Basic Electronics with Radio Applications

Oct 2022

KF4DII

Chapter 1: Electricity

Atomic Structure.

All elements are made up of atoms. An atom has a nucleus which consists of positively charged protons and neutrons which have no charge. Around this, at various energy levels or shells are negatively charged electrons. Each shell likes to have a fixed number of electrons in it. The outer shell, called the valence shell, is the one we are usually most interested in. if the number of electrons in the valence shell is small with respect to the number it would like to have in that shell, then those electrons can be easily "knocked off" by external forces creating a flow of free electrons. An element with such arrangement would be a good conductor. If the valence shell is "full", the atom is relatively stable and few free electrons are generated. Such an element would make a good insulator.

There is a relationship between charges that will play a part in electronics over and over again. This relationship is captured in the following principle:

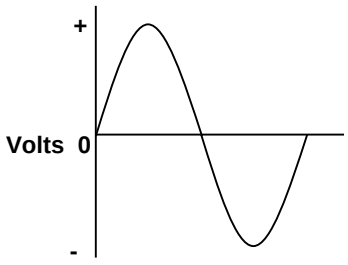
Like charges repel, unlike charges attract.

Electricity, then, is just the flow of free electrons in a conductor. The flow of the electrons is called **current** and it is measured in Amperes or **Amps** for short. The force the "pushes" the electrons is called **voltage** and is measure in **Volts**. **Resistance** impedes the flow of current and is measured in **Ohms**.

There are two types of current: direct current, abbreviated **DC**, and alternating current, abbreviated **AC**. With DC, the electrons always flow in one direction through a circuit. A car battery is a typical example of DC current. With AC, the electrons flow back and forth through a circuit. This is caused by the power source changing its polarity or switching back and forth between positive and negative. This causes the current to flow "back and forth" as well, hence the term, alternating current.

A typical AC signal is a sine wave, shown below. In this figure we start the sine wave at zero Volts and increase smoothly in a positive direction until it reaches peak positive voltage. It then smoothly decreases back to zero Volts, reverses polarity (i.e., changes from positive to negative voltage) and begins to swing in a negative direction until it reaches peak negative voltage. At this point, it begins a smooth swing back to zero volts. This completes *one* cycle of the AC sine wave.

The number of such cycles in one second is the sine wave's frequency, measured in Hertz, abbreviated Hz. One cycle per second is one Hz; 10,000 cycles per second is 10,000 Hz or 10 kHz (where the "k" stands for "kilo" or 1,000); 10,000,000 cycles per second is 10,000,000 Hz or 10 MHz (where "M" stands for "mega" or 1,000,000).



The time it takes to travel one cycle is the period often designated as "T". If the period is known, the frequency can be determined:

$$\text{Frequency} = 1 / T$$

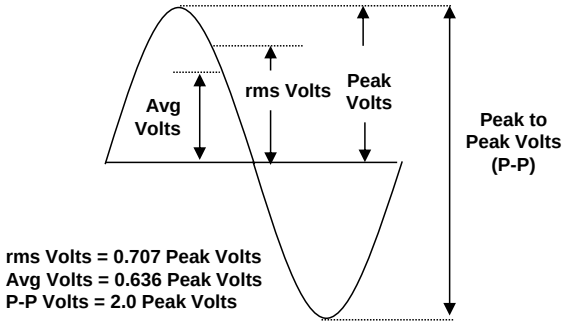
In many electronic circuits, "T" can be much smaller than one second. One millisecond, ms, is $1/1000^{th}$ of a second. One microsecond, μs, is $1/1{,}000{,}000^{th}$ of a second. If "T" is in seconds, frequency will be in HZ; if "T" is in milliseconds, frequency will be in KHz; and if "T" is in microseconds, frequency will be in MHz.

If the period, T, is known, the frequency can be determined:

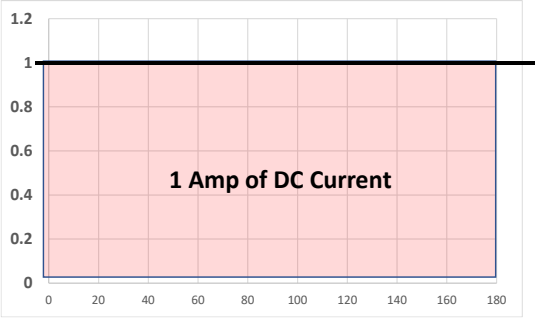
$$T = 1 / \text{frequency}$$

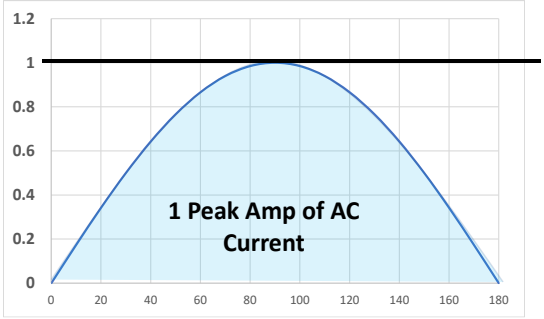
If frequency is in Hz, "T" will be in seconds; if frequency is in KHz, "T" will be in milliseconds; and if frequency is in MHz, "T" will be in microseconds.

Another of the characteristics of AC is its amplitude in Volts. There are four different ways to measure amplitude: (1) peak to peak volts, (2) peak volts, (3) RMS volts, and (4) average volts. The figure below illustrates these four ways to measure amplitude. The most common one used is RMS volts because it gives us the best indication of the useful power in an AC signal.



Consider now 1 Amp of DC current flowing through a 1 Ohm resistor. The power dissipated in the resister would be P, Watts = I^2 x R = 1^2 Amp x 1Ω = 1 Watt. But compare 1 Amp of DC to 1 ***peak*** Amp of AC as shown below –





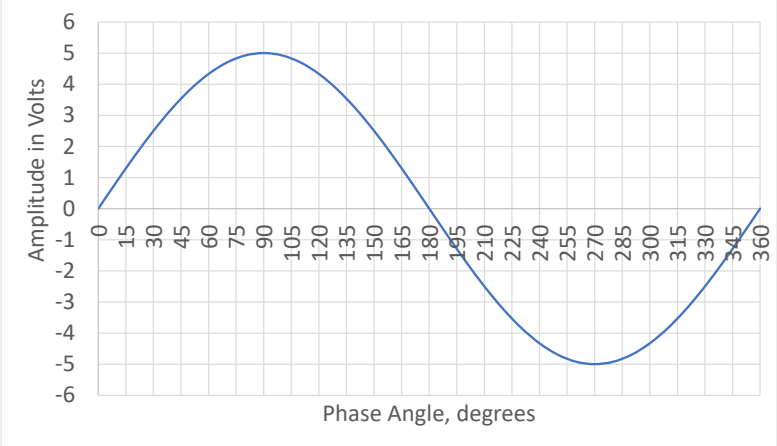
Due to the variations in the amplitude of the alternating current, it does not provide the same amount of current over time as 1 Amp of direct current which never varies in amplitude. This is why rms is used for AC –it provides equivalent performance to DC. For example, it heats the 1Ω resistor as much as the 1 Amp DC, as shown below -

Amp, DC	1
R, Ω	1
P_{DC}, W	1

Amp AC, $_{PEAK}$	1.414
Amp AC, $_{RMS}$	0.999698
R, Ω	1
P_{DC}, W	1.00

Notice in the right-hand table above, if we use 1 Amp AC, ***rms***, we get the same power dissipation in the resistor, 1 Watt, as we do with 1 Amp DC, as shown in the left-hand table above. Thus, the required AC peak current must be 1.414 Amps but when the 0.707 multiplier for rms is applied, we get the same heating power as 1 Amp DC. So, for AC, rms proves to be "equivalent" to DC.

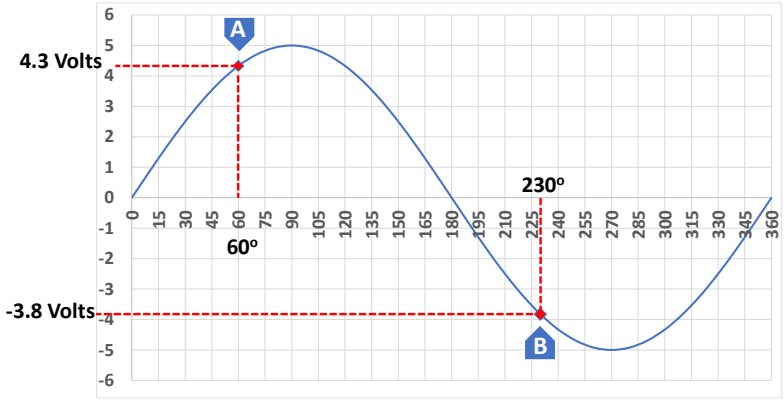
A third characteristic is **phase angle** and it is measured in units of **degrees**. Phase angle defines where on a single sine wave you are. One cycle is divided into 360 degrees. Each phase angle on the sine wave, in degrees, is illustrated below.



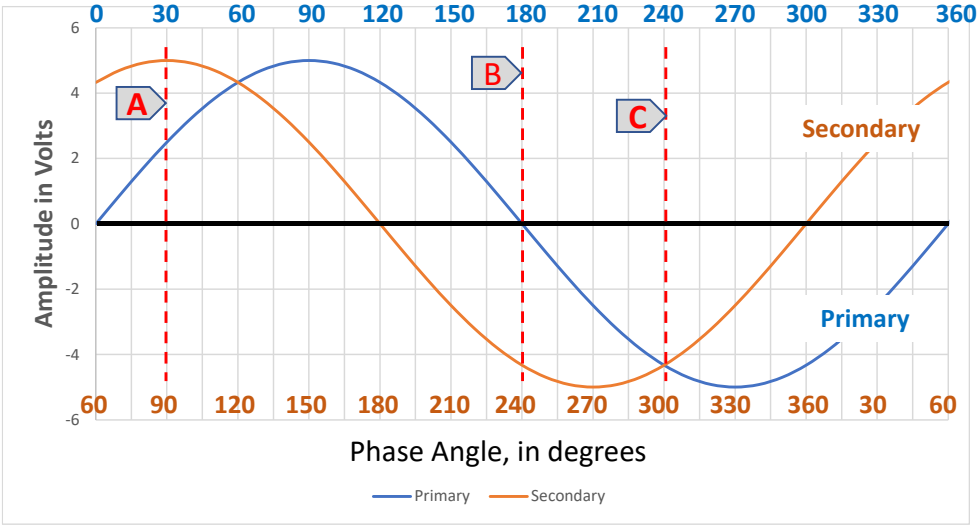
In this example, the sine wave starts at an amplitude of 0 Volts at 0 degrees phase angle and starts to increase in a positive direction. At 90 degrees the sine wave is at peak positive amplitude, which in this example 5 Volts. It now starts to decrease back to zero. It reaches zero at 180 degrees where it reverses polarity from positive to negative. It then continues swinging negative until it reaches its peak negative amplitude at 270 degrees. It then swings back toward zero volts, arriving there at 360 degrees. Thus, every phase angle position, in degrees, defines a specific point on the sine wave with a specific amplitude and polarity, positive or negative.

For example, consider points "A" and "B" on the sine wave below. Point "A" is at a phase angle of 60° with an amplitude of 4.3 Volts. This can be seen by tracing down from point "A" to the phase angle scale where it reads 60° and by tracing from point "A" to the left to the amplitude scale where it reads 4.3 Vols. Thus, point "A" on the sine wave represents a phase angle of 60° and an amplitude of 4.3 Volts. Following the

same process, it can be seen that point "B" represents a point on the sine wave of 230° phase angle and an amplitude of -3.8 Volts.

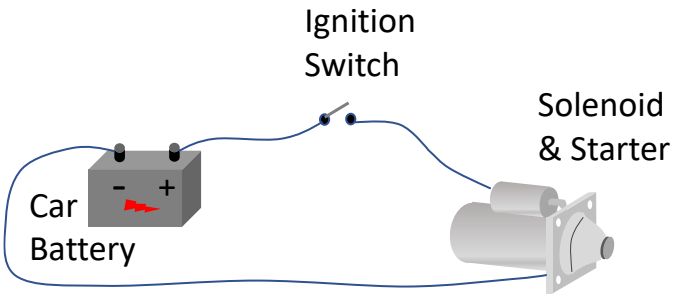


Phase can also be used to define the relationship between two sine waves, as shown below. The blue curve represents one sinewave, and its degrees are shown in blue text across the top of the figure. This sine wave is identified as the primary. The orange curve represents a second sine wave, identified as the secondary. Its degrees are shown in orange text along the bottom. Both sinewaves are present at the same time but they are not in sync. For discussion purposes, three arbitrary points have been selected: A, B, and C. Starting with point A, we can read the phase angle of the primary by tracing the red line up to the top scale where it can be seen that the primary's phase angle reads 30°. Tracing the same line down, it can be seen that at that same time, the secondary's phase angle reads 90°. Thus, these two sine waves are 60° apart (90° – 30° = 60°). Note that the secondary reaches its peak positive *before* the primary has. This means the secondary *leads* the primary by 60 degrees. Or said another way, the primary *lags* the secondary by 60 degrees. Following the same procedure, it can be seen that the same 60° difference occurs at points B and C as well.

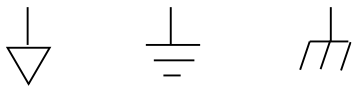


Closed Circuit

In order for current to flow in a conductor, there must be a closed loop or closed circuit. Consider the sketch below of a car battery, ignition switch, and solenoid and starter. Nothing happens until the key is inserted in the ignition switch and turned on. A closed loop is then formed from the positive battery terminal, to – and through – the ignition switch to the solenoid and starter and then back to the battery negative terminal, usually through the car's chassis. Current then flows, the solenoid engages and the starter cranks the motor. Open this circuit in any place – turn off the key, remove a battery terminal, break a wire, etc., and everything stops.



In many electronic systems. one side of the power supply is connected to the metal chassis and serves as a common power return. It is usually connected to Earth through the power cord. This return is called the common side or ground. It is called common because it is the common conductor that all the circuits use to return their electron flow to the power source. It is also called ground because many times it is actually connected to an earth ground, often for safety reasons. Acceptable schematic symbols for common or ground are shown below



How Electricity is Generated

Electricity can be generated chemically or mechanically. A battery is a device that converts chemical energy to electrical energy. The chemicals inside the battery undergo a chemical change to release free electrons. With some batteries this chemical change is reversable by recharging, in other batteries it is not. A carbon-zinc battery is not rechargeable; lead acid, lithium-ion, and nickel – metal hydride batteries are. An important take away is that a battery can provide current for only a fixed period of time. This period of time is measured in **amp – hours**, or how many amps of current a battery can provide in one hour. For example, a typical 9 VDC battery is rated at 550 mAh, meaning it can provide a steady current of 550 mA for one hour. If you use less current, say 100 mA, then it will last 5.5 times longer (550 mAh / 100 mA = 5.5 hours).

AC is generated by mechanical means. The principle behind this is as follows:

If a conductor cuts a moving magnetic field at right angles to the field, a current flow will be induced in the conductor.

Either the magnetic field must be moving (expanding out and collapsing back in) or the conductor must be moving; it doesn't matter which.

This is how power for our homes is generated. Steam is used to providing the required movement. The steam can be generated from the heat of a nuclear reaction or the heat from burning coal or natural gas in a boiler.

Chapter 2: Ohm's Law and Joule's Law

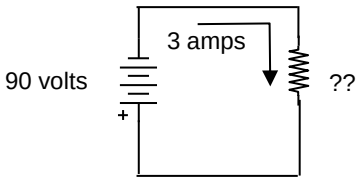
Ohm's Law. Current, voltage, and resistance interact. A simple, and often used, analogy is a water hose connected to household water spicket. Turn on the faucet and water comes out. The water pressure, supplied by the city water system, is the pressure that pushes the water out. This is analogous to voltage. The water flow through the attached hose is analogous to current flow (free electrons) through a conductor (the hose) Squeeze the hose and the water flow decreases. This is analogous to resistance. In direct current, the current is always "pushed" by the voltage in one direction. The water flowing out of our hose is analogous to direct current – the water flows only one way.

Ohm's Law describes the relationship between resistance, current, and volts. This Law has three equations:

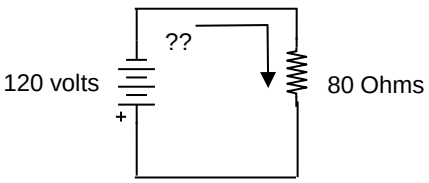
$$E = I * R \qquad I = \frac{E}{R} \qquad R = \frac{E}{I}$$

The letter "I" means current and is measured in Amperes or Amps. One thousandth of an Amp is a milliamp abbreviated mA. One millionth of an Amp is a micro amp, abbreviated μA. The letter "E" means volts ("V" is also used). One thousandth of a Volt is a millivolt, abbreviated mV. One millionth of a Volt is a microvolt, abbreviated μV. The letter "R" means resistance and is measured in Ohms, abbreviated as Ω. A thousand Ohms is designated as kΩ with the "k" indicating units of one thousand. A million Ohms is designated as MΩ with the "M" indicating units of one million.

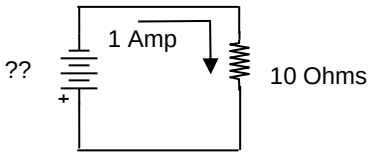
If any two of the three variables (current, voltage, or resistance) is known, the third can be calculated using Ohm's law. Three examples are shown below:



$$R = \frac{E}{I} = \frac{90\ \text{Volts}}{3\ \text{Amps}} = 30\ \text{Ohms}$$



$$I = \frac{E}{R} = \frac{120\ \text{Volts}}{80\ \text{Ohms}} = 1.5\ \text{Amps}$$



$$E = I \times R = 1\ \text{Amp} \times 10\ \text{Ohms} = 10\ \text{Volts}$$

Joule's Law. Power is the rate of doing work. For example, a resistor does work by impeding current flow. When it does this work, it converts electrical energy to thermal energy (heat). An electric motor does work by turning a shaft. The symbol for power is "P". It is measured in Watts, abbreviated "W". Joule's Law describes the relationship between resistance, current, and power. This Law has three commonly used derivations:

$$P = E * I \qquad P = \frac{E^2}{R} \qquad P = I^2 * R$$

One thousand Watts is a kilowatt, abbreviated KW. A million Watts is a megawatt, abbreviated MW. One Watt divided by 1,000 is called one milliwatt, abbreviated mW. One Watt divided by 1,000,000 is one microwatt, abbreviated μW.

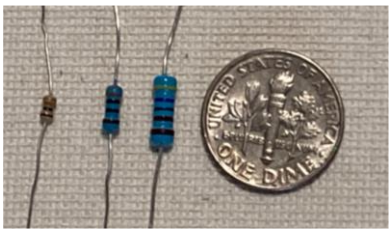
Chapter 3: Resistors

A ***resistor*** is a component that impedes the flow of current and is measured in Ohms, abbreviated Ω. If one Volt pushes one amp of current through a component, that component has a resistance of one Ohm. A resistor may have thousands or millions of Ohms, expressed as kilohms (KΩ) or megohms (MΩ).

Resistors can be made of wire wound on a core, carbon, or even thin metal films. Resistors come in different sizes. The bigger they are, the more power they can handle. When a resistor impedes current flow, it converts electrical energy to thermal energy (heat). The resistor must get rid of this heat or dissipate it before the resistor gets too hot and is damaged. The bigger the resistor is physically, the more heat it can dissipate and the more power it can handle. The picture below illustrates resistor size, compared to a dime, vs power rating.

1/4 W

1/8 W 1/2 W



If 2 Amps is going through a 1 Ohm resistor, the power dissipated by the resistor, in Watts is -

$$P = I^2 R = (2 \text{ Amps})^2 \times (1 \text{ Ohm}) = 4 \text{ W}$$

The 1-ohm resistor must be big enough to dissipate at least 4w. ***To be safe, we generally design a circuit so that the resistor is dissipating no more than ½ its power rating.***

The most basic specifications of a resistor are the resistance value, tolerance, and power rating. The resistance value is how many Ohms the resistor has. The tolerance tells us how close the manufacturer kept the resistance to its specified value, measured in percent. The power rating tells us how much power the resistor can handle. An example of a basic resistor specification is:

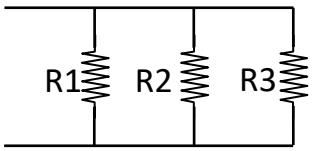
Resistor, Carbon, 10K, 20%, 1W – a resistor made of carbon with a resistance somewhere between 8,000 and 12,000 Ω, capable of dissipating 1 Watt.

Resistors can be fixed or variable. A fixed resistor has a constant resistance. With a variable resistor, the user can vary the resistance value by rotating a shaft on the resistor. There are different kinds of variable resistors. One of the most common is called a potentiometer. The schematic symbols for a fixed and variable resistor are shown below

Fixed Resistor Variable Resistor

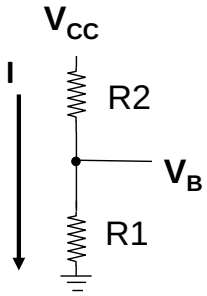
Resistors in Series and Parallel

Resistors in series add. The equivalent circuit of series resistors is the sum of all the resistors in series. Resistors in parallel are a bit more complicated, but the figure below summarizes the process.



$$R = \frac{1}{\frac{1}{R1} + \frac{1}{R2} + \frac{1}{R3}}$$

A practical application of resistors is a voltage divider. This is used to establish the proper bias voltage, V_B, for a transistor amplifier, which will be discussed in more detail later. For now, just assume that VB is the voltage that will be required for a transistor amplifier design and that the required current, I, and power supply voltage, Vcc, are both known. The panel below illustrates the design process.



Resistors are often used as voltage dividers. V_{CC} is a common term used for power supply voltage.

Problem: If V_{CC} = 10 VDC and I = 1mA, what values do R1 and R2 need to be to set V_B = 1.7 VDC?

Solution: By inspection, it can be seen that the voltage across R1 is V_B and the current through R1 is 1 mA. We can use Ohms law to determine the required value for R1.

Ohm's Law: $R = \frac{E}{I}$

So … $R1 = \frac{V_B}{I_C} = \frac{1.7\ \text{VDC}}{0.001\ \text{A}} = 1700\ \Omega \text{ or } 1.7\ \text{k}\Omega$

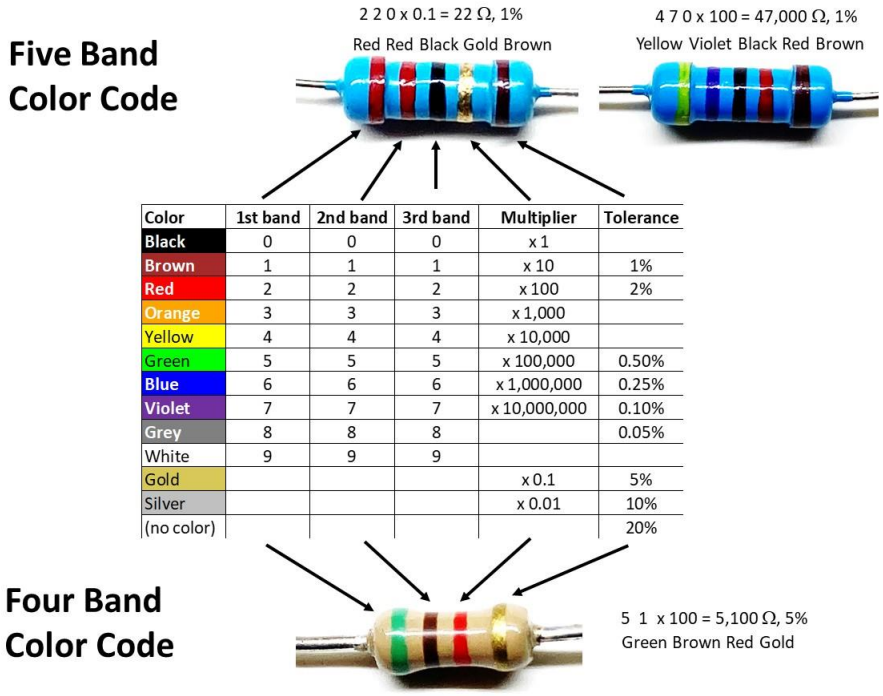
1 mA converted to Amps

By inspection, the voltage across R2 is $V_{CC} - V_B = 10 - 1.7 = 8.3$ VDC. The current through R2 is I mA, so …

$$R2 = \frac{V_{CC} - V_B}{I_C} = \frac{10 - 1.7\ \text{VDC}}{0.001\ \text{A}} = 8300\Omega \text{ or } 8.3\ \text{k}\Omega$$

Resistor Color Codes

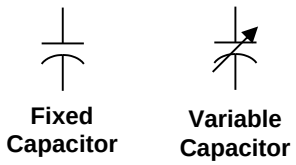
Resistors are marked with bands of colors to denote their value and tolerance. The sketch below illustrates this for five band and four band color coding.



