 150$, $R_C = 10k$ and $R_B = 10k$, then the slope or gain is – 150. This means that a 10mV increase in the input will produce a 1.5V <u>decrease</u> in the output. V_C must satisfy $1V < V_C < V_+$ for this equation to work well. β decreases as V_C approaches V_E, so as V_C drops below 1V the gain will decrease, but V_C can still get down around 0.1V or so. I've plotted this for $0.55V < V_{in} < 0.8V$ where I've assumed V_{in} can get close to 0V and that $V_+ = 15V$. Note that for $V_{in} < 0.6V$ $V_{out} = 15V$ and for $V_{in} > 0.7V$ $V_{out} = 0V$. It is only linear between $0.6V < V_{in} < 0.7V$. This type of amplifier primarily amplifies voltage. It is awkward to use it as an amplifier in this form because β is not a "good" number to design around, i.e. it is not really a constant and it varies from transistor to transistor. However it does

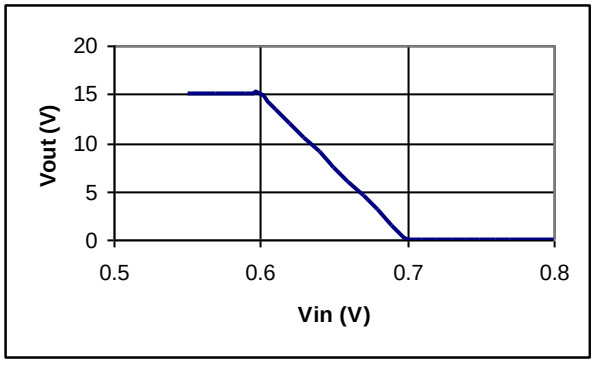


Fig. 4.15 Vout vs Vin for fig. 4.12

form the basis of more complicated amplifier circuits. (Moving the base resistor to the emitter will produce a "better" voltage amplifier.)

The circuit in fig. 4.14 is <u>often used as a switch or logic inverter.</u> If R_B is small, say 1k and R_C is 5k, the voltage gain is – 750. If the input is below 0.6V the output is high, V_+, and if it gets much above 0.6V, the output goes low, close to 0V. If V_+ = 5V, this would be a logic inverter. If V_+ = 5V, the transition would occur for a change of about 7mV in V_{in}.

The second configuration is called an emitter follower. It is used as a current or power amplifier and some form of this is usually found in the output stage of most amplifiers. The collector is connected to the positive power supply, V_+, the input is connected to the base and there is a resistor between the emitter and ground. The output is taken at the emitter. If $V_{in} < 0.6V$, no current flows and $V_E = V_{out} = 0V$. If $V_{in} > 0.6V$, the transistor provides enough current to keep V_E 0.6V below V_{in}, so $V_{out} = V_{in} - 0.6V$. Here the voltage gain of the circuit, $\partial V_{out}/\partial V_{in}$, is +1. The virtue is that the source of V_{in} only needs to supply a small fraction, $\approx 1/\beta$, of the current that goes through R_E. For instance, if $R_E = 8\Omega$ and V_{out} = 8V, I_E = 1.0A and β = 100, The input need only supply about 0.010A of this, the rest is provided by the power supply via the transistor. (This is also called a class A power amplifier.) It is called an emitter follower because voltage changes at the output, the emitter, follow voltage changes at the input, just 0.6V below the input.

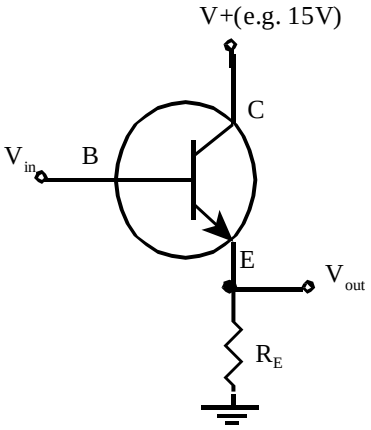


Fig. 4.16 Emitter Follower

As suggested before, a more useful version combines these two with resistors in both the emitter and collector leads. If the output is taken at the collector, the gain of the circuit is approximately $-R_C/R_E$. Again, $V_{out} = 15V - I_C R_C$ and $V_E = V_{in} - 0.6V$ if V_{in} is in the proper range. However, $V_E = I_E R_E$ and $I_E = I_C + I_B$. But since $I_B = I_C/\beta$, we also have $I_E = I_C(1+1/\beta)$. If $\beta >> 1$, $I_C \approx I_E$. At this level of approximation,

$$V_{out} = 15V - \frac{\beta}{\beta+1}\left(\frac{R_C}{R_E}\right)(V_{in} - 0.6V)$$

If you take the derivative of this with respect to V_{in}, you have

$$\frac{\partial V_{out}}{\partial V_{in}} = -\frac{R_C}{R_E}\left(\frac{\beta}{\beta+1}\right)$$

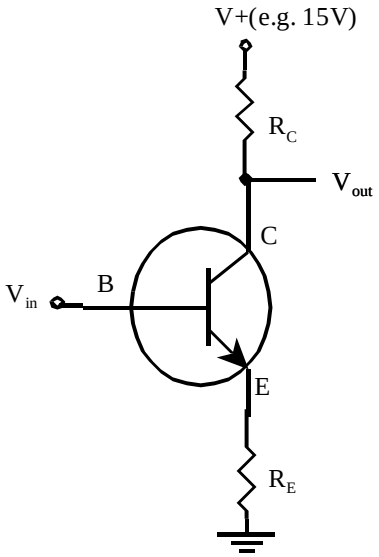


fig. 4.17

This is approximately independent of β if β is >>1. If β = 150 or 200, you get about the same gain. However, there is only a small range of input voltages for which this relation holds. If $V_{in} < 0.6V$, V_{out} = 15V. Also when V_{in} approaches V_C, this relation won't work. For $R_C = 5R_E$, this occurs when V_{in} reaches 3V. So for a gain of – 5, the circuit is useful for $0.6V < V_{in} < 3V$.