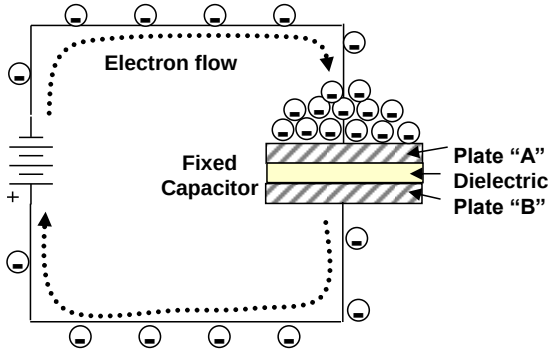
Color	1st band	2nd band	3rd band	Multiplier	Tolerance
Black	0	0	0	x 1	
Brown	1	1	1	x 10	1%
Red	2	2	2	x 100	2%
Orange	3	3	3	x 1,000	
Yellow	4	4	4	x 10,000	
Green	5	5	5	x 100,000	0.50%
Blue	6	6	6	x 1,000,000	0.25%
Violet	7	7	7	x 10,000,000	0.10%
Grey	8	8	8		0.05%
White	9	9	9		
Gold				x 0.1	5%
Silver				x 0.01	10%
(no color)					20%

Chapter 4: Capacitors

A capacitor is a component that stores electrical energy. A capacitor is made up of two metal plates separated by an insulator, called the dielectric. The schematic symbol for a capacitor is shown below. Note, that like a resistor, capacitors can be fixed or variable.



If a battery were connected across the capacitor, electrons would flow out of the battery's negative terminal to the capacitor where the dielectric would act as an insulator and block further electron flow. However, the electrons, pushed by the battery's negative terminal and blocked by the capacitor's dielectric would simply "pile up" on the capacitor's plate, plate "A" as shown below.



This would begin to build up a negative charge on plate "A" which would repeal the negative electrons on the other capacitor plate, plate "B", since like charges repel. The electrons pushed off plate "B" are also attracted to the battery's positive terminal since unlike charges attract. This "current flow" would continue until the negative voltage on plate "A" is equal to the battery voltage and of opposite polarity. At this time, the capacitor is fully charged. If the battery were removed, the capacitor would remain fully charged. That is, the capacitor would now store the energy. If the dielectric were a perfect insulator, the capacitor's charge would stay forever. But the dielectric is not a perfect insulator and a small amount of leakage current would flow through it, eventually discharging the capacitor. If a resistor were placed across the charged capacitor, the electrons now stored on plate "A" would flow through the resistor to plate "B". This would continue until the capacitor is fully discharged. At that point, the voltage across the capacitor would be zero.

Capacitance is measured in Farads. Capacitors can also be measured in microfarad, abbreviated μF; nanofarad, abbreviated nF; and Pico farad, abbreviated pF. It takes l,000,000 μF to make one Farad, 1,000,000,000 nF to make one Farad, and 1,000,000,000,000 pF to make one Farad. Capacitors are specified by their capacitance and working voltage, usually abbreviated WVDC. WVDC is the maximum voltage that can be applied across the capacitor plates without arcing through the dielectric and damaging the capacitor.

There are many kinds of capacitors. Some capacitors are polarized. This means that one lead of the capacitor is marked negative. This lead must always be connected to the most negative side of the circuit. An example of such capacitors are electrolytic and tantalum capacitors. Other capacitors do not have polarity constraints and are called, appropriately, non—polarized capacitors. These are often made of ceramic, mylar, and even silver and mica.

Capacitors in Series and Parallel

The total capacitance of capacitors in parallel is the sum of them all. The total capacitance of capacitors in series is calculate with the equation shown below. Note that this is exactly like the formula for resistors in parallel and the same calculator key strokes would apply.

$$C = \frac{1}{\frac{1}{C1} + \frac{1}{C2}}$$

Capacitors in AC

So far, we have looked at capacitor performance when subjected to DC. When an AC is applied across a capacitor, the capacitor behaves differently. First the generator piles electrons onto one plate of the capacitor, plate "A", just as the battery did. But instead of leaving them there, the AC reverses polarity, pulls the electrons off plate "A", and pushes them up on the other capacitor plate, plate "B". Because of this, the capacitor appears to "pass" AC. Electrons don't actually flow through the dielectric but the circuit will respond as though they did. Note that the negative electrons "piled" onto one plate *oppose* the source voltage, which reduces current flow. In this sense, the capacitor can act as a resistor.

Capacitors offer resistance to AC due to the opposing plate voltage discussed above. This resistance is called **capacitive reactance**, abbreviated X_c. It is calculated by the following equation:

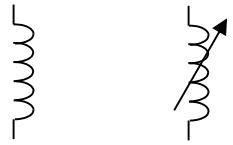
$$X_C = \frac{1}{2 \pi f C}$$

Where F = frequency, in Hertz
C = capacitance, in Farads
$2 \pi \sim 6.28$

Note that X_c varies inversely to both frequency and capacitance. This means that if the frequency or capacitance goes up, X_c goes down. If frequency or capacitance goes down, X_c goes up. X_c is measured in Ohms, just like a resistor is. In fact Ohms Law can be used with X_c just like resistance. Instead of putting the resistor value in the Ohms Law Equations, we put X_c. However, you must remember that the results you get for X_c are only good for that frequency.

Chapter 5: Inductors and Transformers

A third electronic component is an inductor. An inductor is a component that resists changes to current flow. It consists of several turns of wire usually wrapped around a piece of metal called a core. They can also be self-standing coils of wire with no metal core at all (air coils). The core is usually made of sheets of iron or powdered iron. The unit of inductance is the Henry. Most inductors are rated in millihenries, abbreviated mH or microhenries, abbreviated μH. It takes 1,000 mH to make one Henry and 1,000,000 μH to make one Henry. Its schematic symbol is shown below. Note that inductors can also be fixed or variable.



Fixed Inductor Variable Inductor

Inductors in Series and Parallel

Since inductors operate on magnetic fields, if two or more inductors are too close to each other or lack shielding, they will interact with each other. So, the total inductance of inductors in series is the sum of them all, *as long as their magnetic fields are shielded from each other or they are far enough apart that they don't interact.* Otherwise, their mutual coupling must be taken into account. Assuming there is no such interaction, a 1 mH inductor in series with a 4 mH inductor will provide a total inductance of 1 mH + 4 mH = 5 mH. The total inductance of the same two inductors in parallel would be 0.8 mH using the equation for inductors in parallel shown below, again, provided there is enough shielding or spacing so that their magnetic fields do not interact.

$$L = \frac{1}{\frac{1}{L1} + \frac{1}{L2}}$$

Inductors in AC

When current flows through a wire, it generates a magnetic field around the wire. If this magnetic field is moving and a second wire cuts across this moving magnetic field at a right angle, a current will be induced in the second wire. This is the basic principle on which inductors operate. Consider an AC voltage applied across an inductor that is starting at zero volts and begins moving towards peak positive voltage. As current begins to flow in the inductor windings, a magnetic field is generated which begins to expand out from the wire in the windings. As this magnetic field moves outward, it cuts across its *own* windings, inducing a current flow in itself. This self-induced current flow generates a counter electromotive force (CEMF) that *opposes* the original current flow. An inductor tries to resist a change in current flow. If current is not flowing, an inductor will resist external forces trying to make current flow. If current is already flowing, it will try to keep it flowing.

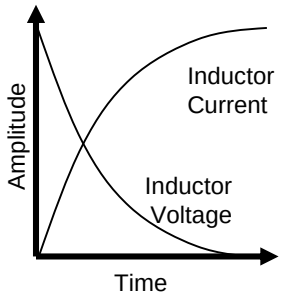
Like a capacitor, inductors offer resistance to AC, due to the CEMF described above. This resistance is called **inductive reactance**, abbreviated X_L. It is calculated by the following:

$$X_L = 2 \pi f L$$

Where $2\pi = 6.28$
F = frequency, in Hertz
L = inductance, in Henrys

Note that X_L is opposite to X_C in that it varies directly with frequency and inductance. If the frequency goes up, X_L goes up. If the frequency goes down, X_L goes down. The same is true of inductance. If the inductance goes up, X_L goes up. If the inductance goes down, X_L goes down. Like X_C, X_L is also measured in Ohms and can also be used in Ohm's Law instead of resistance.

Inductors are opposite to capacitors in another way. A graph of the inductor voltage and the inductor current, is shown in the graph below. Note that the voltage reaches peak amplitude before the current does. The current lags the voltage by a phase angle of 90°.



As mentioned, inductors can be fixed or variable. Variable inductors are usually adjusted by moving the core up and down inside the center of the coil. This movable core is often called a slug.

Due to the self-generated CEMF, it takes a certain amount of time for current flow in an inductor to reach its maximum level. This time depends on the size of the inductor and the amount of resistance in the circuit. As with capacitors, this time is measured in time constants. The symbol for time constant is τ (Tau). One Tau for an inductor is calculated by:

$$\tau = L/R$$

Where R = the resistance of charging circuit, in Ohms, and L = the inductance, in Henrys. One time constant is the time it takes for the inductor current to reach approximately 63% of its full value. Typically it takes 5-time constants for an inductor to reach full current flow. Inductance is measured in Henrys. Inductors can be measured in millihenry, abbreviated mH and microhenry, abbreviated μH. It takes 1,000 mH to make one Henry and l, 000,000 μH to make one Henry.

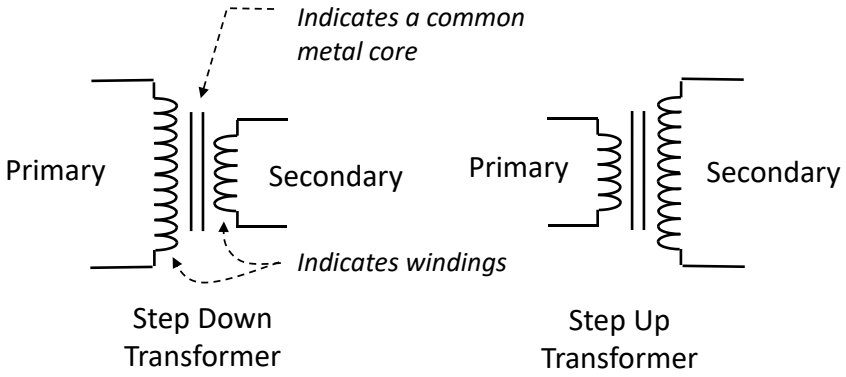
Saturation

As discussed, inductor performance depends on a moving magnetic field. For any given inductor, there is a certain coil current at which there will be no further increase in the magnetic field. Without a changing magnetic field, there is no CEMF and hence no inductive reactance. At this point, called *saturation*, the

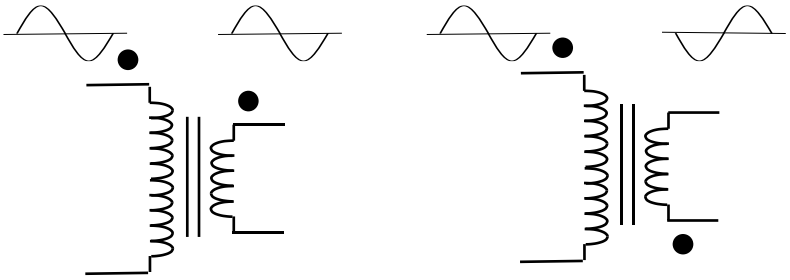
only resistance in the inductor is that offered by the resistance of the inductor wires themselves, which compared to the inductive reactance, is very small. At this point, the inductor no longer acts as an inductor; it just becomes a piece of low resistance wire. Because of this, the design process must assure that the inductor is operating below its saturation point.

Transformers

A transformer consists of two inductors wound on a common metal core. They are normally designated with a "T", e.g., "T1" or "T2". Transformers are used to magnetically couple two circuits together. The two inductors of the transformers are called windings. One winding is called the primary; the other winding is called the secondary. The ratio of the number of primary windings to the number of the secondary windings determines the magnitude of voltage across the secondary winding. If there are more turns of wire around the secondary than the primary, the output (secondary) voltage will be greater than the voltage across the primary. This is a step-up transformer in that it "steps up" the output voltage. Since the total power across the transformer must remain the same (disregarding transformer losses) if the secondary voltage of the step-up transformer increases, the secondary current must decrease since power = voltage x current and power remains constant (again, ignoring transformer losses). If there are fewer turns of wire on the secondary than the primary, the output (secondary) voltage will be less than the primary voltage (and the current more). This is a step-down transformer. If the number of windings is the same on the primary and secondary, then the voltage on the secondary is the same as the voltage on the primary (in an ideal transformer with no losses). Such a transformer is used to isolate two circuits. The sketch below illustrates the schematic symbol for a step-down and a step-up transformer.



The secondary output can be in phase or 180 degrees out of phase with the primary. The black dots on the schematic symbol are used to indicate the phase relationship. Refer to the sketch below. In the left-hand sketch, the black dot on the primary and secondary windings are both on top, indicating the top leads of both primary and secondary are in phase as indicated by the two sine waves in sync. In the right-hand sketch, the black dot on the primary is on top and the black dot on the secondary is on the bottom, indicating the *top* lead of the primary and *bottom* lead of the secondary are in phase. Not that the sine wave shown on the *top* lead of the secondary is out of phase with the *top* lead of the primary, as would be expected.



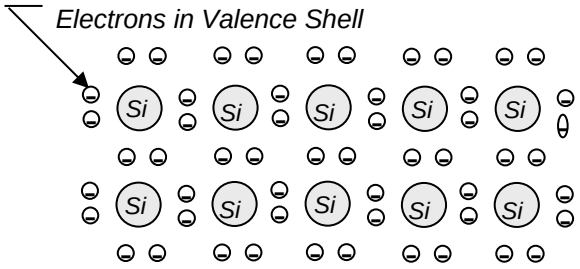
As mentioned, the ratio of the primary to secondary windings determines the output voltage on the secondary for a given voltage on the primary. If the number of windings on the primary and secondary are known as well as the input voltage on the primary, then the output voltage on the secondary can be calculated as follows –

$$V_S = V_P \frac{N_S}{N_P}$$

Where V_S = secondary voltage
V_P = primary voltage
N_S = number of secondary windings
N_P = number of primary windings

Chapter 6: Semiconductors

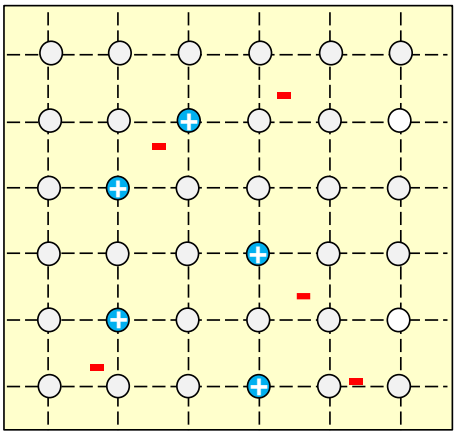
Semiconductors are made up mostly of silicon. Each silicon atom is made up of a nucleus with negative electrons positioned around it. The nucleus is made up of protons, which have a positive charge, and neutrons, which have no charge. Since the number of electrons equals the number of protons, the positive and negative charges cancel each other and the silicon atom has no net charge. The electrons around the nucleus are at different energy levels which we can think of as “shells”. A detailed discussion of this is beyond the scope of this paper so we will confine our discussion to the “outermost” shell called the valence shell. This is where reactions with other atoms occur to form different compounds. Silicon has four electrons in its valence shell but it likes to have eight. The silicon atoms accomplish this by lining up so as to share one electron from each of its four neighboring atoms. By sharing electrons, each atom thinks it has eight electrons in its valence shell. This form of bonding is called covalent bonding and it forms a crystalline structure as shown below.



Now suppose an atom with *five* electrons in its valence shell is added to the silicon crystal. This atom would line up with the rest of the silicon atoms and share four of its electrons in its valence shell. The fifth electron, having no covalent bond, would roam all about as a free electron. Such an atom is called a donor because it "donates" a free electron. However, once it does this, the positive protons now out-number the negative electrons in the donor atom and it has a net positive charge. Because the donor atom now has a charge, it is called an ion. And because it is locked into the silicon matrix it does not move. The process of placing the donor impurity in the silicon is called doping. A bar of silicon doped in this manner is called an N-type semiconductor where the "N" stands for the free, negative electron.

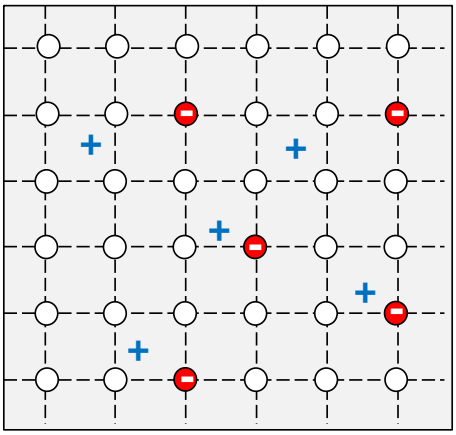
Now suppose an atom with *three* electrons in its valence shell is added to the silicon crystal. This atom will line up with the other silicon atoms as well. But it only has three electrons to share instead of four. To make up the fourth electron, it pulls an electron away from a neighboring silicon atom. That silicon atom now has a "hole" in its valence shell and the number of protons and electrons are no longer equal. With one electron gone, there is one more positive proton than negative electrons and that silicon atom now has a net positive charge. The silicon atom that lost an electron will take an electron from a second silicon atom. Now the second silicon atom has a hole and a net positive charge. In this way, the positive "hole" roams around just as the free electron did in the N-type semiconductor. Such an atom with only three electrons in its outer orbit is called an acceptor impurity because it accepts or takes electrons from other atoms. However, once it does this, the negative electrons now out-number the positive protons in the acceptor atom and it has a net negative charge. It becomes an ion, too. And because it is locked into the silicon matrix it does not move. A bar of silicon doped in this manner is called a P-type semiconductor. The "P" stands for the free, positive hole. These are illustrated below.

N – Type Semiconductor



(donor has 5 valence electrons)

P – Type Semiconductor



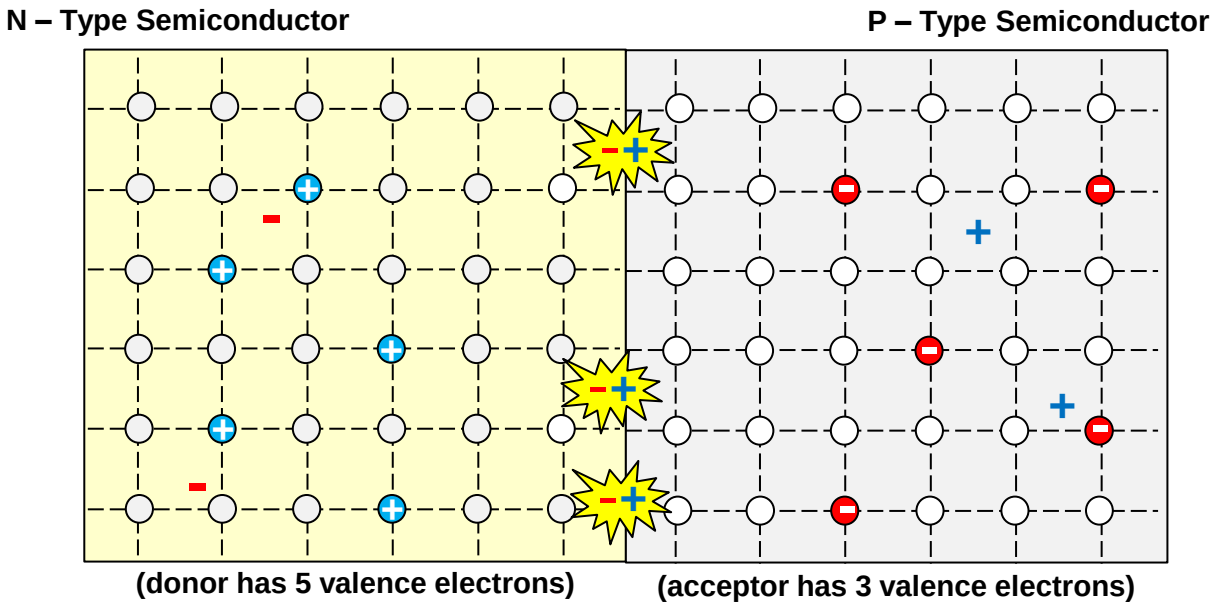
(acceptor has 3 valence electrons)

In each diagram the white circles represent the silicon atoms locked in a crystal matrix by covalent bonds, represented by the grid of dotted lines. In the N-type material, the donor atoms are represented as blue circles with a "+" sign inside, denoting their positive charge. The free donor electrons are represented as the red, free-floating "-" signs. In the P-Type material, the acceptor atoms are represented by the red circles with a "-" sign inside. The acceptor holes are represented as the blue, free-floating "+" signs. The free electrons and holes are the majority carriers.

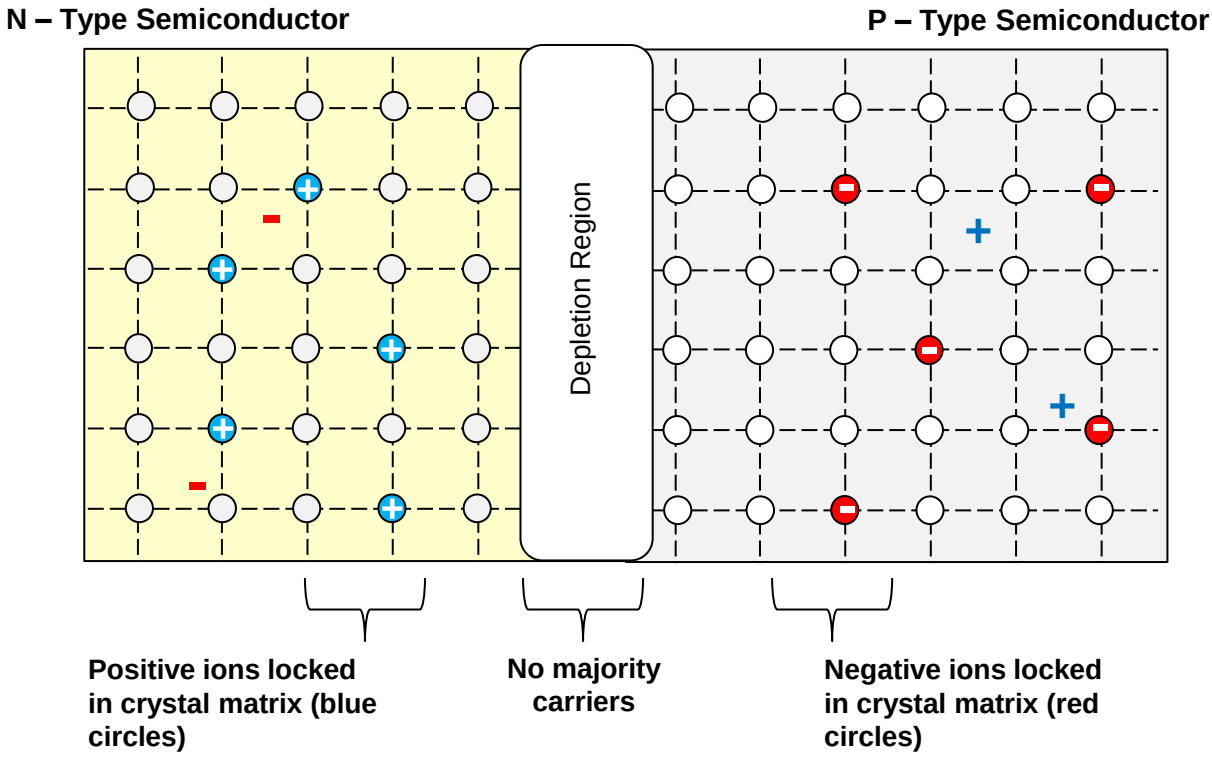
There are some minority carriers as well. Due to natural impurities in the silicon and thermal agitation, the N-type material has some positive "holes" roaming free and the P-type material has some negative free electrons. These are small in number and will not be discussed further.

P – N Junctions

What happens an N-type and P-type semiconductor are brought together. Refer to the figure below.



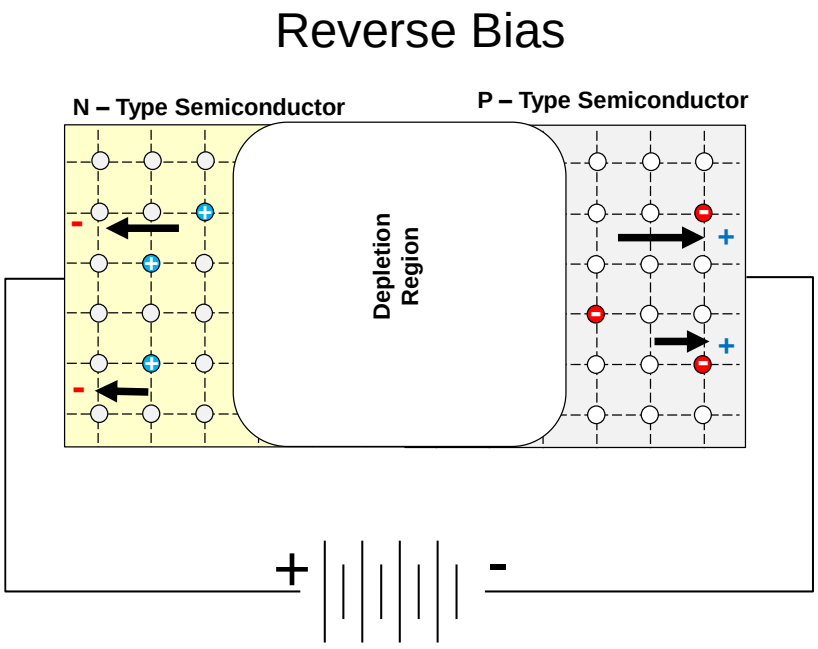
The free negative electrons in the N-type material and the free positive holes in the P-type material are drawn towards each other since they are unlike charges and unlike charges attract. When they collide in the middle, they recombine and cancel each other out. The net result is shown in below



The junction of the two semiconductors is now deplected of majority carriers due to the recombination of free electrons and holes. This is known as the depletion region. The fixed positive ions in the N-type material (blue circles with "+" sign) repel any positive holes (blue "+") from the P-type material. The fixed negative ions in the P-type material (red circles with "-" sign) repeal any negative electrons (red "-") from the N-type material. The net result is there is no current flow across the junction.

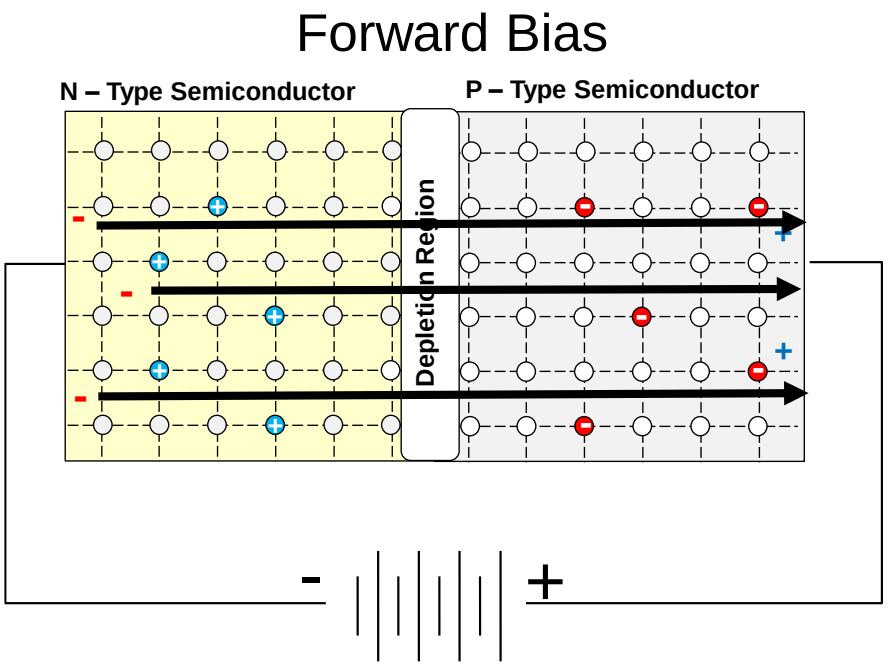
P – N Junction Bias

Consider now a battery placed across this N-P semiconductor with the polarity as shown below.



Note that the battery's positive terminal is connected to the N-type material which attracts the free negative electrons even further away from the N-P junction. The battery's negative terminal is connected to the P-type material which attracts the free positive holes even further away from the junction. This increases the depletion region further ensuring no current flow due to majority carriers. This condition is called reverse bias.

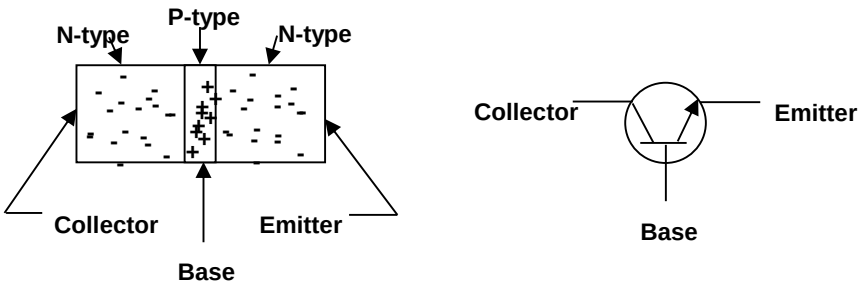
Now reverse the battery polarity as shown below.



With the negative battery terminal connected to the N-type material, the free negative electrons are repelled across the P-N Junction and attracted to the positive battery potential on the P-type material. They go so fast that they fly past the fixed ion's that used to repel them. The net result is that there is now current flow due to majority carriers. This condition is called forward bias.

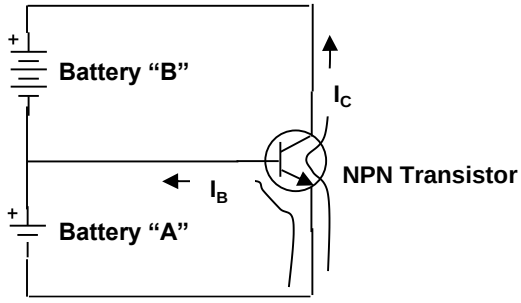
Chapter 7: Transistors and Transistor Amplifiers

A thin slice of P-type semiconductor is placed between two pieces of N-type semiconductor to form an NPN transistor. One of the N-type semiconductors is called the emitter; the other is called the collector. The P-type semiconductor in between is the base. This structure is illustrated in the figure below along with the schematic symbol for an NPN transistor.



(A) NPN Transistor Structure **(B) NPN Transistor Symbol**

Consider now the NPN biasing configuration as shown below. Note that the voltage of Battery "B" has a greater positive voltage than battery "A". This makes the collector more positive than the base. Thus, the base to collector (B-C) junction is *reverse* biased. Note also that battery "A" has the base to emitter (B-E) junction *forward* biased. The negative voltage on battery "A" pushes the electrons *across* the PN junction into the base. The more positive terminal of battery "B", connected to the collector, pulls most (about 95%) of the electrons across the base to the collector. The remaining 5% leave the base and go to the positive terminal of battery "A".



If we can increase the base current, I_b, we increase the current flow to the collector (I_C); if we decrease I_B, we decrease I_C. We can increase the base current, and thus increase I_C, by increasing the forward bias voltage across the B-E junction. Conversely, decreasing the forward bias decreases the base current and I_C. Thus, a small change in base current causes a large change in collector current. Because of this, the transistor is a current amplifier and has gain. The current gain of a transistor is called Beta. The formula for Beta is:

$$\text{Beta, } \beta = \text{delta } I_C \text{ / delta } I_b$$

Where delta I_C = the change in collector current
Delta I_b = the change in base current that caused the change in collector current

For short, it is often written as

$$\beta = I_C / I_b$$

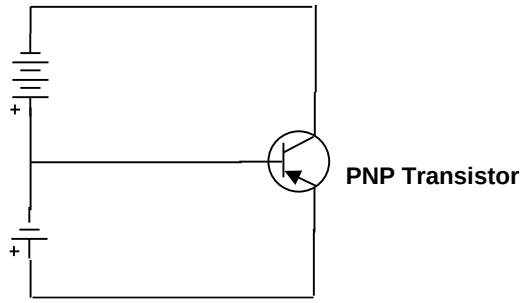
For example, suppose an NPN transistor had a current gain or Beta = 10. Suppose that I_b increased by 1 milliamp (a milliamp, abbreviated mA, is 1/1000 of an amp). With a Beta of 10 and a 1 mA increase in I_b, then I_C would increase by I_B x 10 = 10 mA. This is how a transistor amplifies. Transistors can have betas from 10 to 10,000. Note that on transistor data sheets, h_{FE} (note the capital "FE") may be used, which is forward DC current gain. For our purposes, we can consider beta and h_{FE} essentially the same.

Let's summarize how an NPN transistor works:

1) To forward bias an NPN transistor, the Base - Emitter junction must be *forward* biased (base voltage must be positive with respect to the emitter) and the Base -Collector junction must be *reverse* biased (collector voltage must be positive with respect to the base).

2) The Base-Emitter forward bias controls the amount of base current and the base current controls the amount of collector current, by a factor equal to Beta. Thus, as forward bias increases, the base current increases which increases the collector current.

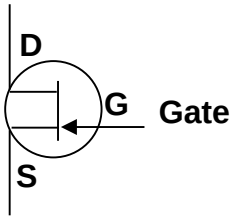
The most common transistor is the silicon NPN. But we can also make a transistor by putting an N-type semiconductor between two pieces of P-type semiconductor, forming a PNP transistor. It operates the same as an NPN transistor, only the polarities of the bias voltages are reversed. The Base-Emitter junction is still forward biased and the Base-Collector junction is still reverse biased. This configuration is illustrated below along with the schematic symbol for a PNP transistor. On schematics, transistors, either NPN or PNP, are identified by the letter "Q". For example: Q1, Q2.



It must be pointed out that the circuits above are grossly simplified to emphasis transistor operation. In a real circuit, resistors are used to limit current flow, establish bias voltages, and control impedance. Multiple batteries are not normally used.

FETs

A FET is a field effect transistor. They are characterized by a single "bar" of semiconductor material with one end being the source and the other the drain. This "bar" could be P-type or N-type, but for purposes of this discussion, we will assume an N-type, A small "button" of P-type material is injected into the side of the bar but does not go all the way through. In this way all the current flow is through the single bar of N-type semiconductor, in a channel from source to drain rather than through an N-type emitter, P-type base, and N-type collector as with an NPN transistor. Further, the source to gate is reversed biased so there is no gate current. As the reverse bias is varied, the depletion zone (area of no carriers) is varied across the channel. When the point is reached that the depletion zone goes completely across the channel, the channel is "pinched off" and there is no current flow. The schematic symbol for such a device is shown below.



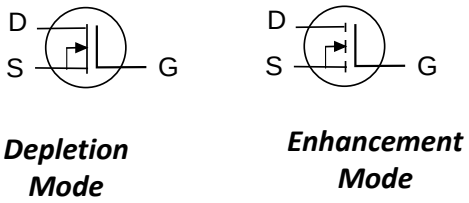
The source would be negative and the drain positive so the electron flow is from source to drain. The gate to source is reverse biased. Since there is no – or little - current flow in the gate, the input impedance is high.

Another type of FET is the metal oxide semiconductor or MOSFET. In this case, the gate is well insulated from the body of the semiconductor. The voltage on the gate controls the amount of current flow by virtue of its electrostatic field. Recall that like charges repel. If the gate has a negative voltage on it, the gate repels the negative electrons in an N-type semiconductor body thus "pinching off" current flow from the source to drain. Thus, as a negative voltage is applied to the gate, the current flow from source to drain decreases. The higher the negative voltage, the less the current flow. This type of MOSFET is a depletion mode MOSFET. Another type is the enhancement mode MOSFET. This one is normally off, with voltage on the gate "creating" a channel for current flow.

An N-channel, enhancement mode MOSFET has a high input impedance, is normally off, and requires a positive voltage to turn on. An N-channel, depletion mode MOSFET has a high input impedance, is normally on, and requires a negative voltage to turn it off.

Symbols for depletion mode and enhancement mode MOSFETs are shown below:

N channel MOSFETs

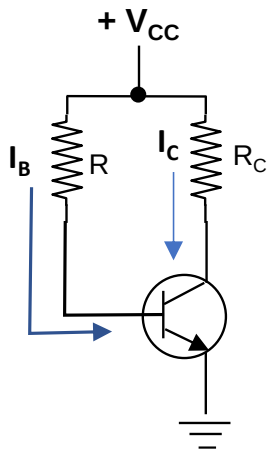


Depletion Mode ***Enhancement Mode***

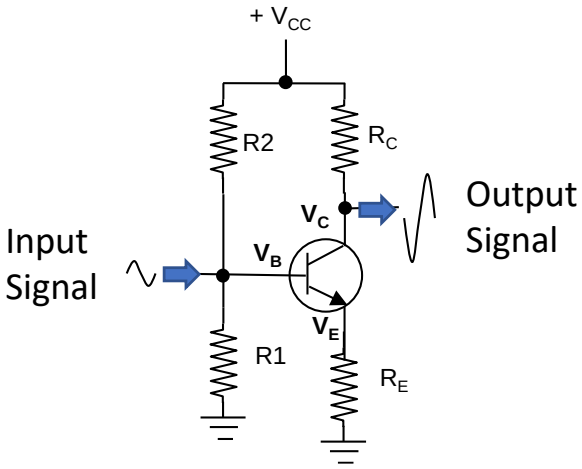
Transistor Biasing Design Considerations

Bias Stability. A simple bias configuration using resistors and a single power supply, V_{CC}, is shown below for an NPN transistor. R_C sets the collector current, I_C, and R sets the base current in accordance with:

$$I_B = I_C / \beta$$



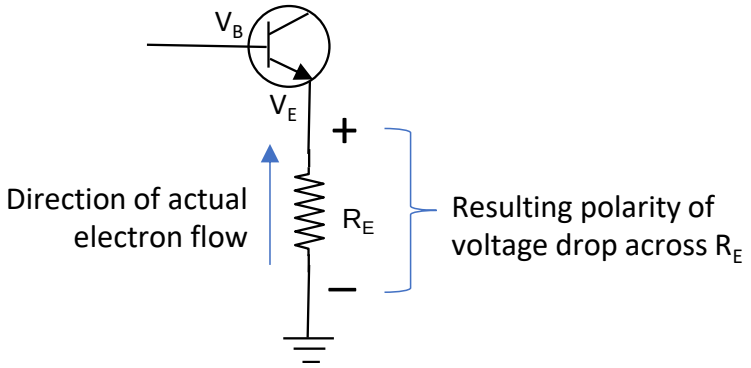
With R ad R_C selected appropriately, this would technically meet the criteria for "proper" biasing, as discussed above, but it would not be a very useful circuit. The base current would be subject to variations due to heating, aging, and production run variations of the transistor. This would affect the collector current, I_C. A more stable configuration is shown below:



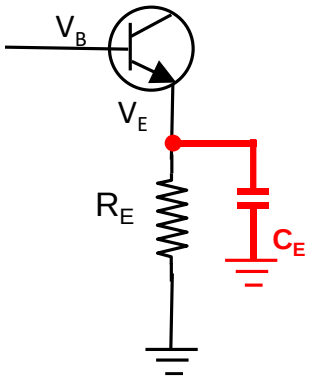
To keep the Base-Emitter forward bias stable, we keep the base voltage, V_B, stable, by *assuring that the current through R1 and R2 is at least 10 x I_B.* This keeps the base voltage, V_B, reasonably impervious to changes in I_B that occur with heat, aging, and production variability of the transistor. The ratio of R1 to R2 is chosen to assure the Base-Emitter junction is forward biased, that is, the base is more positive than the emitter.

Degenerative Feedback. R_E provides degenerative feedback which further stabilizes the circuit. Should the transistor base current increase, say due to heating, then the collector current would increase in accordance with the equation for β. The emitter current is essentially the same as the collector current so an increase in the collector current also means an increases in the emitter current through R_E. In terms of actual electron flow, the electrons would flow from the ground through R_E to the emitter, as shown in the partial schematic below. This would make the *emitter* voltage, V_E, more *positive* due to the increased voltage drop across R_E, in accordance with Ohms law, in this case $V_E = I_E \times R_E$. This would *reduce* the forward

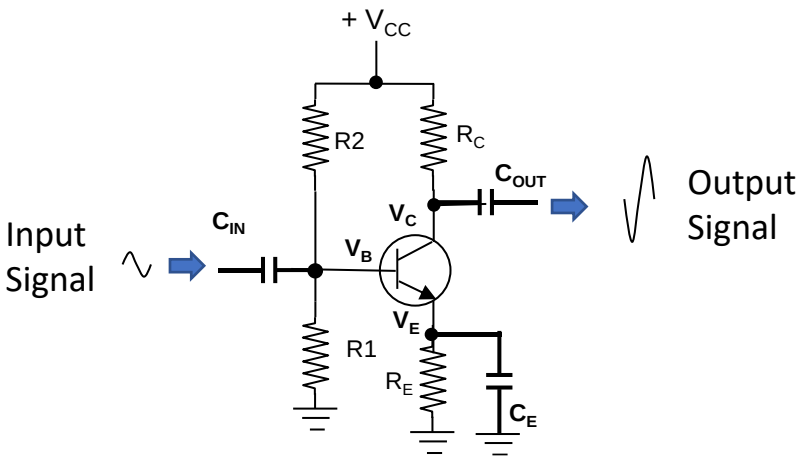
bias across the Base-Emitter junction which would tend to offset the base current increase due to junction heating and thus return I_B to a stable value.



To further maintain stability, a capacitor, C_E, is placed across R_E as shown below. This allows the AC signal to be passed around R_E through the capacitor, C_E. Without this, the AC signal being amplified would cause the voltage across R_E to go up and down in step with the AC signal. This would cause the transistor biasing to vary as well.



Capacitive Decoupling. In addition, this amplifier circuit would normally have a preceding circuit providing the input signal and a following circuit driven by the output signal. Each of these circuits would have their own DC biasing scheme. To block the DC voltage from these two circuits from reaching our amplifier, and interfering with our DC biasing, the input and output would have decoupling capacitors, C_{IN} and C_{OUT}. These two capacitors block any stray DC voltage from the previous or following circuits while allowing the AC signals to pass. The final configuration for a transistor amplifier, with these improvements, is shown below-



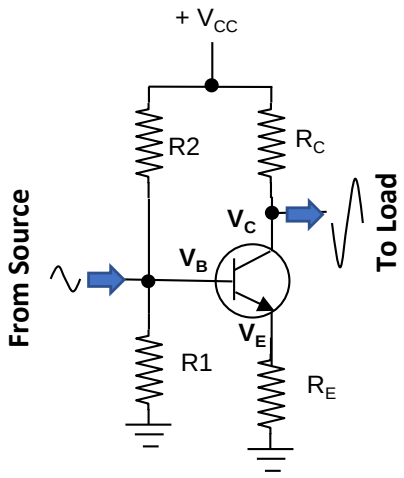
Finally, note that in this configuration, the input and output are out of phase. As the input signal goes positive, it increases V_B which increases the Base-Emitter junction forward bias, causing an increase in base current, I_B. And in accordance with the equation for β, this will increase I_C., which, in turn, will increase the voltage drop across R_C. This will *decrease* the voltage at the collector, V_C, which is also the output signal. So an *increase* in the AC input signal causes a *decrease* in the AC output signal and vice versa.

CE, CC aka emitter follower, CB Configurations.

This amplifier configuration we have been using is known as a common emitter (CE) amplifier. It is one of three types; the others being Common Collector (CC) also known as an Emitter Follower and Common Base (CB). All three are used for different purposes. Their configurations and use are shown below. Of the three, the CE is probably the most common and is often used as a general-purpose amplifier, so we will focus our attention on this one.

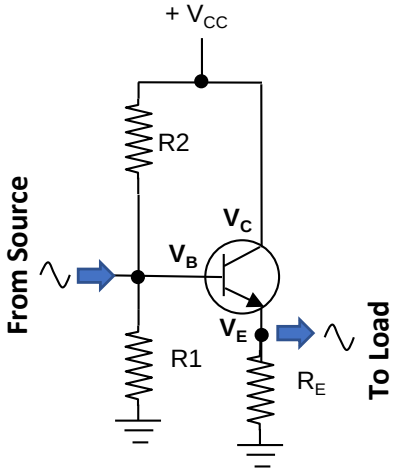
	CE	CC	CB
Input Impedance, Z_{IN}	*moderate*	*high*	*low*
Output Impedance, Z_{OUT}	*moderate*	*low*	*high*
Voltage Gain, A_V	*high*	*<1*	*high*
Current Gain, A_I	*high*	*high*	*<1*
Output Inverted	*yes*	*no*	*no*

CE



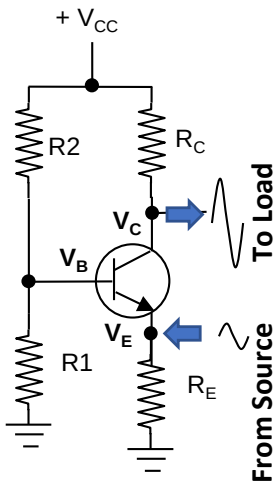
General purpose amplifier

CC



Normally used to connect high Z source to a low Z load

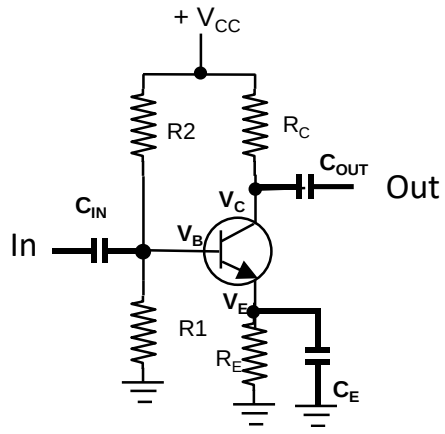
CB



Normally used to connect low Z source to a high Z load

Chapter 8: CE Amplifier Design Example

Consider a Common Emitter amplifier configuration as shown below. We will be using a 2N3904 transistor with the key specifications as shown.