I have described an npn bipolar transistor. The pnp is a mirror image of an npn. The collector of the pnp is connected to the negative power supply and the base has to get 0.6V

below the emitter to turn on the transistor. I'm not going to discuss it further, but if you are going to use pnp transistors, consult one of the references.

D: Field Effect Transistors

Field effect transistors use an applied voltage V_{Gate} to control the resistance of the channel connecting the Source and Drain. Almost no current flows into the gate, except to charge and discharge the gate's capacitance. There are many subtypes of FETs, JFETS, MOSFETS, and subdivisions within these subtypes, p-channel, n-channel, enhancement mode, depletion mode etc. Again, I'm not going to discuss these in any detail. Consult a reference if you need to use them. JFETs and MOSFETs are often used as the input stage for op amps because they draw so little current through the gates, often in the pA (10^{-12}A) range. MOSFETs are also used in very large scale integrated circuits, e.g. microprocessors, memory chips. They are also used as analog switches because they really do act like a variable resistor between the Drain and Source. When the transistor is on, the resistance between the drain and source is typically 50Ω to 300Ω for small MOSFETS and roughly 0.01Ω to 2Ω for power MOSFETS. When they are off, the resistances are often greater than 10^9Ω, especially for the smaller MOSFETS. The symbol for a NMOS transistor (n-channel MOSFET) is shown at the right. For an NMOS the primary charge carriers in the channel are electrons (−), for a PMOS transistor they are holes (+). The source (S) is the source of electrons for an NMOS transistor and should be negative relative to the Drain (D). The motion of electrons in the transistor is normally from the source to the drain. The voltage at the gate controls the conductivity of the channel between the source and the drain. (The channel is the region under the gate and I haven't shown the "Bulk" connection.) In an enhancement mode NMOS the channel normally has very few charge carriers to carry current if $V_G \approx V_S$. (V_G = gate voltage and V_S = source voltage, V_{GS} is the voltage of the gate relative to the source and V_{DS} is the voltage of the drain relative to the source. V_{DS} is usually ≥ 0V.) To get electrons to move into the channel, the gate must be made positive relative to the source which attracts electrons into the channel and it becomes conducting. The gate voltage where the conductivity of the channel starts increasing dramatically is called the threshold voltage. Over a small range of gate voltages and drain-source voltages, the channel acts like a resistor, but in digital circuits they are usually used in a mode where the channel is either highly conducting or highly insulating, i.e. ON or OFF.

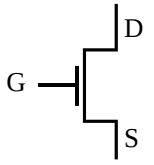


fig. 4.18: An N channel MOSFET

The structure of an NMOS transistor is shown at the right with the geometry of the channel and oxide layer exaggerated. The wires to the Source and Drain connect to heavily doped n-type regions which are very conducting. The channel has very few free charges of its own and only becomes conducting when charges (electrons) are induced to enter that region when the gate is made positive enough. The Oxide layer insulates the metal gate from the channel, and it is very thin. The Bulk would be a p-type material.

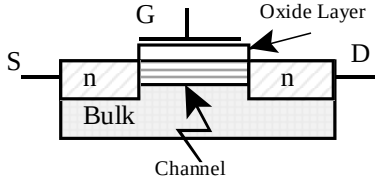


fig. 4.19 Structure of a n channel MOSFET

An enhancement mode PMOS transistor works just the opposite, the gate must be made negative relative to the source (of holes this time) to draw holes into the channel so that it becomes conducting. Here, V_{DS} is usually negative, $\leq$)V. The threshold voltage in both of these types can be controlled by the manufacturing process.

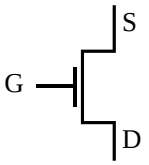


fig. 4.19: A PMOS transistor

In fig. 4.20 I have shown an NMOS operated in the common source mode, the equivalent of a common emitter configuration for an npn BJT. Here the input is to the gate and as the gate voltage rises the and channel becomes conductive, its resistance might change from $10^9\Omega$ to 50Ω as the gate voltage goes through the threshold voltage. With a 5k resistor between the drain and the +5V, the drain voltage will appear to change a lot with a small change in V_G. In this case the threshold voltage is around 2.5V. When V_G is small, the transistor is OFF and there is no current through the 5k resistor so there is no drop across the resistor and therefore $V_D = 5V$. When V_G approaches the threshold, the resistance of the channel drops dramatically and the current through the 5k resistor increases dramatically and V_D drops towards 0V as the resistance of the channel approaches 50Ω or so. If the channel resistance is 50Ω, then V_D would be about 50mV at that point. (A PMOS transistor is just the opposite in the way it works.)

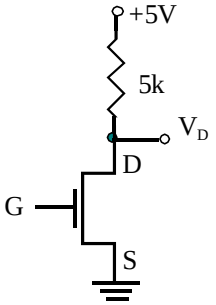


fig. 4.20 A common source NMOS transistor

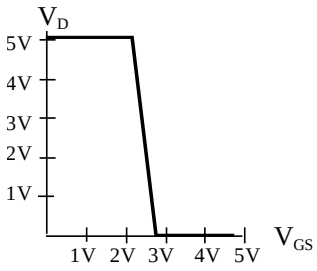


fig. 4.21 The variation of V_D with V_G