- Max collector to emitter voltage, V_{CEO} = 40 VDC
- Max continuous collector current, I_C = 200 mA
- Total Power Dissipation, P_T @ T_A 25C = 500 mW

I_C, mA	h_{FE} (aka β)	
	minimum	maximum
1	80	300
10	100	300
50	60	300
100	30	300

Using this data, we can make some simplifying assumptions to set some upper bounds to work with. Often the power supply voltage available for V_{CC} is already known. Assume that V_{CC} = 12 VDC, which is well below V_{CEO}. So far, so good. To be conservative, we will derate the maximum power dissipation that we want to use to ½ P_T or 250 mW. Given a V_{CC} of 12 VDC and a desired maximum power dissipation of 250 mW, we can determine I_C max for our purposes. Recall that power = voltage x current. Solving for current we get-

I_C, Max = 0.250 W / 12 VDC = approx. 20 mA

We know we don't want to exceed 20 mA, but what current should we use for I_C and what would be the required I_B to produce it? Referring to the table above for β at different values of collector current, I_C, we see an entry at 10 mA. This is ½ our max desired I_C, so this seems a good choice. There are more detailed techniques to make this selection such as plotting a load line, but for our purposes, this will be good enough. The table shows that β can vary between a minimum of 100 to a maximum of 300 at I_C = 10 mA. So our required base current, I_B, can be as follows:

I_B, min = 10 mA / 300 = 33 μA
I_B, max = 10 mA / 100 = 100 μA

So, let's summarize where we are at this point: V_{CC} = 12 VDC, I_C = 10 mA, and I_B = 33 to 100 mA

So now we design a bias network to provide a collector current of 10 mA at a max base current of 100 μA, with a V_{CC} of 12 VDC using the steps outlined below:

Step1) V_E = 0.1 V_{CC} or you can just set it to 1 Volt

Step 2) $R_E = V_E / I_C$

Step 3) To maximize the output voltage swing, $R_C = (V_{CC} /2) / I_C$

Step 4) I_{R1} = 10 x max I_B (this assures the base bias voltage, V_B, is not changed by any undesired changes in the base current)

Step 4) I_{R2}= 11 x max I_B (this includes I_{R1} and I_B)

Step 5) $V_B = V_E + V_{BE}$ (V_{BE} is a nominal 0.65 VDC for an NPN transistor)

Step 6) $R1 = V_B / I_{R1}$

Step 7) $R2 = (V_{CC} – V_B) / I_{R2}$

Now add the capacitors using the steps below

Step 1) Add R_E bypass capacitor C_E. This keeps the AC signal from changing the bias point. Select C_E so its capacitive reactance is 1/10 the value of R_E at the lowest frequency the amplifier will operate at … say 500 Hz

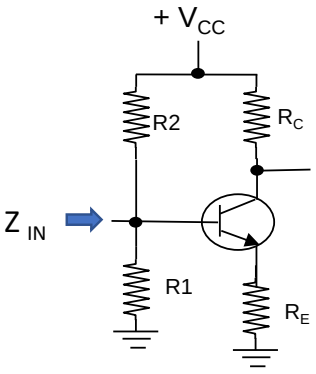
$$C_E = 1 / (2 \pi f (0.1 R_E))$$

Step 2) Add decoupling capacitor C_{IN}, which keeps the base bias voltage from being changed by the DC voltage on the previous stage. Select C_{IN} so its capacitive reactance is 1/10 the value of input impedance, Z_{IN} , at the lowest frequency the amplifier operates at … say 500 Hz. See the panel below to determine Z_{IN}.

$$C_{IN} = 1 / (2 \pi f (0.1 Z_{IN}))$$

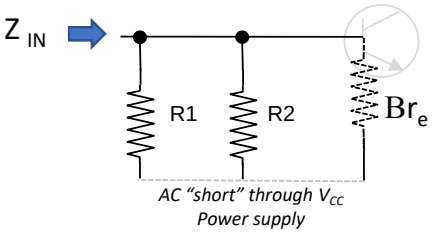
Step 3 Add decoupling capacitor C_{OUT}, which keeps the base bias voltage from this amplifier from changing the bias on the next stage. Select C_{OUT} so its capacitive reactance is 1/10 the value of Z_{OUT} at the lowest frequency the amplifier will operate … say 500 Hz. Note that Z_{OUT} = approximately R_C when there is no load on the transistor output. If there is a load. Z_{OUT} = approximately R_C || R_L.

Input Impedance, Z_{IN}, for a Common Emitter Amplifier



For AC signals, a power supply appears as a short. So for AC, R1 and R2 are in parallel via the "short" through the V_{CC} power supply. For AC, the resistance of the transistor base can be approximated as βr_e where $r_e = 26 / I_C$, ***in mA*** (r_e is the dynamic resistance of the transistor)

Thus, for Z_{IN}, this circuit reduces to R1, R2, and an equivalent "resistor", βr_e, in parallel. Note that "||" means "in parallel with".



$$Z_{IN} = R1 \,||\, R2 \,||\, \beta r_e$$

$$Z_{IN} = \frac{1}{\frac{1}{R1} + \frac{1}{R2} + \frac{1}{\beta r_e}}$$

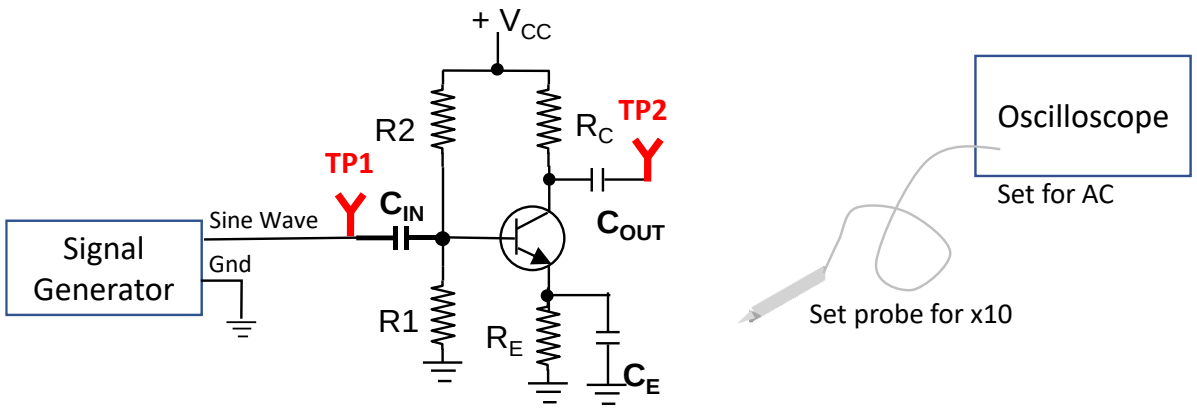
Analysis and test

Calculate the amplifier AC gain, A_V, using the following equation:

$$A_V \sim - \frac{R_C \,||\, R_L}{r_e}$$

- where $r_e = 26 / I_C$, in mA (r_e is the dynamic resistance of the transistor)
- "||" means "in parallel with"
- R_L is the input impedance of the next stage

If there is no R_L, just use R_C. Measure the actual gain by connecting a signal generator to the input as shown below. Set up the oscilloscope to read AC with a x10 probe. Adjust the probe for x10 (this reduces the probe loading on the circuit).



Connect your probe to TP2 and experiment to find the point of maximum gain with no clipping by adjusting the signal generator in amplitude and frequency. Record the peak voltage reading on the oscilloscope as V_{OUT}. Some scopes will tell you what the signal voltage is in V rms, Volts peak (V_P), or Volts peak to peak (V_{PP}); some will not, requiring you to "read" it using the gradients on the display. *It doesn't matter which you choose (rms, peak to peak, or peak) as long as you use the same thing throughout the test*. Move the probe to TP1 and record the input voltage in as V_{IN}. The voltage gain, A_V, = V_{OUT} / V_{IN}. Compare this to your calculated A_V. Remove capacitor C_E and see what that does to the voltage gain.

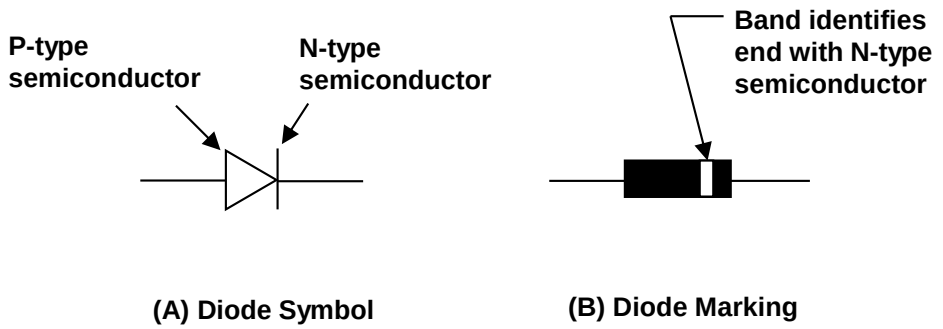
Next, determine the frequency bandwidth of your amplifier. From the point of maximum gain, step the signal generator *up* in frequency, in increments, until the voltage at TP2 drops to a value equal to V_{OUT} at max gain x 0.707. This point will be the amplifier upper frequency limit. At each increment, record the frequency, V_{IN} and V_{OUT}. Make sure V_{IN} remains constant throughout this test by adjusting the signal generator amplitude as required. Repeat this process going *down* in frequency until you find the lower frequency limit.

To assure repeatability, these tests should be repeated at least three times.

Chapter 9: Diodes

A diode is a semiconductor component composed of P-type and N-type semiconductor joined together. The diode will only conduct when the diode is forward biased.

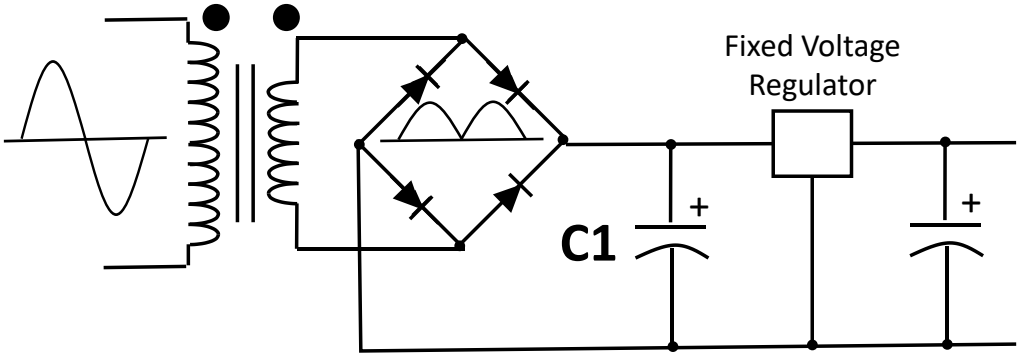
The schematic symbol for the diode is shown below. The vertical bar indicates the N-type semiconductor, and the P-type semiconductor is indicated by the triangle. Since diodes are sensitive to the polarity of the applied bias voltage, we need to be able to identify which end of the diode is the N-type semiconductor. This is usually done by band around the end of the diode that is the N-type semiconductor, as shown in (B) below.



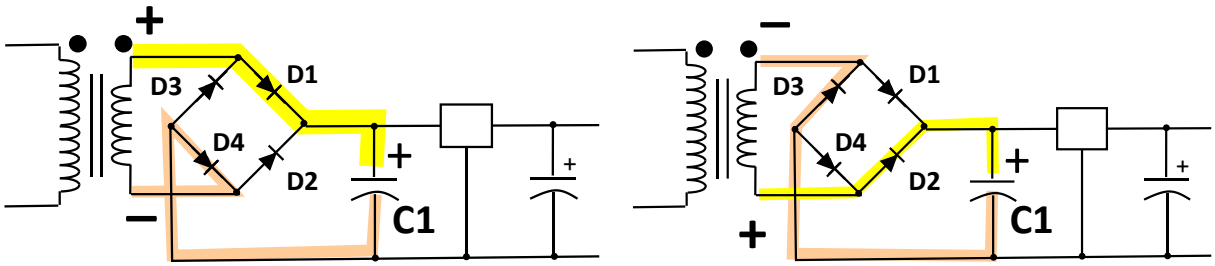
(A) Diode Symbol **(B) Diode Marking**

Since diodes conduct current in only one direction, they are used convert AC to DC. When the AC signal is applied to the diode such that the diode is forward biased, current will flow. When the polarity of the AC signal reverses, the diode is reversed biased and no current flows. In this manner the current changes from one that goes "back and forth" (AC) to one that goes in only one direction (DC). This is call rectification.

A common application for diodes is in DC power supplies. Consider the circuit below.

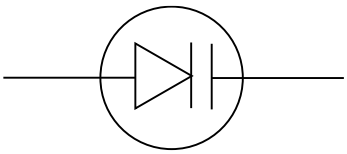


This is a full wave bridge rectifier with positive DC voltage fixed regulator. It is found in computers, TVs, and almost any electronics that runs on house hold AC. A step-down transformer drives four diodes arranged in a diode bridge as shown. The unique feature of the bridge rectifier is that it produces a positive pulse when the input VAC swings positive *and* when it swings negative. Thus, each cycle of the AC input is transformed into two pulses by the bridge rectifier. These are filtered by C1 to remove the ripple and produce a constant, unregulated DC voltage. C1 is very large (several hundred μF) electrolytic capacitor and has a polarity as shown. The unregulated DC voltage is fed to a fixed voltage regulator. This is an integrated circuit the maintains a fixed DC voltage with varying current demands. Typical values are +5 VDC, +9 VDC, +12 VDC, and +15 VDC. Fixed negative voltage regulators are also available. These IC's deal with a lot of power and often require heat sinks to help dissipate the heat. The two sketches below illustrate how the full wave bridge rectifier works.



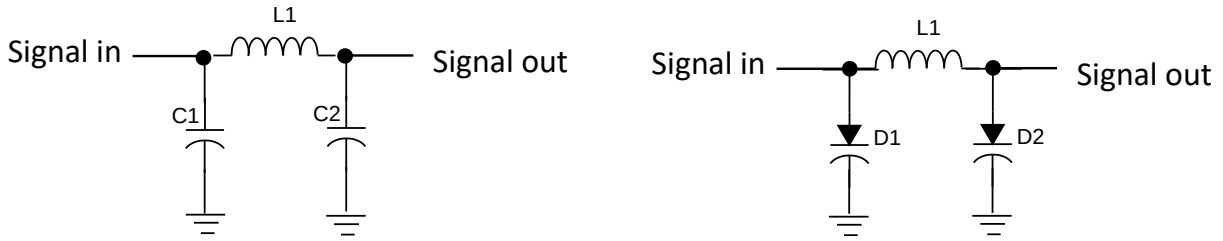
When the top of the secondary is positive, diodes D1 and D4 are both forward biased and conduct as shown in the left-hand sketch. This applies a positive pulse to positive side of C1. When the top of the secondary is negative, diodes D3 and D2 are now forward biased and conduct, applying another positive pulse to the positive side of C1.

Some diodes are designed to act as variable capacitors. Such diodes are called varactor diodes. Recall that a capacitor has two metal plates separated by an insulator called a dielectric. When a varactor diode is reverse biased no current flows across the PN junction. The PN junction depletion region acts as the dielectric and the P-type and N-type semiconductors act as capacitor plates. Changing the amount of reverse bias voltage can vary the width of the depletion region which varies the amount of capacitance. This allows us to change the capacitance of the varactor by changing the voltage applied to it. The schematic symbol for a varactor diode is shown below. On a schematic, the letters “CR” or “D” may identify diodes. For example: CR1, CR22, D4.

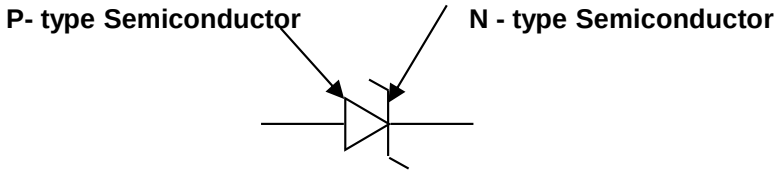


Varactor Diode

One application for varactors is in electronically tunable filters. Consider the circuit shown below. The left-hand schematic is a traditional low pass filter with two capacitors and an inductor. As the “Signal In” goes up in frequency, the capacitive reactance of C1 and C2 goes down, shunting the signal to ground and reducing the level of “Signal Out”. Also as the frequency goes up the inductive reactance of L1 goes up, blocking “Signal In” from reaching “Signal Out”. The net effect of C1, C2, and L1 is to pass low frequencies and block high frequencies. The exact frequency at which attenuation begins to increase is determined by the component values. This a low pass filter. Now consider the right-hand circuit below where C1 and C2 have been replaced with varactor diodes. By varying the reverse voltage across the varactor diodes, the capacitance can be changed, which varies the frequency at which attenuation sets in. Thus, we have an electronically tunable low pass filter.

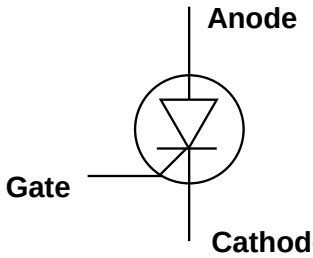


A Zener diode also operates in reverse bias. It is designed to break down and clamp at a specified voltage. That voltage can range from *5* to 50 volts. Zener diodes are often used in conjunction with a resistor as a simple voltage regulator. Although it does not maintain as good a voltage regulation as the previously mentioned fixed voltage regulator IC, it still is useful as a voltage regulator in less demanding situations. It can also be used as a voltage clamp to control signal amplitudes. The schematic symbol for a Zener diode is shown below.

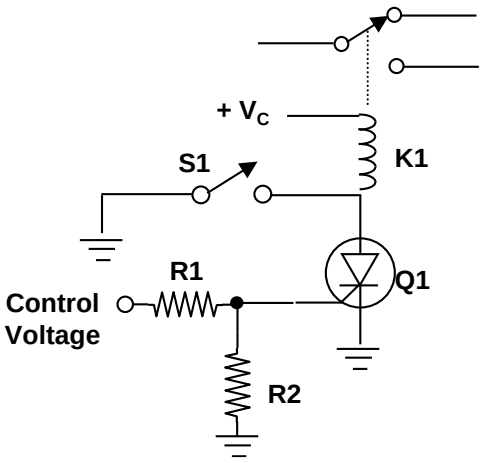


Symbol for Zener Diode

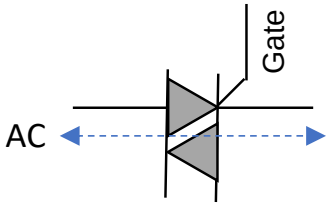
An SCR is a silicon-controlled rectifier, which operates on direct current only. The schematic symbol for it is shown below. Note that it is a diode with a third lead added, called a gate. The SCR does not conduct until the voltage at the gate goes positive by specified amount. When the gate reaches that point, the SCR conducts. Once it starts conducting, the SCR will not turn off even if the gate voltage is removed; the SCR has latched. Once an SCR has latched, the only way to turn it off is to break the current flow through the anode and cathode.



The figure below shows a typical SCR circuit. When the control voltage goes positive, the SCR begins to conduct and latches. This energizes relay K1. K1 will remain energized after the control voltage is removed. To turn off the SCR, and de-energize K1, push in pushbutton S1. This puts a ground on the SCR anode, stopping current flow through the SCR. This unlatches the SCR off and de-energizes K1.



A similar device that operates on alternating current is a Triac. The schematic symbol is shown below.

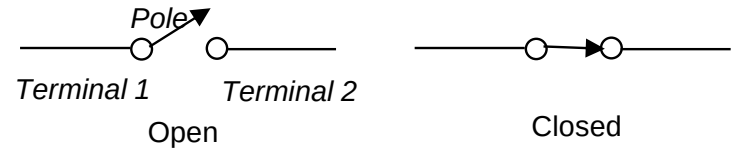


Note that like the SCR, the Triac has a gate as well. Unlike the SCR, the Triac does not latch. Instead, varying the Gate current will vary the alternating current (AC) that passes through it. Thus it can be used to control motor speed, dim lights, etc.

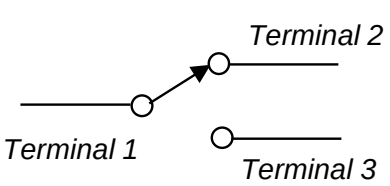
Chapter 10: Hardware

A switch is a mechanical device that opens or closes a circuit. All switches have an arm, called a pole that can connect or open two terminals on the switch. A switch with only two terminals and one arm is called a single pole, single throw switch. It is abbreviated SPST and is shown schematically in the figure below (a). Note that it has only one arm or pole. In the closed position, the pole connects terminals 1 and 2 together. In the open position, the pole disconnects terminals 1 and 2. If we add a third terminal to the switch, we now have a single pole, double throw switch, abbreviated SPDT shown in figure (b) below. Note that the pole can now connect terminal 1 to either terminal 2 or 3.

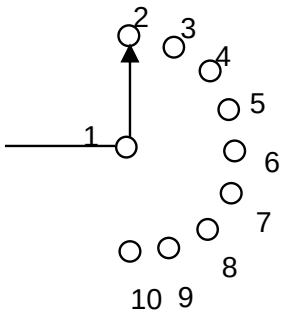
A rotary switch can connect one pole to several different terminals as shown on (c). Note that the pole can connect to any one of the other terminals on the switch depending on the rotational position of the pole. Sometimes two rotary switches are put together in a single switch. This is shown in (d). The dashed line between the two poles means the poles are ganged. This means pole “A” and “B” rotate together. If pole “A” were rotated to connect terminals 1A and 3A, pole “B” would rotate along with it to connect terminals 1B and 3B.



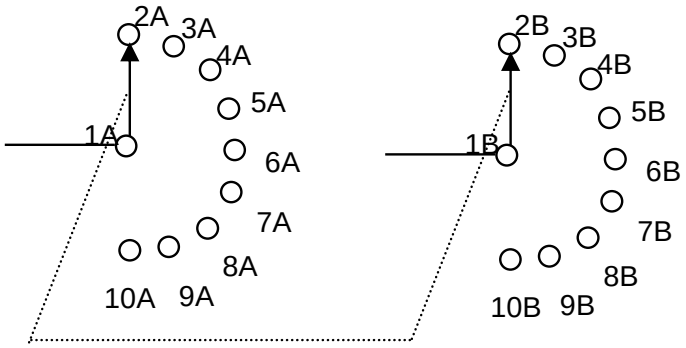
(a) SPST Switch



(b) SPDT Switch



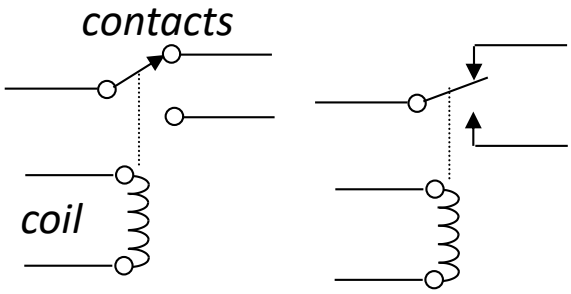
(c) Rotary Switch



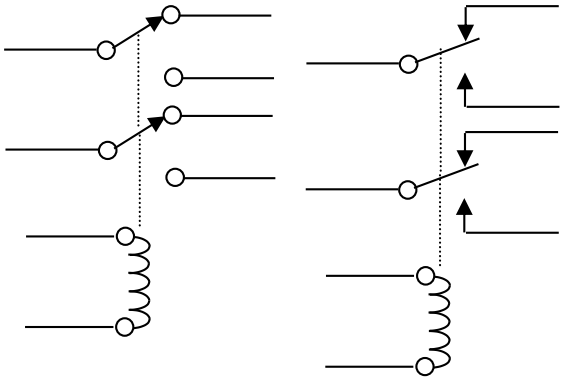
(d) Ganged Rotary Switch

Another switch is called a pushbutton switch. It can be SPST or SPDT and is spring loaded, meaning it has to be held in to move the pole. When it is released, the spring pushes the pole back to its original or rest position. The push-to-talk (PTT) button or transmit button on a radio is usually a pushbutton switch. Radio on/off switches are usually SPST switches. On a schematic, the letter “S” usually identifies switches. For example: S1, S4.

A relay does the same thing as a switch but by remote control. A relay basically consists of two parts: a coil of wire (the coil) and the metal switches (contacts). When a voltage is applied to the relay coil, current flows creating a magnetic field. This pulls down the metal switch moving the pole from one contact to another. Relay contacts can be SPST, SPDT, or they can be ganged. Two SPDT contacts ganged together are called a double pole, double throw. The schematic symbols for relays are shown below. On a schematic, the letter “K” usually identifies relays. For example: K1, K3.

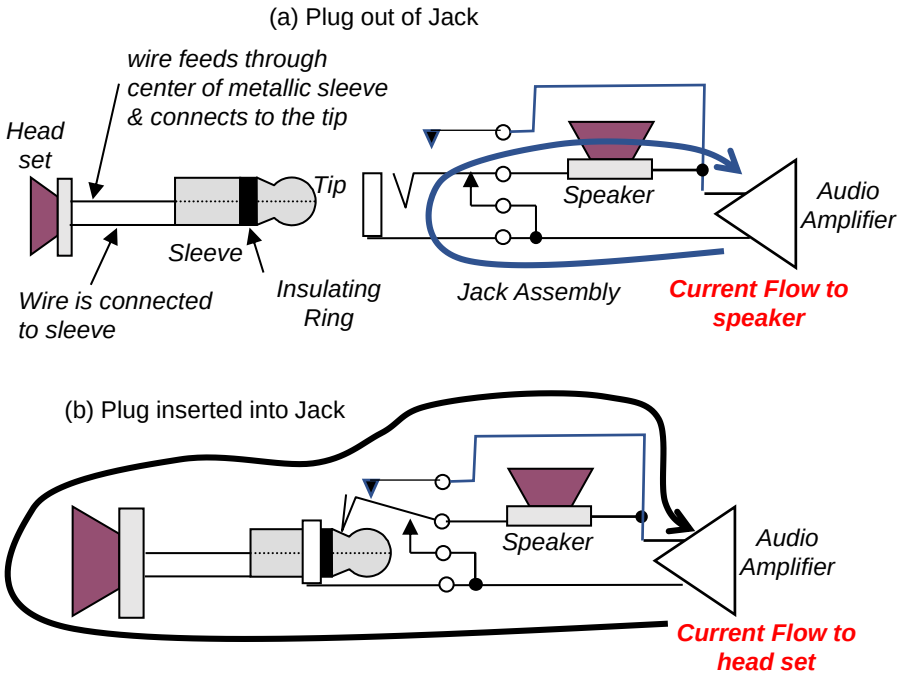


(a) SPDT Relay



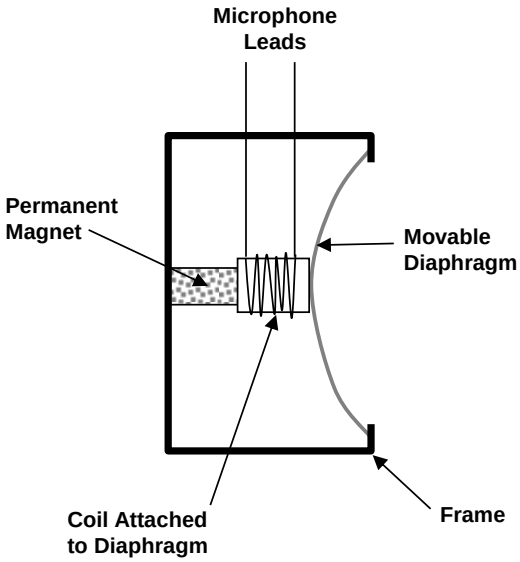
(b) DPDT Relay

A jack is a device that allows the insertion of an external piece of equipment into a circuit. For example, a jack on a head set will open the circuit to the speakers and change the path from the speaker to the head set, as shown below. On a schematic, the letter “J” usually identifies jacks. For example: J1, J5.



A fuse is a component that opens a circuit if too much current goes through it. It opens the circuit by causing a wire inside it to melt into two pieces. The figure below shows the schematic symbol for a fuse. On a schematic, the letter "F" usually identifies fuses. Fuses are used to protect against shorts or too much current draw.

A microphone is a transducer that changes sound energy into electrical energy. There are many types of microphones (abbreviated "mic"). The figure below illustrates a typical dynamic microphone.



Sound is caused by changes in air pressure at a rate that can be detected by the human ear. The movable diaphragm in the mic is connected to a small, suspended coil of wire that does not touch anything else except the diaphragm to which it is attached. In this way, both the diaphragm and coil can move freely back and forth. The coil moves within a magnetic field of a permanent magnet. Recall that if we move a wire in and out of a magnetic field, we will induce a current in the wire. When the sound pressures strike the diaphragm, the diaphragm moves back and forth, moving the coil back and forth in the magnetic field. This induces a current in the coil and a voltage appears at the output of the coil leads. This voltage varies with the movement of the diaphragm caused by the sound waves in both frequency and amplitude, thus converting the sound wave to an electrical audio signal.

A speaker is a transducer that changes the audio electrical signal to a sound wave. It is similar in construction to the dynamic mic, only instead of using the coil leads as an output, they are used as an input. The audio signal is applied to the microphone leads in the figure above. This applies the audio signal to the coil of wires (voice coil). As in the mic, the voice coil is attached to the flexible diaphragm and is suspended in the magnetic field of the permanent magnet. The voltage of the audio signal, applied to the voice coil, causes a current to flow in the coil generating a magnetic field around the voice coil. This magnetic field varies in frequency and amplitude with the audio signal. The magnetic field of the voice coil and permanent magnet interact to move the diaphragm back and forth creating sound.

Another type of microphone is a condenser (capacitor) mic. In this case, there are two metal plates just like in a capacitor; one is fixed, the other is movable. Sound, in the form of varying air pressure strikes the movable plate causing it to move slightly back and forth. This changes the capacitance which is detected by electronics embedded in the condenser mic, amplified, and sent out as an electrical audio signal. The device requires an external voltage to operate and has polarity. That is, one lead of the mic has to be connected to a specified positive voltage, the other to ground. An Electret microphone is such a device.

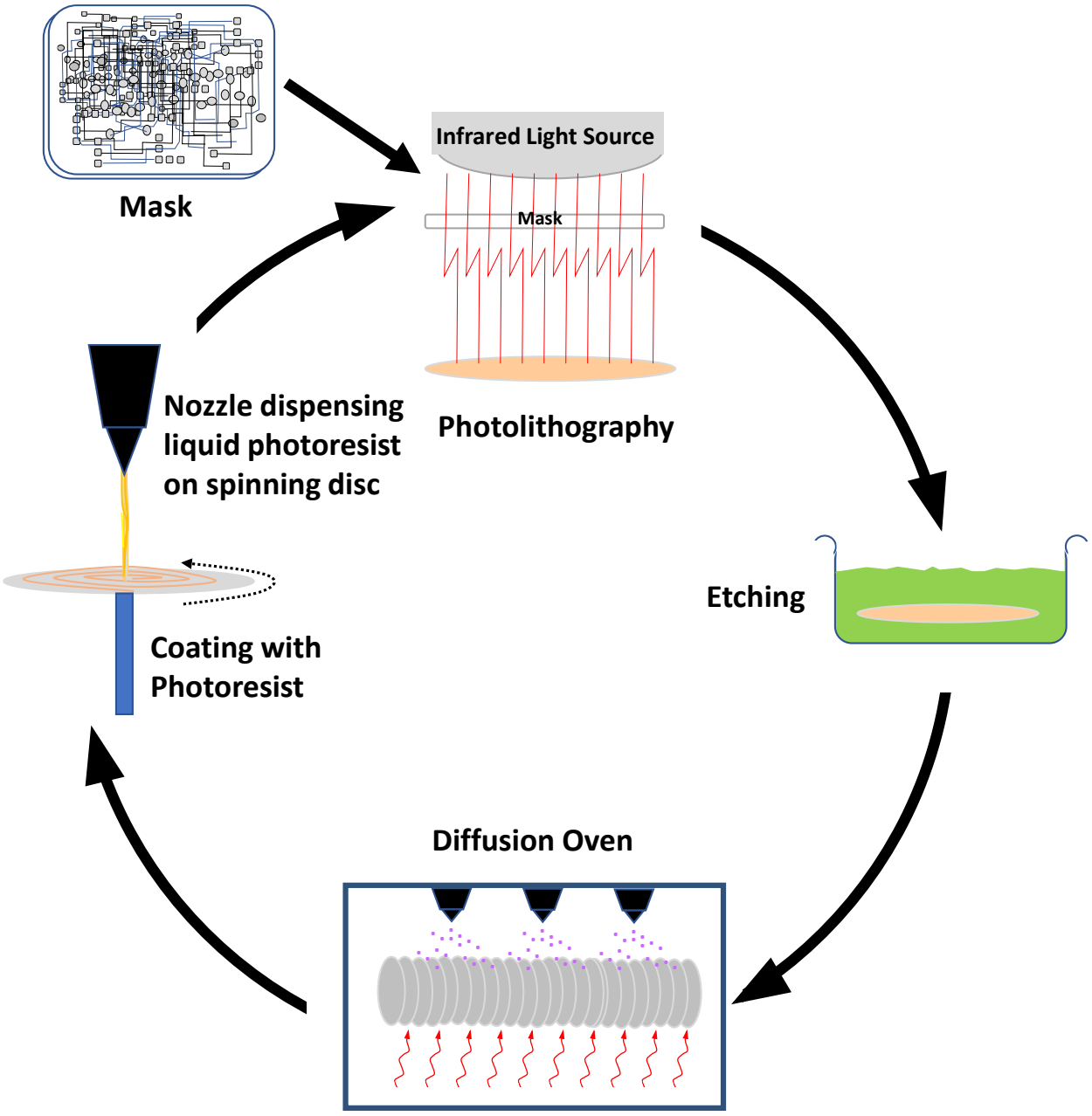
Chapter 11 Integrated Circuits Fabrication

Integrated circuits, abbreviated ICs, are made up of tiny silicon wafers that have a large number of separate transistors build in and arranged so as to perform logic functions or analog operations. Each IC, sometimes called chips, contains a tiny silicon die that has a complete circuit build right into the die. The die is then packaged or mounted inside a protective covering. The circuit in each chip may contain anywhere from 20 to 70,000 transistors.

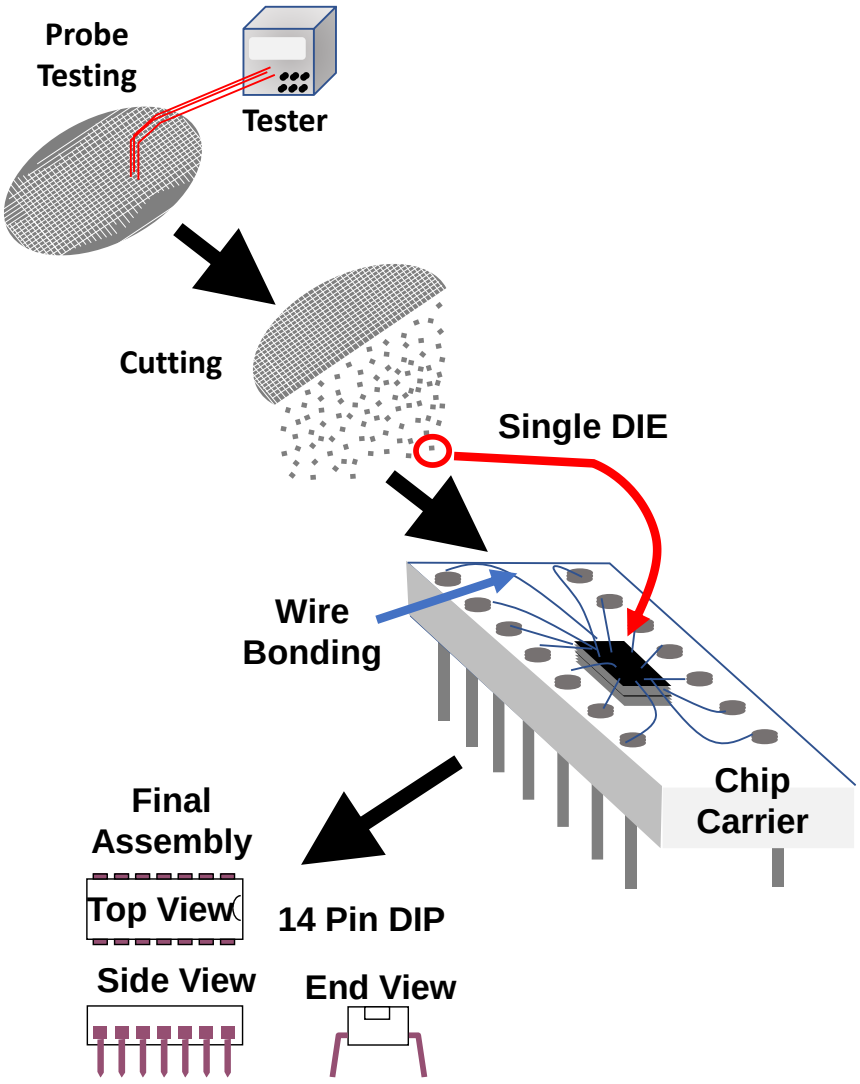
The basic manufacturing process is discussed in the following text.

Rods of silicon, 1 ½ to 4 inches in diameter, are sliced into thin, round wavers. Photoresist is applied to each waver by securing it to a vacuum chuck and dribbling liquid photoresist onto the center of the waver while it is spun at high speed. The spinning assures the photoresist is spread evenly across the waver. In the mean-time, design engineers have prepared glass masks that define the structure and placement of transistors, circuit traces, etc. for the IC. In the photolithography step, the mask is placed between an infrared light source and the waver covered with photoresist. The light is turned on, “imprinting” the mask design onto the waver photoresist. Where the light was blocked by the print on the mask, the photoresist

remains soft; where it was fully exposed to the light, the photoresist hardens. The waver is then place into a chemical bath which etches off the soft photoresist. The processed wavers then go to the diffusion ovens. Here, in contained enclosures, the wavers are exposed to gaseous dopants, such as arsenic (which by the way smells like garlic) and are heated to high temperatures. This diffuses the required dopants into the exposed silicon forming the desired components such as transistors. This process is usually repeated several times with different masks to build up all the required circuit layers. This process is summarized below:

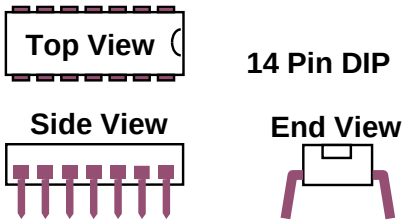


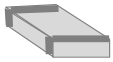
Finally, the waver undergoes a metallization process (not depicted in the figure) which adds the required metal connections and traces. The completed waver goes through probe testing. Tiny probes are used to access test points on the integrated circuits and automated testers use these probes to apply test signals and measure test results to verify the circuits are working. The waver is then scored and cut into individual chips or dies. Each die is the packaged onto a chip carrier and the external pins are connected to the appropriate points on the die by wire bonding. The top of the chip carrier is then secured making the final IC assembly. See below.



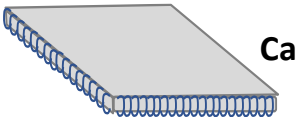
There are a number of IC chip carriers or packages and more are being developed every day. Some of the early technology was dual inline package or DIP. These could have 8, 14, 16, 24 or more pins in a “double row” configuration. This is often referred to as “pin in hole” technology because it requires a hole to be drilled through the printed circuit board for the pins to go through for soldering. Although this is old technology, it is still used today. Most modern IC packages use some form of surface mount technology

(SMT), allowing the IC to be soldered directly to the PCB without going through holes. This eliminates the drilling process during fabrication. The trend is toward smaller packages and finer pitch (lead spacing), allowing more and more leads per inch. Three examples – of many – are depicted below.





Resistor Surface Mount
1.2 mm x 1 mm



Plastic Leaded Chip Carrier (PLCC) 28 pins, 12 mm square

More modern ICs are packaged without leads and are soldered directly to the printed wiring board via solder "bumps" on the IC chip carrier.

There are many different kinds of IC's. Basically, ICs are either analog or digital. Analog ICs are designed to work with AC signals. They are sometimes called linear IC's. Generally, analog and linear refer to the same thing. Digital ICs are designed to work with computers or logic circuits and are characterized by internal components having only two states – on or off, often referred to as "high" or "low", logic "1" or logic "0", or just "1" or "0"'

Chapter 12 Integrated Circuits: Digital

Digital Logic

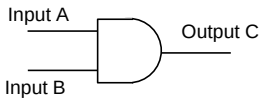
Digital logic has only two states - "on" or "off", low or high. The binary number system has only two numbers, 0 or 1. Therefore, digital ICs operate on the binary number system. Instead of "on" or "off", we refer to logic 1 or logic 0. Logic 1 can be +VDC and logic 0 ground (positive logic) or logic 1 can be ground and logic 0 +VDC (negative logic). Digital ICs are building blocks for larger logic circuits.

Basic Logic Gates

AND & NAND Gates. The schematic symbol for an AND gate is shown below on the left. It is called an AND gate because both input A *and* input B must be 1 before the output will go to 1. The truth table associated with the schematic symbol tells us how the AND gate works. Since there are two inputs, there are four possible combinations of inputs A and B: 00, 01, 10, and 11. These input combinations, along with the outputs generated, are shown on the AND gate truth table below. Row 1 on the AND gate truth table tells us that if A=0 (low) and B= 0 (low), the output, C, will be 0 (low). Row 2 tells us if A=0 (low) and B=1 (high), C=0 (low). Row 3 tells us if A=1 (high) and B=0 (low), C=0 (low). Row 4 tells us if A=1 (high) and B=1 (high), C=1 (high). Another gate is the NAND gate, shown below on the right with its truth table. This gate acts

the same as an AND gate only it inverts the output. Note the small circle on the output side of the NAND gate symbol. This indicates the output is inverted. Compare the truth table of the NAND gate with that of the AND gate and you will see that for any input combination, the output to the NAND gate is opposite or the inverse of the AND gate's output.

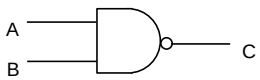
AND Gate



Truth Table - AND

Row	Inputs		Outputs
	A	B	C
1	0	0	0
2	0	1	0
3	1	0	0
4	1	1	1

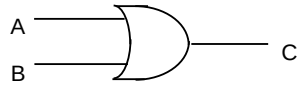
NAND Gate



Truth Table - NAND

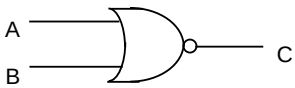
Row	Inputs		Outputs
	A	B	C
1	0	0	1
2	0	1	1
3	1	0	1
4	1	1	0

OR & NOR Gates. The schematic symbol for an OR gate is shown below on the left. It is called OR because a 1 on input A *or* input B will cause the output, C, to go to 1. The schematic symbol and truth table for the OR gate is shown below on the left. The NOR gate acts the same as an OR gate only it inverts the output. The symbol and truth table for this gate is shown below on the right on the right.



Truth Table - OR

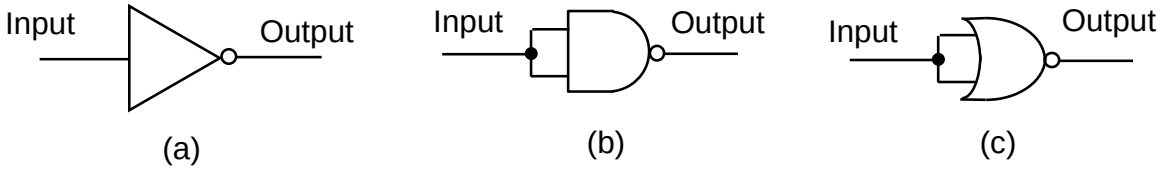
Row	Inputs		Outputs
	A	B	C
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	1



Truth Table - NOR

Row	Inputs		Outputs
	A	B	C
1	0	0	1
2	0	1	0
3	1	0	0
4	1	1	0

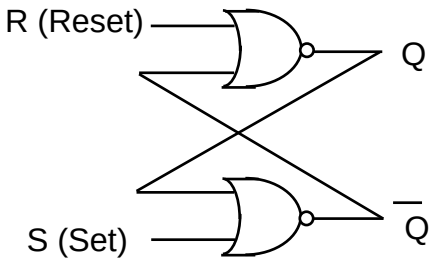
Inverter. An IC that inverts a logic 1 or 0 is called an inverter. Its schematic symbol is shown below on (a). We can easily accomplish the same thing using NAND and NOR gates as shown in (b) and (c) below.



Basic Logic: Flip Flops (FF)

RS FF. A flip-flop can store a logic state (1 or 0). The outputs of flip-flops are labeled "**Q**" and "**Q not**" (illustrated by a "Q" with a bar over the top). "**Q not**" is *always* the inverse of "**Q**".

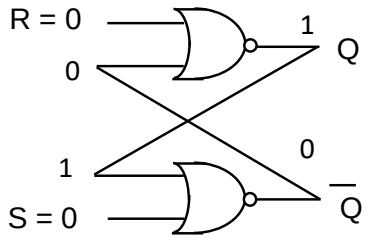
The figure below illustrates an R-S flip flop and its truth table.

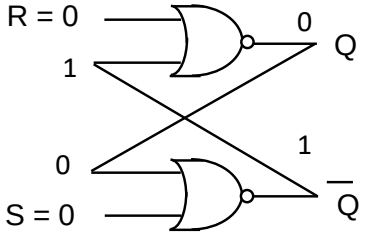


Truth Table - RS FF

Inputs		Outputs
Reset	Set	Q
0	0	no change
0	1	1
1	0	0
1	1	illegal

The R-S FF can be made of various gates; the one above is made of two NOR gates. If both Reset and Set are 0, the output is unchanged. To verify this, assume **Q**=1 and **Q not** =0 (see below, left hand schematic). Following the feedback paths on the RS flip flop, we see that the 1 on **Q** is fed back to one of the inputs on the bottom NOR gate. From the *NOR gate truth table*, on the right, we see that this will force **Q not** to a 0, *regardless of what the Set (S) input is*. Now follow the **Q not** feedback path to the input of the top NOR gate. This places a 0 on one input of the top NOR gate and the other input (R) is already 0. The *NOR gate truth table* tells us the **Q** output will be 1. From this it can be seen that the assumed output, **Q**=1 and **Q not** = 0, has not changed with Reset=0 and Set=0.



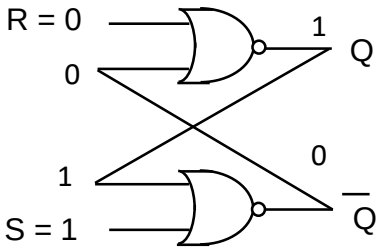


NOR Truth Table

Inputs		Output
A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

Now assume **Q** = 0 and **Q not** =1 and R = S = 0. Refer to the middle schematic above. The 0 on **Q** is fed back to one of the inputs on the bottom NOR gate whose other input, S, is already 0. The *NOR gate truth table* indicates the output, **Q not**, will be 1. This 1 is fed to an input of the top NOR gate whose other input, R, is already 0. The truth table indicates the output **Q** will be 0. Again, no change in the state of **Q** and Q not.

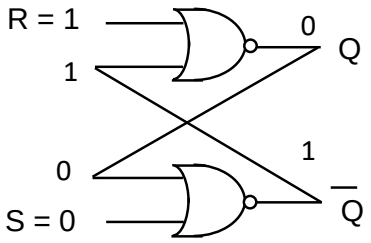
Consider now the state where Reset=0 and Set=1. Refer to the figure below. Again referring to the *NOR gate truth table*, we see that a 1 on any input of a NOR gate will cause the output to go to 0. In this case, with Set=1, **Q not** will go to 0 *regardless* of the state of the other input on the bottom NOR gate. With **Q not** = 0 and R = 0, the *NOR gate truth table* indicates **Q** will go to 1. This signal is fed back to the other input at the bottom NOR gate thus assuring **Q not** remains latched at 0 *even if the S signal goes back to 0*.



NOR Truth Table

Inputs		Output
A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

The operation with Reset = 1 and Set = 0 is the same only **Q** is latched at 0 and **Q not** at 1. Using the same logic as above, trace through the circuit below to verify that for yourself.

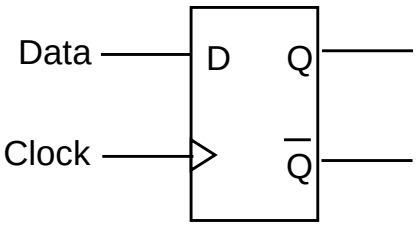


NOR Truth Table

Inputs		Output
A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

Notice in the *RS FF truth table* that a state with Reset =1 and Set=1 is declared illegal. Recall again from the NOR gate truth table that a 1 on *any* input causes the output to go to 0. In this case, this would cause both **Q** and **Q not** to go to 0. By definition, **Q** and **Q not** *must* be reciprocals of each other. Therefore, this would be an illegal action for the R-S FF. It is up to the designer to assure that such a violation (i.e., Reset=1 and Set=1) is never presented to the RS FF.

D FF. A D flip flop is illustrated below along with its truth table. The Clock can be a series of pulses that change states (1 to 0 to 1, etc.) at a fixed rate or it can be a single pulse to "latch" data into the flip-flop. With a D flip flop, the **Q** output will assume the value of the D input (1 or 0) upon the *leading* edge (0 to 1 transition) of the clock pulse.



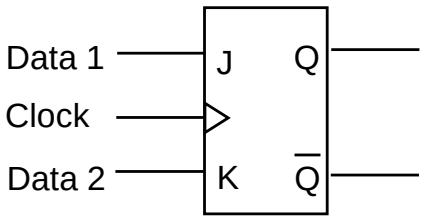
D Flip Flop

Truth Table - D FF

Inputs		Outputs
D	Clock	Q
0	⌐	0
1	⌐	1

⌐ Means "leading edge" of clock pulse

J-K FF. The J-K flip-flop is a bit more complex. Refer to schematic and truth table below. With this circuit, the **Q** and **Q not** transitions take place on the *trailing* edge of the clock. From the J-K FF truth table we can see that if J=0 and K= 0, there will be no change in the value of **Q** or **Q not** upon the trailing edge of the clock pulse. If J =1 and K = 0, **Q** will latch to 1 upon the trailing edge the clock pulse and **Q not** will latch to 0. If J = 0 and K = 1, **Q** will latch to 0 upon the trailing edge the clock pulse and **Q not** will latch to 1. If both inputs are 1, the output will toggle. That is to say, if **Q** were equal to 1 and **Q not** to 0 before the clock pulse, **Q** would go to 0 and **Q not** to 1 upon the trailing edge of the clock pulse.



J-K Flip Flop

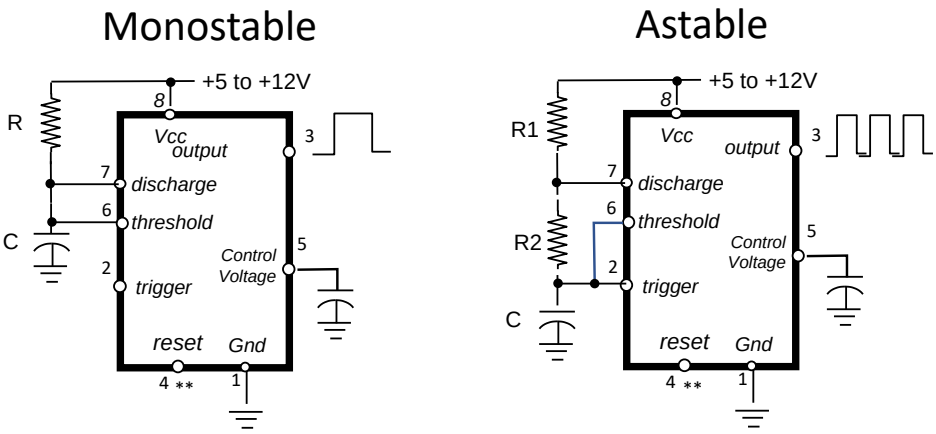
Truth Table - JK FF

Inputs			Outputs
J	K	Clock	Q
0	0	⌐	no change
0	1	⌐	0
1	0	⌐	1
1	1	⌐	toggle

⌐ Means "trailing edge" of clock pulse

Timer

A 555-timer chip can be configured as a one shot (monostable) or as a clock (astable) as shown below.



Pulse Width = 1.1 RC

Frequency = $\frac{1.443}{(R1 + 2R2)\,C}$

** "reset" normally held high, 555 resets when goes to ground

As a one shot, the output of this circuit is at zero volts until it is triggered by a positive going pulse. Once it is triggered, its output will go to a positive voltage for a time specified by 1.1 RC. At rest, the trigger input is held at V_{CC}. When it drops to approximately 1/3 V_{CC}, it will trigger.

As a clock, the 555 will provide a periodic square wave with a frequency as shown in the sketch above.

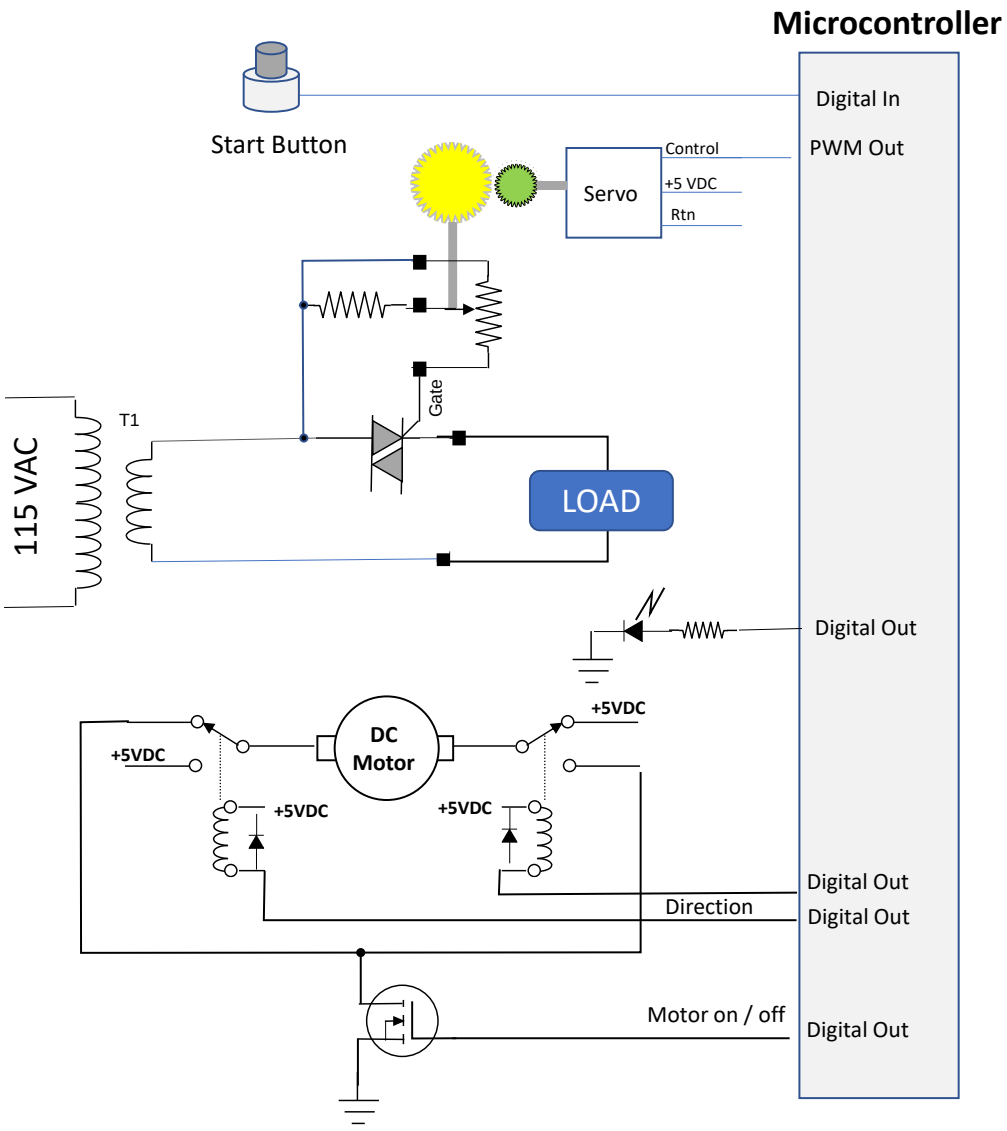
Microcontrollers

Almost all modern radio transceivers utilize a microcontroller of some sort. Microcontrollers have the advantage of small size, low power consumption, and low cost, making them ideal for embedded applications. A microcontroller has a CPU plus all the required memory and input / output circuits that allow it to be used as a "stand alone" system whereas a microprocessor is only a Central Processing Unit (CPU) and requires external components (memory, input / output, etc.) to form a functioning computer such as a lap top. Generally, a microcontroller is dedicated to a specific task such as controlling a piece of hardware.

Microcontrollers have a large number of digital input / output (I/O) pins that can be programmed as an output providing 0 or +5 VDC, or as an input that can accept 0 or +5 VDC. Some pins can also be programmed to be pulse width modulators that can provide DC pulses of various widths under program control. They usually have analog input pins that drive an internal Analog to Digital Converter (ADC). This allows analog voltages to be converted into binary, digital words that the microcontroller can process. The microcontroller must be programmed. To support this, an integrated development environment (IDE) is provided to allow the user to write, debug, and download the program. Once the application is complete, it can be download onto the microcontroller for storage in its internal non-volatile memory.

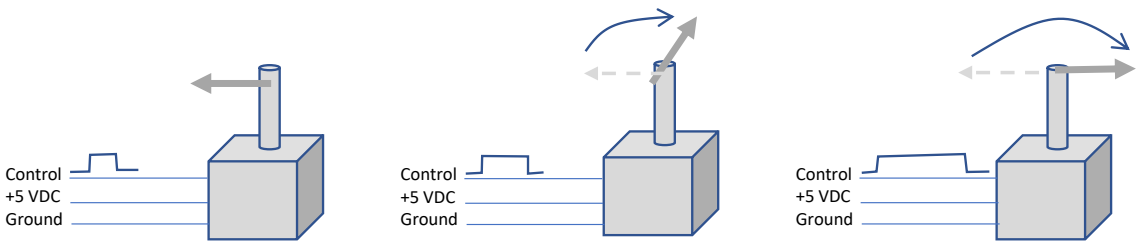
Now for some examples of radio applications. Most transceivers will provide an analog signal indicative of receive signal strength often called an RSSI (received signal strength indication). This can be applied to the microcontroller's analog inputs and passed on to the internal ADC which prepares a digital word equivalent of the analog RSSI input. The microcontroller can then process this data and provide a receive strength indication on the radio display panel for the user. The radio push to talk (PTT) button can be tied to a digital I/O pin configured as input pin. When the PTT is pressed, indicating a transmission, +5 VDC is applied to this input pin. The microcontroller can be programmed to respond to this by turning on another digital pin, programmed for an output, to close the transmit / receive relay, thus connecting the transmit power amplifier to the antenna. The microcontroller can be programmed to monitor radio battery state of charge, change frequencies, look for improper impedance match between the transmitter power amplifier and the antenna and take protective measures is necessary. The possibilities are limitless.

The sketch below illustrates some other types of controls that are possible with a microcontroller. Starting from the top there is a digital I/O pin configured for a digital input. It monitors the status of the Start Button; at rest the input is 0 VDC, when pressed the input is +5VDC. When pressed, the microcontroller could be programmed to respond by executing a specified sequence of events. Next, a servo is connected to a digital I/O pin configured as a pulse width modulator (PWM).



Servo

A servo is a device that uses the width of a control pulse to control the rotational position of a mechanical shaft on the servo. This is illustrated in the three sequences below where the control pulse width gets wider and the shaft rotational position increases. Pulse width may vary from 1.25 milliseconds to 1.75 milliseconds for a shaft rotational position from 0 to 180 degrees.



In this case, the microcontroller varies the pulse width to the servo to rotate a series of gears which, in turn, moves the wiper on a potentiometer. This changes the voltage applied to the gate of the Triac. This, in turn, varies the AC current to the load. The load could be an AC motor, a heater, etc. In this manner, the microcontroller can control the AC to the load.

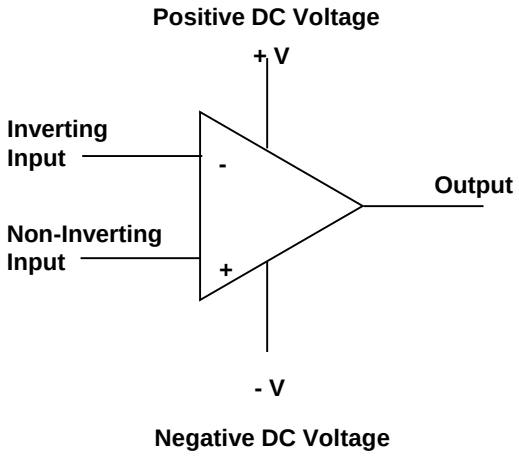
Continuing down, the next pin is a digital I/O configured as an output. The microcontroller can use this to turn on or off an LED providing a visual indication to the user.

The last three digital I/O pins are also configured as output pins to turn a DC motor on or off and change its direction of rotation. As shown, the motor would rotate in one direction, say clockwise (CS), when the microcontroller puts +5 VDC on the "Motor on / off" pin. This would turn on the enhancement mode MOSFET, applying a ground to the motor and causing it to rotate. To turn it off, "Motor on / off" is set to 0 VDC. To change direction, the microcontroller applies +5 VDC to the two "direction" pins, energizing the two relays. By inspection, it can be seen that this will reverse the polarity of the VDC applied to the DC motor. Now when +5 VDC is applied to the "Motor on/off" pin, the motor will change direction and rotate counter clockwise (CCW). The configuration of the two relays is an H bridge motor control.

Chapter 12 Integrated Circuits: Analog

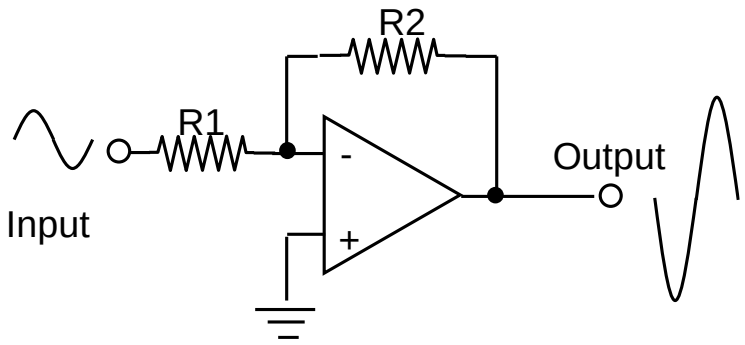
Operational Amplifiers

Another kind of analog IC is an operational amplifier, or op amp for short. These are amplifiers with very high gain, often as high as 200,000, high input impedance and low output impedance. Op amps have two inputs. One is called the non-inverting input and is designated with a "+" symbol. A signal applied to this input will come out of the op amp in phase with the input signal. The other input is called the inverting input and is designated with a "-" symbol. A signal applied to this input will come out of the op amp inverted. "Inverted" means the output signal is 180 degrees out of phase with the input signal. Do not confuse the "+" and "-" signs on an op amp input pins with positive or negative power supply voltages. Although the symbols are the same, they have different meaning when applied to the input pins of an op amp. The figure below shows the schematic symbol for an op amp. Many op amps are powered by positive and negative DC voltages; some required a single power supply voltage.



Op Amp as Amplifiers

The most common use of op amps is in amplifier circuits. There are two kinds of amplifier circuits: inverting and non-inverting. The figure below is the schematic for an inverting amplifier. Since the input signal is fed to the (-) input, the output signal will be inverted. Part of the output is fed back to the inverting input. Since the output is out of phase with the input, it will subtract from the input signal. This type of feedback is called negative feedback since it subtracts from the input signal and allows us to control the gain of the amplifier circuit.



Gain is calculated as follows:

Gain = - R2 / R1

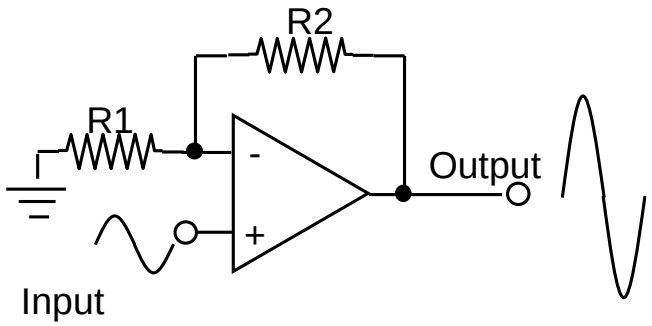
The **minus** sign (-) means the output is inverted. But why do we need negative feedback'? Recall that an op amp can have gain up to 200,000. With that much gain, even the tiniest signal on the input would cause the op amp output to saturate. If the input signal increased any more, the output voltage could not follow it; the output voltage would already be as high as it could be. To solve this problem we use negative feedback. This allows the op amp to follow the input signal and to amplify it without going into saturation. Suppose the input signal was 1 millivolt, abbreviated 1 mV. One millivolt is 1/1000 or 0.001 volt. Suppose R2 = 1 M Ohm and RI = 1 K Ohm. Gain would be:

Gain = -R2 / R1 = -1 M / 1 K = - 1000

This means the 1 mV input signal would become 0.001 Volt x -1000 = -1 Volt at the output.

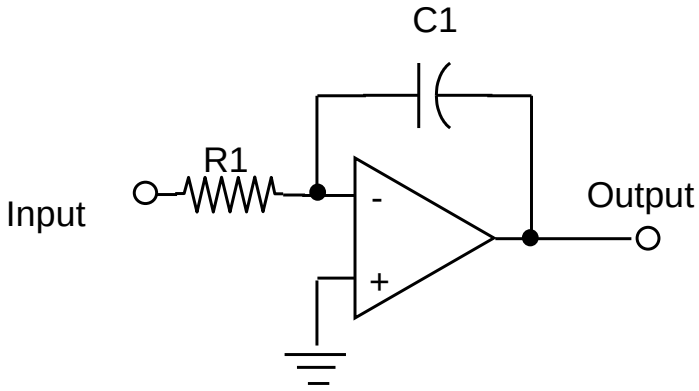
A non-inverting amplifier is shown below. This time the input signal is applied to the **(+) input**, so the output signal will be in phase with the input signal. The feedback is still to the **(-) input**. Since any signal applied to the **(-)** input will be inverted at the output, the feedback signal will subtract from the input signal. Therefore, we still have negative feedback. The gain is approximately as follows:

Gain = (R2 / R1) + 1



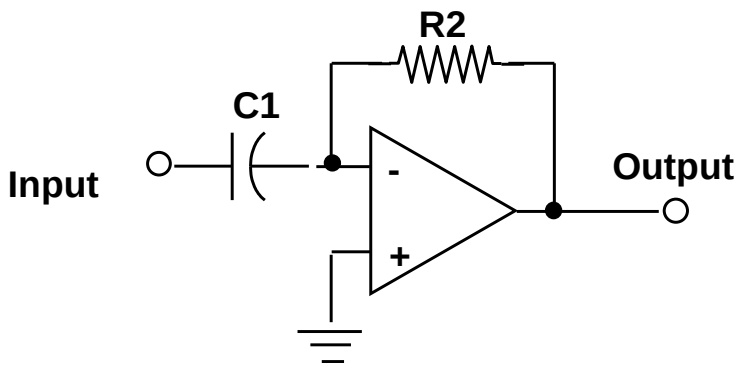
Op Amps as Active Filters

The configuration below is a low pass filter. Gain = - Xc / R1



Recall that X_C goes down as the frequency goes up. Therefore, the gain will go down as the frequency goes up. Thus, the low frequency signals are amplified but the high frequency signals are not. This is a low pass filter.

Consider now the configuration below. Gain = - R2 / Xc



Recall that X_C goes down as the frequency goes up. Therefore, the gain will go up as the frequency goes up. Thus, the high frequency signals are amplified but the low frequency signals are not. This is a high pass filter.

These are just a few uses for op amps. They can also be used as oscillators and to detect voltage levels.

Op Amps as Comparators

Op amps can be used to detect DC voltage levels. They are called comparators. They are called comparators because they compare the voltage on their non-inverting (+) inputs and inverting inputs (-). A typical comparator works like this:

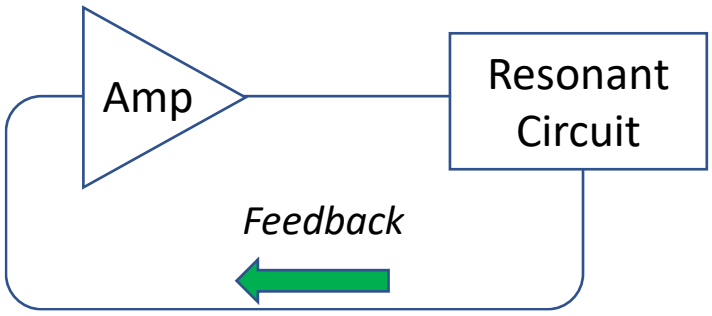
1) If the voltage on the (+) input is more positive than the voltage on the (-) input, the output will be a positive voltage.

2) If the voltage on the (-) input is more positive than the voltage on the (+) input, the output will be a negative voltage for a dual power supply op amp or ground for a single power supply op amp.

Chapter 13: Signal Sources

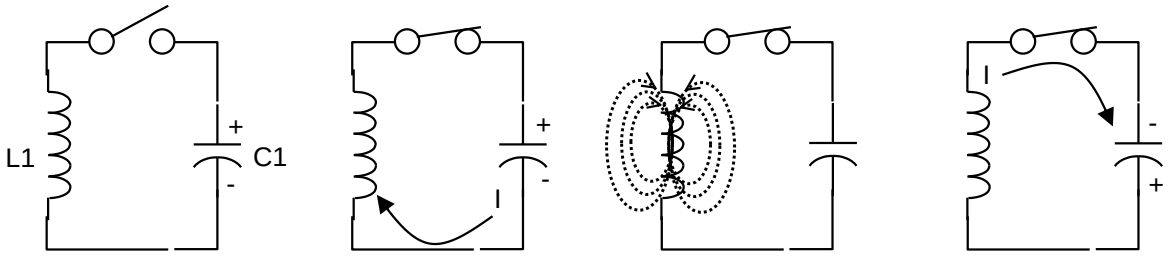
Oscillators

An oscillator is a circuit that generates an AC signal, typically at a single frequency. Oscillators are actually complex circuits to analyze and design and are often developed by plagiarizing an existing design and modifying it to suit current purposes. There are many kinds of oscillators but they all have two common features: (1) gain > 1 and (2) in phase positive feedback, meaning a sample of the output is fed back to the input in phase with original input signal to the amp. The basic architecture is illustrated below:



The resonant circuit, resonator, or filter as it would be known in a closed loop system, of which the oscillator is one, determines the oscillator frequency. Part of the oscillating signal is feed back to the input of the amplifier (Amp) in phase with the resonant circuit's oscillating signal. If the amplifier is an inverting amplifier which introduces a 180-degree phase shift, then the resonant circuit must provide an additional 180-degree phase shift to assure in phase feedback. The amplifier provides the necessary gain to make up for circuit losses and sustain oscillations but not too much so as to drive the circuit into saturation. Noise starts the oscillations which grow until they stabilize at amplitude and frequency.

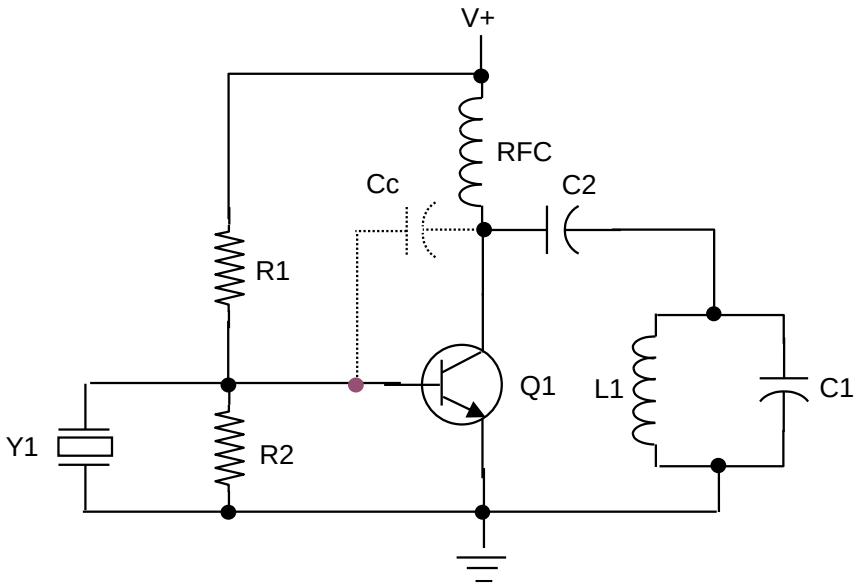
With these things in mind, let's look at how an inductor – capacitor resonant tank circuit produces oscillations. Refer to the sketch below.



Assume C1 is fully charged and S1 is open as shown in figure (a), above. This is the initial state and the energy in the circuit is stored in the electrostatic field of C1. When S1 is closed, figure (b), C1 begins to discharge through L1. Recall that inductors resist changes to current flow. L1 resists the current flow from C1, so the current through L1 and the magnetic field around it builds up slowly. Maximum current through L1 occurs when C1 is fully discharged. At this point the magnetic field around L1 is fully developed. All the energy in the circuit is now stored in this magnetic field as shown in figure (c). Now the magnetic field begins to collapse and induces a current in L1 that charges C1, figure (d), but with a *polarity opposite to that in the initial state* (compare figure (a)). At this point, C1 is again fully charged but with an opposite polarity and the process repeats itself. From this we can see the tank circuit oscillates, generating the desired AC signal, as determined by C1 and L1.

If there were no losses in the tank circuit, it would continue to oscillate indefinitely. But there are losses, primarily due to the resistance of the inductor windings, and unless we periodically "pump in" some energy to make up for these losses, the oscillations would cease. How do we add this energy? First, take a sample of the tank circuits AC signal and amplify it. Next, feed the amplified signal back to the tank circuit, in phase with the oscillating signal. This method is called positive feedback. Now we have an "energy pump" and the circuit continues to oscillate.

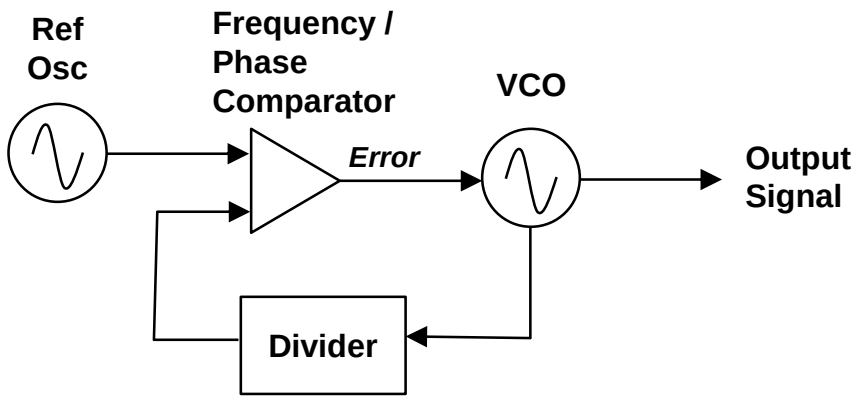
Consider now a crystal oscillator. When an AC voltage is placed across a crystal, the crystal will vibrate at a particular frequency depending on its cut, size, and structure. The smaller and thinner the crystal, the higher the frequency at which it will vibrate. Since higher frequencies demand smaller and thinner crystals, there is a limit to how high one can go with crystal oscillators. We will see later how PLL VCOs get around that problem. If the crystal is placed in an oscillator circuit tuned to the crystal's natural vibration frequency, the crystal will vibrate strongly. The crystal acts like a resonant circuit only it is much more stable than one made from inductors and capacitors. By using crystals in an oscillator, we can make much more stable frequency sources. The figure below shows one kind of crystal oscillator circuit.



R1 and R2 provide the proper bias voltage for Q1. The tank circuit is L1 and C1. Recall that a reverse biased P-N junction can act like a capacitor. Since the base to collector of Q1 is a reversed bias, it acts like a capacitor. This is shown in the figure as a capacitor Cc (dotted lines). This capacitance provides the path for the positive feedback for the Q1 amplifier. Crystal Y1 now has the positive feedback signal across it, so Y1 vibrates strongly. The signal from Y1 is amplified and "pumped" into the LC tank circuit. This keeps the circuit oscillating, developing a sine wave AC signal.

PLL VCO

Traditional LC based oscillators have pretty well been replaced by Phase Lock Loop Voltage Controlled Oscillators. A basic block diagram is shown below.



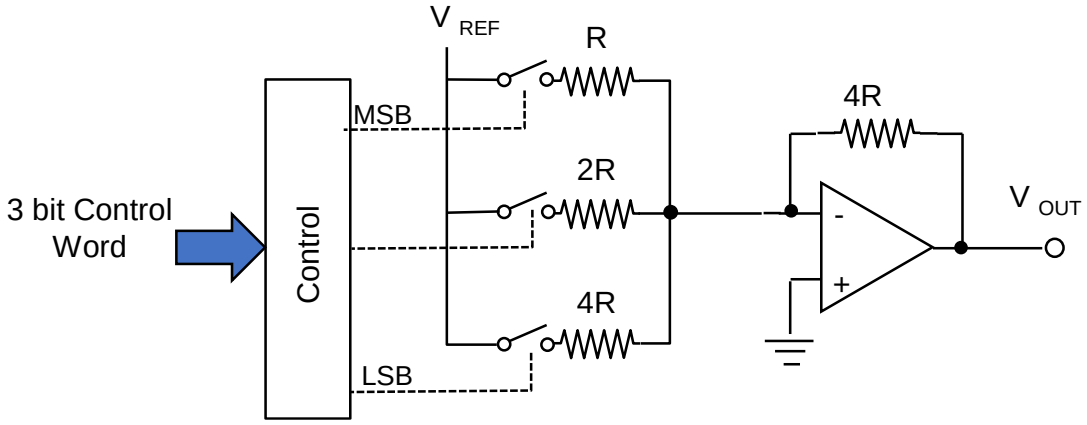
The reference oscillator (Ref Osc) is normally a relatively low frequency crystal oscillator at a fixed frequency that is very stable over time and temperature. Typical values are in the 10 to 20 MHz range. The voltage-controlled oscillator (VCO) uses a varactor diode as the capacitor in an LC tank circuit of the VCO. This LC tank circuit determines the frequency of the VCO. A DC voltage applied to the varactor diode will change its capacitance causing the resonant frequency of the capacitor – inductor circuit to change thus changing the frequency of the VCO output. Part of the VCO output is fed back through a frequency divider ("Divider" block in the figure) and then, after being divided down in frequency, is passed on to the frequency / phase comparator. The other input to the frequency / phase comparator is the output of the reference oscillator. The frequency / phase comparator develops an error voltage indicative of the magnitude and direction of the difference between the reference oscillator and the VCO output frequency, as divided down by the Divider.

Assume the reference oscillator is 10 MHz. Assume the desired output frequency is 145 MHz. To achieve that frequency, the divider is programmed to divide by 14.5 (note that these are notional values to illustrate circuit operation). Assume at start up the VCO output is 100 MHz (again arbitrarily chosen to illustrate circuit operation). The 100 MHz is divided by 14.5 to 6.896 MHz, which is fed into the frequency / phase comparator along with the 10 MHz reference signal. Since the frequency and phase are not the same, the frequency / phase comparator generates an error voltage with a magnitude indicative of the difference between the two signals and a polarity indicative of the direction (above or below the ref osc). This voltage is applied to the varactor diode in the VCO thus changing its capacitance. This change in capacitance shifts the frequency of the VCO toward 145 MHz. This continues until the VCO settles at 145 MHz. At this point, the output of the divider is 10 MHz and the output of the reference oscillator is still 10 MHz. Since both inputs to the frequency / phase comparator have the same frequency and phase, the error voltage is nil and the PLL VCO output is "locked" at 145 MHz.

To allow changes in frequency, the divider is programmable in response to user selected frequency settings. When the user dials in a new frequency, the divider is changed to provide a 10 MHz output to the frequency / phase comparator when the VCO is at the desired output frequency.

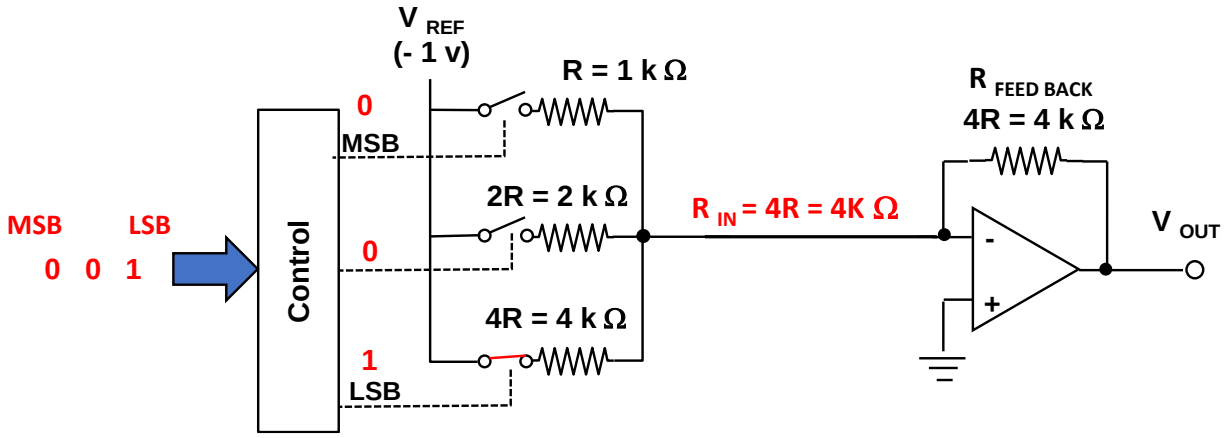
Chapter 14: Digital to Analog (DAC) and Analog to Digital (ADC)

Consider the circuit below which illustrates the basic concept of one type of DAC. Note that this illustration, and those that follow, are simplified to facilitate understanding. "MSB" means "most significant bit" and "LSB" means "Least Significant Bit". For example, in arithmetic, in the number 324, the 3 is the most significant bit, representing three 100s. The number 4 is the least significant bit, representing four 1's.



This represents a fictitious 3- bit DAC. It consists of an operational amplifier in an inverting amplifier configuration (see lesson on Op Amps) with a fixed feedback resistor and three different, selectable input resistors, R, 2R, and 4R. Note that the selected combination of R, 2R, and 4R form the input resistance, R_{IN} for the inverting op amp. The selection of the input resistors is done by the control block as directed by the 3-bit Control Word. V_{REF} provides a fixed, stable reference voltage.

Now consider the circuit shown below. R has been set to 1000 Ohms (1kΩ), 2R to 2000 Ohms (2kΩ), 4R to 4000 (4kΩ) Ohms, and $R_{Feed\ Back}$ to 4000 (4kΩ) Ohms. Again, these are fictious values for discussion purposes only. The Control Word "001" closes the relay on the 4R resistor as shown. The value of 4R, 4000 Ohms, now becomes R_{IN} for the inverting amplifier.



Knowing R_{IN} and $R_{Feed\ Back}$, the gain and V_{OUT} can be calculated as shown below:

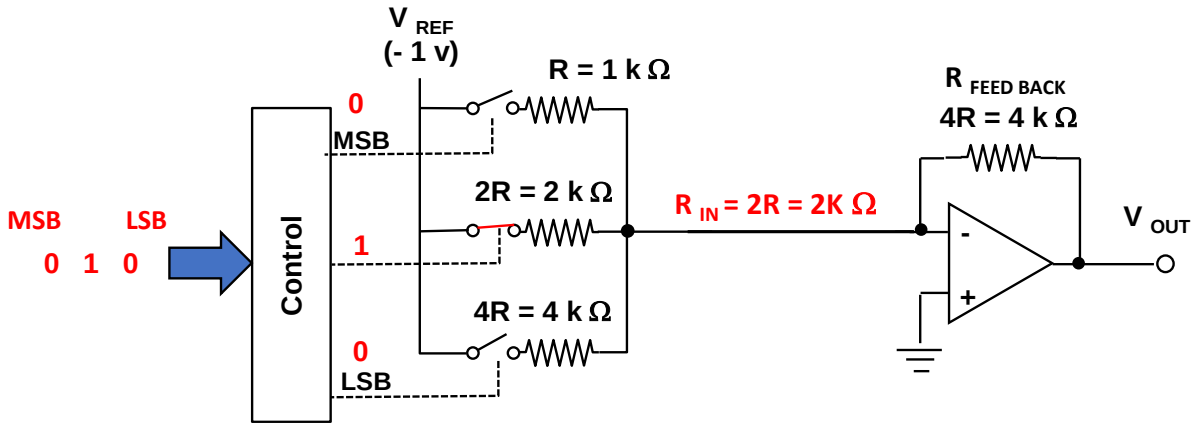
$$\text{Gain} = -\frac{R_{\text{FEED BACK}}}{R_{IN}} = -\frac{4\ \text{k}\Omega}{4\ \text{k}\Omega} = -1$$

$$V_{OUT} = \text{Gain} * V_{REF} = -1 * -1\ \text{V} = 1\ \text{V}$$

This state is summarized in the table below:

Control Word			Relays			R_{IN}	$R_{\text{Feed Back}}$	Gain	V_{REF}	V_{OUT}
MSB		LSB	R (MSB)	2R	4R (LSB)					
0	0	1	0	0	4000	4000	4000	-1.00	-1	1

Consider the next state, with the Control Word of "010" as shown below. This closes the relay on the 2R resistor. The value of 2R, 2000 Ohms, now becomes R_{IN} for the inverting amplifier.



Knowing R_{IN} and $R_{\text{Feed Back}}$, the gain and V_{OUT} can be calculated as shown below:

$$\text{Gain} = -\frac{R_{\text{FEED BACK}}}{R_{IN}} = -\frac{4\ \text{k}\Omega}{2\ \text{k}\Omega} = -2$$

$$V_{OUT} = \text{Gain} * V_{REF} = -2 * -1\ \text{V} = 2\ \text{V}$$

The summary table becomes –

Control Word			Relays			R_{IN}	$R_{Feed\ Back}$	Gain	V_{REF}	V_{OUT}
MSB		LSB	R (MSB)	2R	4R (LSB)					
0	0	1	0	0	4000	4000	4000	-1.00	-1	1
0	1	0	0	2000	0	2000	4000	-2.00	-1	2

Taking another step, with Control Word "011", R_{IN} is now a parallel combination of 2R and 4R or 2000Ω in parallel with 4000Ω which equals approximately 1.333 KΩ. Again, knowing R_{IN} and $R_{Feed\ Back}$, the gain and V_{OUT} can be calculated as shown below:

$$\text{Gain} = -\frac{R_{FEED\ BACK}}{R_{IN}} = -\frac{4\ k\Omega}{1.333\ k\Omega} = -3$$

$$V_{OUT} = \text{Gain} * V_{REF} = -3 * -1\ V = 3\ V$$

And the summary table is updated as shown:

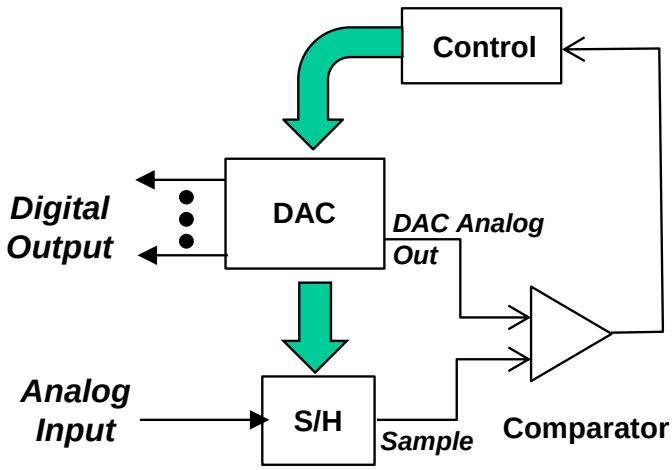
Control Word			Relays			R_{IN}	$R_{Feed\ Back}$	Gain	V_{REF}	V_{OUT}
MSB		LSB	R (MSB)	2R	4R (LSB)					
0	0	1	0	0	4000	4000	4000	-1.00	-1	1
0	1	0	0	2000	0	2000	4000	-2.00	-1	2
0	1	1	0	2000	4000	1333	4000	-3.00	-1	3

Completing the analysis sequence for the remaining possible combinations of the 3-bit Control Word, the complete summary table is as follows:

Control Word			Relays			R_{IN}	$R_{Feed\ Back}$	Gain	V_{REF}	V_{OUT}
MSB		LSB	R (MSB)	2R	4R (LSB)					
0	0	1	0	0	4000	4000	4000	-1.00	-1	1
0	1	0	0	2000	0	2000	4000	-2.00	-1	2
0	1	1	0	2000	4000	1333	4000	-3.00	-1	3
1	0	0	1000	0	0	1000	4000	-4.00	-1	4
1	0	1	1000	0	4000	800	4000	-5.00	-1	5
1	1	0	1000	2000	0	667	4000	-6.00	-1	6
1	1	1	1000	2000	4000	571	4000	-7.00	-1	7

From this, it can be seen that stepping the 3-bit Control Word from "001" to "111", V_{OUT} can be stepped linearly from 1 Volts to 7 Volts.

DACs can be used to do the inverse, convert analog to digital (ADC – analog to digital converter). Consider the successive approximation ADC, which is essentially a "trial and error" process, shown below -



The S/H is a sample and hold circuit capable of capturing and storing, for a brief period, the amplitude of the analog input signal to be "digitized". *The comparator output will go high if the DAC analog output is higher than the sample and low if it is lower.* For simplicity, we will assume the DAC is a 4-bit DAC with a full range of one Volt. Thus, it has sixteen steps ($2^4 = 16$) with a resolution of 0.0625 VDC (1 V / 2^4) meaning it can provide sixteen distinct output voltages as multiples of 0.0625 VDC. Refer to the example below. The Control block directs S/H to sample and hold the amplitude of the analog input, which we will assume is 0.384 volts. The basic operation of this ADC is shown below:

DAC Digital Word	Equivalent Decimal Value	DAC Out, VDC	Sample, VDC	Difference, VDC
0000	0	0.0000		
0001	1	0.0625		
0010	2	0.1250		
0011	3	0.1875		
0100	4	0.2500	0.384	-0.1340
0101	5	0.3125		
0110	6	0.3750	0.384	-0.0090
0111	7	0.4375		
1000	8	0.5000	0.384	0.1160
1001	9	0.5625		
1010	10	0.6250		
1011	11	0.6875		
1100	12	0.7500		
1101	13	0.8125		
1110	14	0.8750		
1111	15	0.9375		

First, Control set the DAC digital word to 1000 (decimal equivalent 8) which produces a DAC Out of 0.500 VDC or ½ its full range of 1 VDC (yellow blocks on table). This is higher than the Sample of 0.384 VDC so the comparator output goes "high". Control responds by cutting the DAC Out in half to 0.250 VDC. It does this by sending the control word 0100 (decimal equivalent 4) which produces a DAC Out of 0.250 VDC (green blocks). This is lower than the Sample of 0.384 VDC so the comparator output goes "low". Control responds by increasing DAC Out to a value half way between 0.500 and 0.250 VDC. It does this be sending the control word 0110 (decimal 6) which produces a DAC Out of 0.375 VDC. This is lower than the Sample of 0.384 VDC. However, the difference between the DAC's output and the Sample is 0.009 VDC which is less than the resolution of this 4 bit DAC (0.0625 VDC) so this is the best this ADC can do. Thus, with this ADC, the digitized value of 0.384 is 0110 (blue blocks).

Chapter 15: Radio Frequency (RF) Principles and Common Circuits

Impedance

Impedance is the combination of all the capacitive reactance, inductive reactance, and DC resistance in a circuit. Since the reactance's affect AC only, impedance is the total effective AC resistance in a circuit or component. At first, one might think that to find the impedance of a circuit we would simply add X_C, X_L, and R together. However, recall that in a capacitor the current leads the voltage by 90 degrees and in an inductor, the current lags the voltage by 90 degrees. Because of this phase difference, the current in X_C and X_L are 180° out of phase and cannot simply be added together to get impedance. Impedance in a series resistor – inductor - capacitor circuit (abbreviated RLC) is calculated as follows:

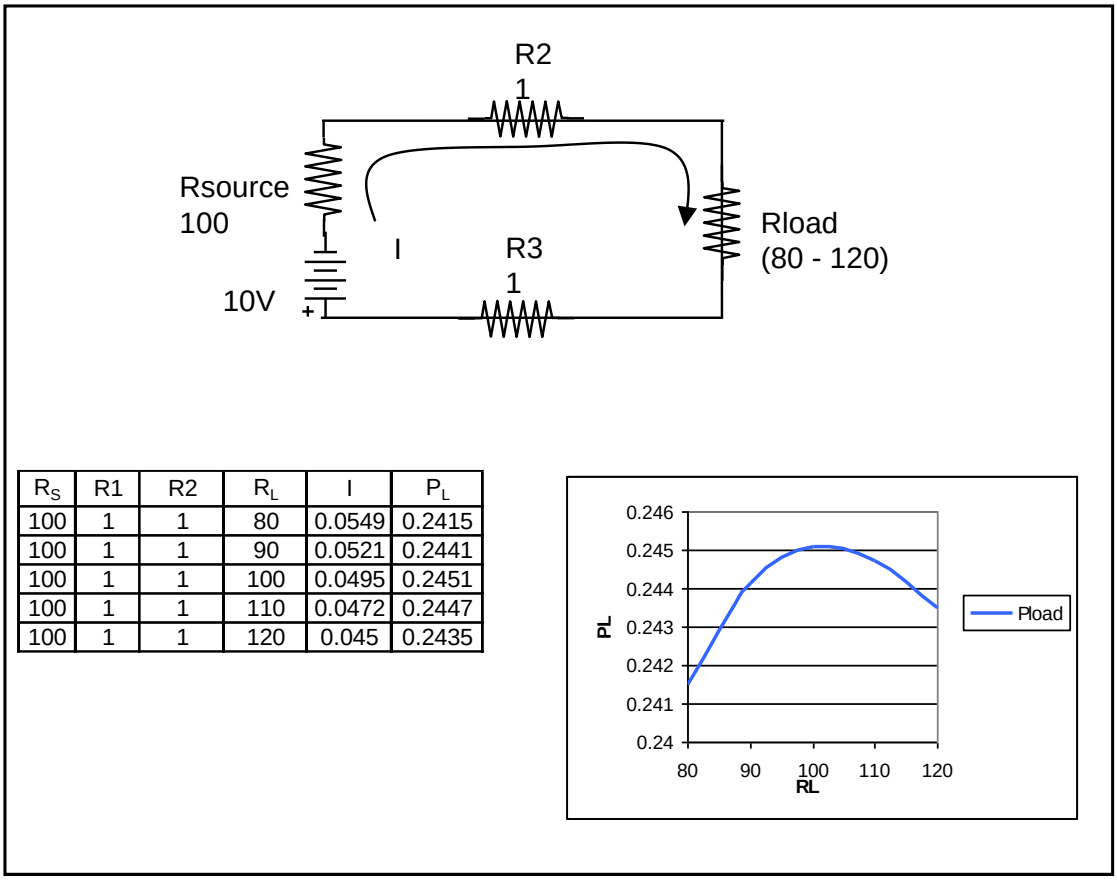
$$Z = \sqrt{R^2 + (X_L - X_C)^2}$$

Impedance (Z*)* is also measured in Ohms.

Impedance Matching

Impedance is very important in radio. In radios, we transfer, or couple, a signal from one circuit to another. We want to get the maximum signal transfer from one circuit to another. If the source impedance or R_{source} is equal to the load impedance or R_{load}, maximum power transfer will occur. The circuit and associated table below illustrate this. Notice that the source resistance, R_{SOURCE}, is fixed at 100Ω. The feed lines to the load, R_{LOAD}, both have a resistance of 1Ω. In the table, R_{load} is varied from 80 to 120 Ω. *Note in both*

the table and the graph that the maximum power in R_{load} occurs when $R_S = R_L$. Maximum power transfer always occurs when the source impedance matches the load impedance.

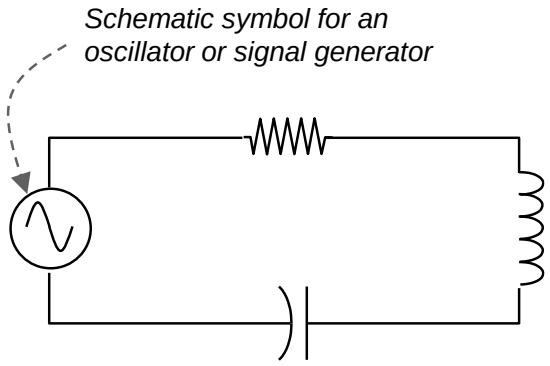


R_S	R1	R2	R_L	I	P_L
100	1	1	80	0.0549	0.2415
100	1	1	90	0.0521	0.2441
100	1	1	100	0.0495	0.2451
100	1	1	110	0.0472	0.2447
100	1	1	120	0.045	0.2435

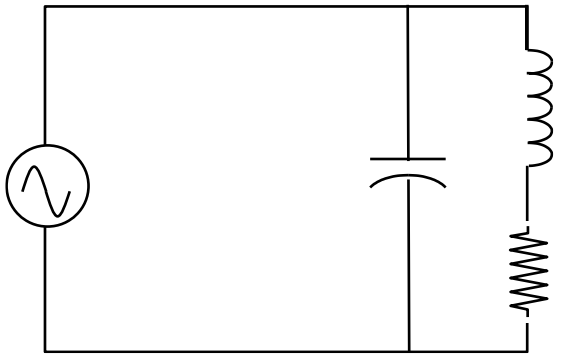
Resonance

Resistors, capacitor, and inductors work together to achieve **resonance**, which is the point of discussion of this section, more particularly, resonant circuits. There are two types of resonant circuits: series and parallel.

A series RLC circuit being driven by a signal generator is shown below. Note that the DC resistance is primarily that due to the windings of the inductor and is usually very small.



A parallel RLC circuit being driven by a signal generator is shown below. Again, the DC resistance is primarily that if the inductor windings and is small.



At some frequency, $X_L = X_C$. The frequency at which that occurs is the **resonant frequency** and is calculated as shown below:

$$f = \frac{1}{2\pi\sqrt{LC}}$$

In a series resonant circuit at resonance, X_L and X_C are equal and opposite and $X_L - X_C = 0$. The impendence, Z, of the series RLC circuit would be:

$$Z = \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{R^2 + 0} = \sqrt{R^2}$$

$$Z = R$$

Typically R is a small value in series RLC circuits. *Therefore in a series resonant circuit, Z is minimum at resonance. Notice further that at resonance, the impedance is all resistive (i.e., it has no frequency dependent characteristics such as X_C or X_L since they have cancelled each other out)*

A parallel LC circuit is a bit trickier. In a parallel circuit, the voltage is the same across all the legs of the circuit. At resonance, the currents in the inductive and capacitive legs are equal and opposite and essentially cancel each other out, minimizing current in the circuit. So a simple way to look at this is through Ohms law.

$$R = E / I$$

If E, the voltage remains constant and I, the current, is very low, then R, the resistance, will be high. And this is the case in a parallel resonant circuit. *Thus, in a parallel resonant circuit, Z is maximum at resonance and is also resistive.*

This feature, i.e., a very low impedance at series resonance or a very high impedance at parallel resonance at a *specific* frequency, allows radios to "tune" to the desired frequency. For example, in the receiver, a series resonant circuit would have low Z at the resonant frequency and a high Z at other frequencies. Thus, it could be used to block all frequencies except the desired one. In reality, these circuits are not perfect and may operate over a broad or narrow range of frequencies depending on how much non-reactive resistance is in the circuit. This is quantified by "Q", the quality factor. Q is the ratio of the reactance to the resistance in the resonant circuit. The resistance is primarily due to the DC resistance in the windings of the inductor. The lower the Q, that is the closer the resistance is to the reactance, the broader the range of frequencies that will "pass" through a resonant circuit. The higher the Q, that is the lower the resistance with respect to the reactance, the smaller the range of frequencies that will "pass" through the resonant circuit.

The choice of Q depends on the application. If it is desirable to operate over a broad range of frequencies, then a low Q is desirable. If it is desirable to operate over a very narrow range of frequencies and have a "sharp" frequency response, then a high Q is desirable.

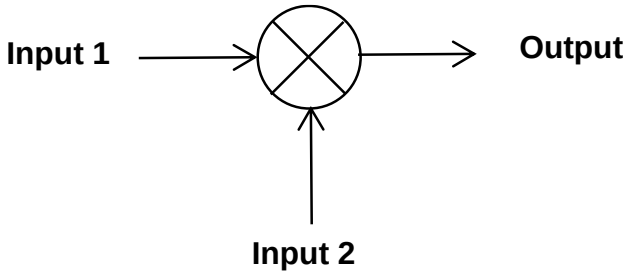
A crystal is a component that can act like a series or parallel resonant circuit. They are cut from quartz crystals. Crystals are generally used in radios as resonant circuits because they are more stable than inductors and capacitors and have high Q's. The schematic symbol for a crystal is shown below. On a schematic, the letter "Y" usually identifies crystals. For example: Y1 or Y22.

Harmonics & Multipliers

Harmonics are signals whose frequencies are multiples of the resonant frequency. The resonant frequency is called the fundamental frequency. If the fundamental frequency was 100 kHz, the second harmonic (second multiple) would be 200 kHz, the third harmonic (third multiple) would be 300 kHz, and so forth. These harmonics are not near as strong as the fundamental. But we can use harmonics to generate higher frequency signals. A multiplier does this. A multiplier is an amplifier with a tank circuit in its output. The output tank circuit is tuned to a frequency that is two or three times the input signal frequency (2^{nd} or 3^{rd} harmonic). Thus, the tank circuit selects the second or third harmonic for its output. Multipliers are often used with oscillators to generate higher frequency signals.

Mixers

Mixers are devices, often passive, that have two inputs and one output. Two analog signals are fed into to the inputs, one signal on each input. The output will have four signals: the two input signals, the sum of the two inputs, and the difference of the two inputs. The schematic symbol for a mixer is shown below:



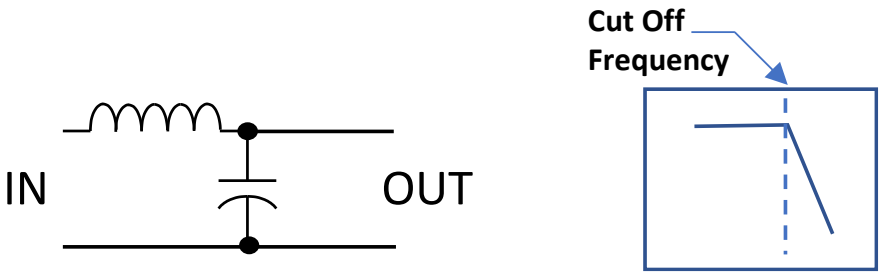
Mixers are used in radios to shift frequencies, up or down. Consider a VHF radio receiving at 145 MHz. If we fed the 145 MHz VHF signal into input 1 and a local oscillator output at say 125 MHz into input 2, we would have four signals at the output of the mixer: 145 MHz, 125 MHz, 270 MHz (sum), and 20 MHz (difference). Using a low pass filter on the output, we could isolate the difference frequency (20 MHz) for further processing and demodulation.

A mixer can also be used in a transmitter to reach higher frequencies. Using the same frequencies from the example above, a high pass filter on the output would isolate the higher 270 MHz sum signal for transmission.

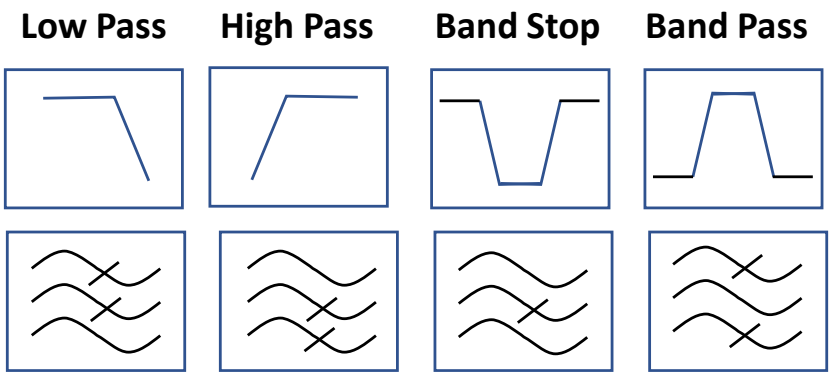
Filters

There are four types of filters: low pass, high pass, bandpass, and band stop. A low pass filter will pass all signals below its cutoff frequency and block all signals above its cutoff frequency. The reciprocal is a high pass filter: all frequencies below cut off are blocked while all frequencies above are passed. Bandpass and band stop have two cutoff frequencies. A bandpass will pass all frequencies between its lower and upper cutoff frequencies and block all others. A band stop will block frequencies between its lower and upper cutoff frequency and pass all others.

Inductors and capacitors are used to make filters because of their frequency sensitive reactance. Consider the simple low pass filter shown below made of an inductor and a capacitor.

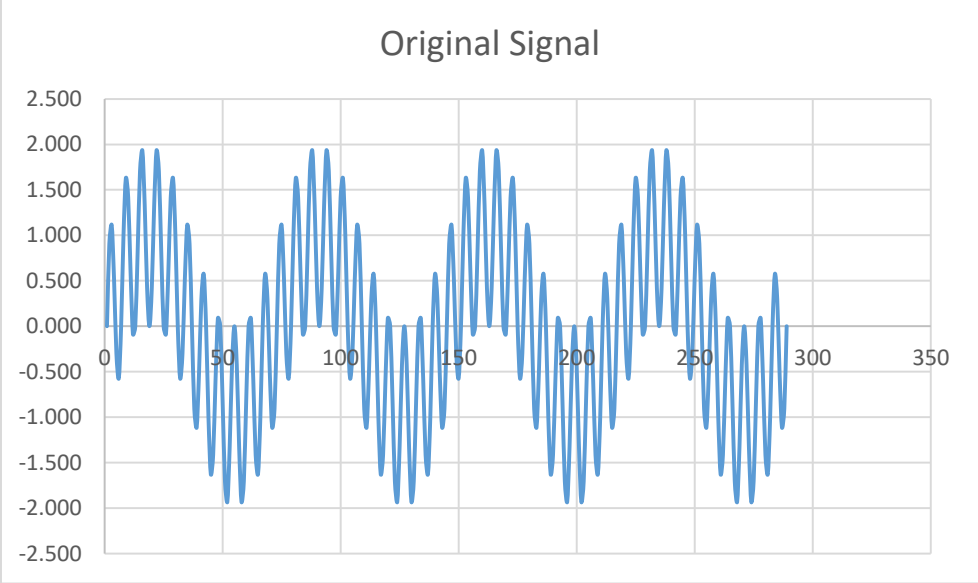


As the frequency goes up, the inductive reactance increases which reduces the signal going from “IN” to “OUT”. As the frequency goes up, the capacitive reactance goes down shunting signals to ground. These two actions tend to pass frequencies below cut off and block frequencies above. A high pass filter places the capacitor is series and the inductor in parallel. Now as the frequency goes up, the capacitive reactance goes down allowing signals to pass from “IN” to “OUT”. Also, as the frequency goes up, the inductive reactance goes up, which reduces how much of the signal is shunted to ground. The net effect is that the high pass filter tends to block frequencies below cut off and pass frequencies above cut off. Commonly used symbols for each filter type are show below:



Chapter 16: Digital Signal Processing

Consider a simple but specific DSP application: a moving average filter. The signal to be filtered is shown below.



Amplitude
0.000
0.906
1.113
0.518
-0.301
-0.574
0.000
0.996
1.628
1.414
0.592
-0.087
0.000
0.819
1.706

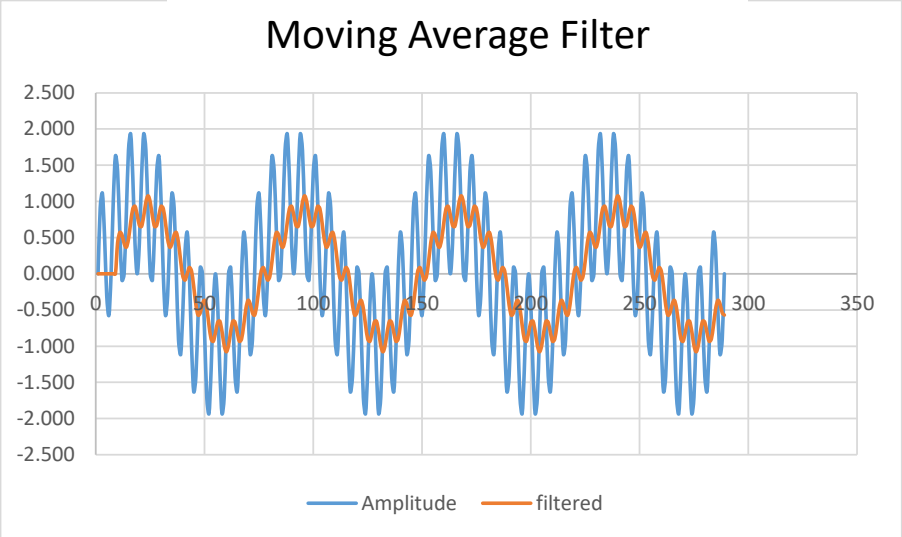
•
•
•

0.000
0.574
0.301
-0.518
-1.113
-0.906
0.000

The signal to be processed is periodically sampled. That is to say, the signal amplitude is measured at specific, repeated time intervals. The measured signal amplitude is converted from an analog value (Volts), to digital value using an analog to digital converter (ADC). A partial listing of the digitized samples for this example are shown in the table to the left. Although the actual digitized value would be in logic ones and zeros (e.g., 111010), they are shown with their equivalent analog value to simplify the discussion. The computer would, of course, operate on digital words.

The processing with this simple low pass filter is a moving average. In this example, we average over ten data points. Refer to the table below, on the left. Since we are averaging over ten samples, the first nine rows in the "filtered" column have been set to zero. The value in row 10, "filtered" column, is the sum of "Amplitude" in rows 1 – 10, divided by ten. The value in row 11, "filtered" column, is the sum of "Amplitude" in rows 2 – 11, divided by ten. This process continues until all the sampled data has been filtered. The filtered output signal is reconstructed by taking each filtered value and converting it to an analog voltage via a digital to analog converter. Note that the table illustrates only a partial listing of the moving average data. The resulting signal is shown below with the input for comparison.

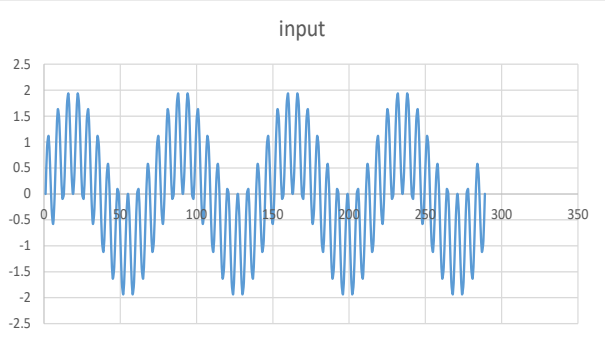
row	Amplitude	filtered
1	0.000	0
2	0.906	0
3	1.113	0
4	0.518	0
5	-0.301	0
6	-0.574	0
7	0.000	0
8	0.996	0
9	1.628	0
10	1.414	0.429
11	0.592	0.570
12	-0.087	0.539
13	0.000	0.419
14	0.819	0.367
15	1.706	0.479

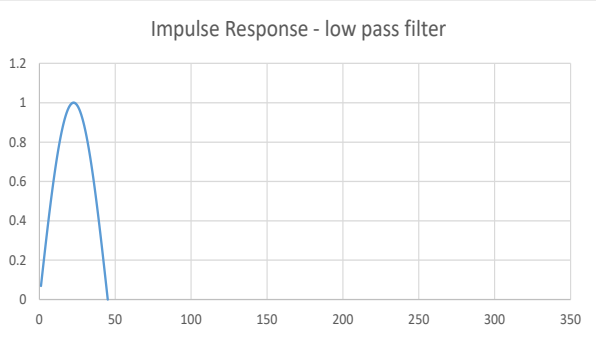


A note about sampling. In order to reproduce a signal, the signal needs to be sampled at rate at least twice its highest frequency. This is known as the Nyquist criteria. Put another way, the signals to be sampled should not have a frequency greater than ½ the sampling rate. If this principle is violated, frequencies greater than the Nyquist limit will introduce false readings, or aliases, into the sampled data. To prevent

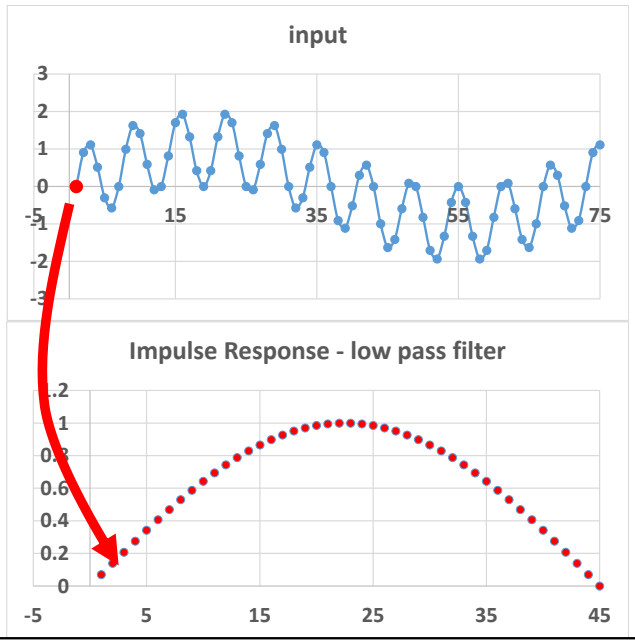
this, a low pass filter or anti-alias filter is placed before the sample and hold circuit to prevent unwanted, higher frequencies. This is an art unto itself but in keeping with the limited scope of this volume, it is sufficient at this time to make the reader aware of this limit.

Convolution is also used to develop DSP filters. The symbol for the convolution process is "*" which is unfortunate because the same symbol is often used in computer programs to mean multiplication. Consider two sets of digitized data, shown below: one is for a "noisy" signal you wish to filter and the other is the impulse response of the desired low pass filter. An impulse response is a measure of how a circuit, in this case a low pass filter, would respond if subjected to an impulse. An impulse can be visualized as a pulse of infinite amplitude and zero width. The filter's impulse response can be used to see how the filter would respond to other types of input signals.





Each of these inputs, the signal to be filtered and the low pass filter impulse response are both digitized and loaded into computer arrays for processing. Think of arrays as post office boxes in a post office. Each box has a specific address and can hold letters addressed to it. An array is similar only instead of a wall of metal post office boxes, an array is a block of computer memory set apart under a single name. Instead of an address, each memory block in the array has a specific index number. In our example, the array for digitized signal values is called "X" and it has an array index of "I", providing a format of X(I). The "I" varies incrementally as follows: 0, 1, 2, 3, 4 … to the maximum size of the array. So the first digitized signal input is stored in array X(0), the next digitized signal input is stored in X(1), and so forth. The digitized low pass filter impulse response is similarly stored in an array H(J). The first digitized impulse response value is stored in H(0), the next in H(1), etc. Finally, the filtered output is stored in array Y(I+J). This is a little tricky because its index is a function of the index for both the input array (I) and the impulse response array (J). The processing is summarized in the panel below.



Each digitized signal value is multiplied by ***every*** digitized value for the low pass filter impulse response.

In the figure to the left, the first digitized input value is represent by the single red dot on the upper graph. This value is multiplied by *every one* of red dots in the lower graph. Each of these red dots represents the digitized values for the Impulse Response – low pass filter. The results of these multiplications are added together, along with previous values already in the filtered output array Y(I+J) per the code snippet below. The algorithm then steps to the next input digitized value and repeats the process.

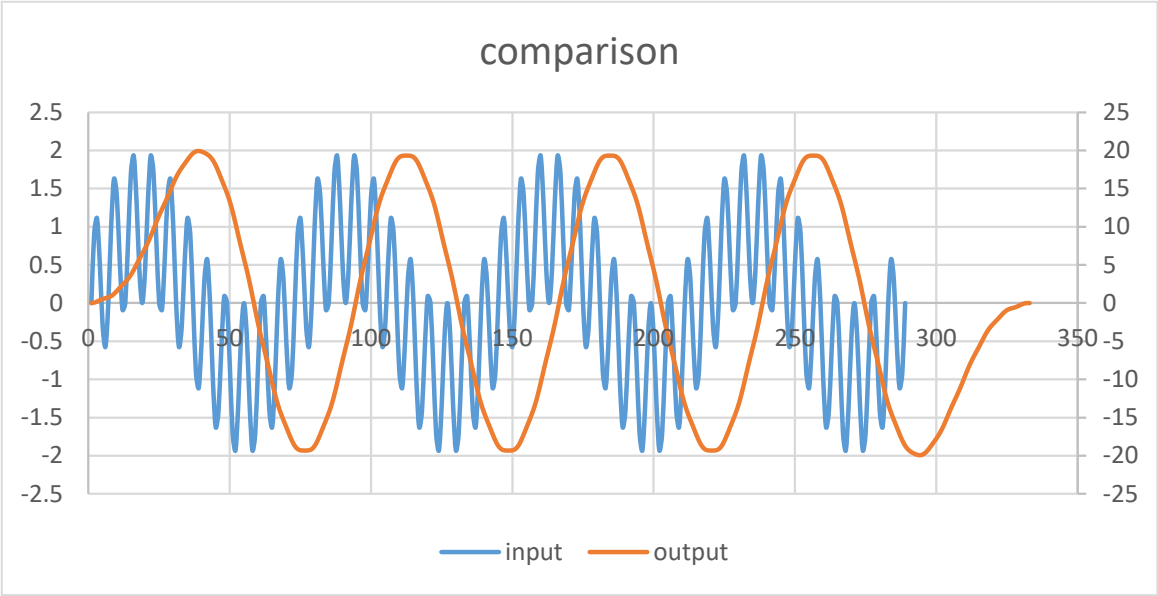
Although the basic concept is simple, the array handling can get tedious. For now, it is enough just to grasp the basic concept described above. The code snippet below describes the array handling and is provide FYI.

```
Do Until X(I) = 999  'go through each input value
   J = 0
   Do Until H(J) = 999  'go through each impulse response value
      Y(I + J) = Y(I + J) + X(I) * H(J) 'filtered output signal
      J = J + 1
   Loop
   I = I + 1
Loop
```

Array X(I) contains the sampled amplitudes of the input signal
Array H(J) contains the sampled amplitudes of the impulse response
Array Y(I+J) contains the output or filtered signal

Note that when each array is full, dummy values of 999 have been inserted to indicate the end of the sampled data

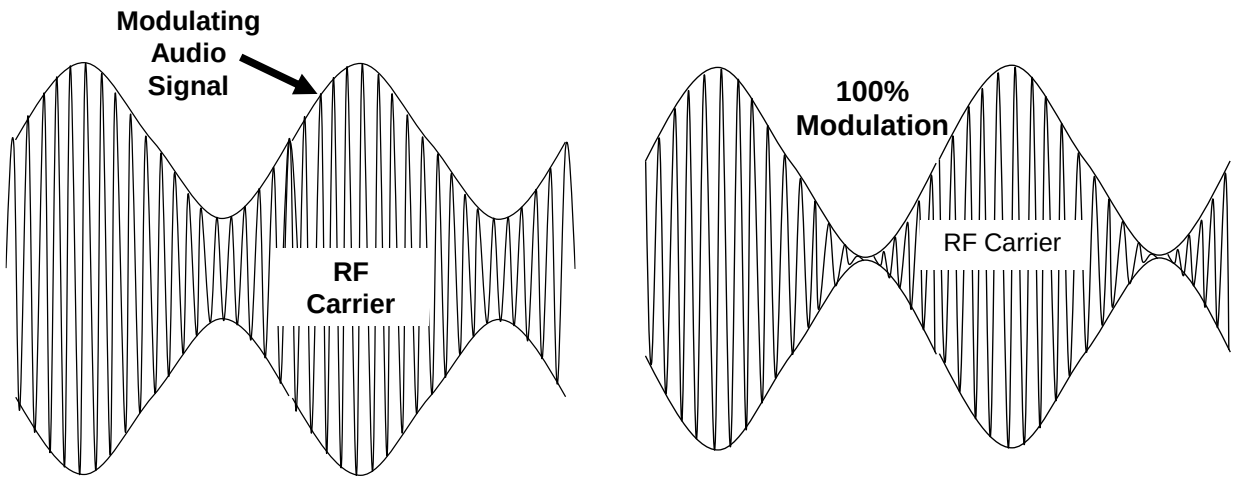
Again, the code snippet is provided for your information but it is not necessary to understand it in detail at this level. Sufficient to know is that the basic algorithm is fairly simple and it provides the results as shown below.



Chapter 17: Modulation and Demodulation

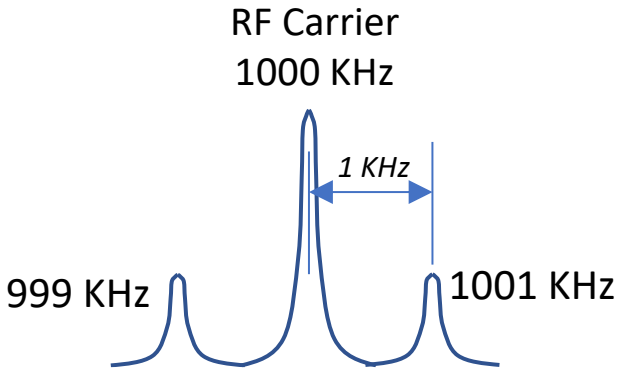
Amplitude Modulation (AM)

Modulation is the process of superimposing information (voice or data) on an RF carrier for purposes of transmission to a distant receiver. Amplitude modulation does this by varying the amplitude of the RF carrier in sync with the modulating signal, usually audio, as illustrated below in the left-hand sketch. As the modulation signal amplitude increases, the peaks and valleys of the RF carrier will increase until they "bottom out" as shown below in the right-hand sketch. At this point, the amplitude modulation is 100%. Any attempts to increase modulation beyond this point will only lead to distortion.



With AM, there is an RF carrier center frequency and two sidebands, one above and one below the center frequency. The separation between the center frequency and the two sidebands is equal to the

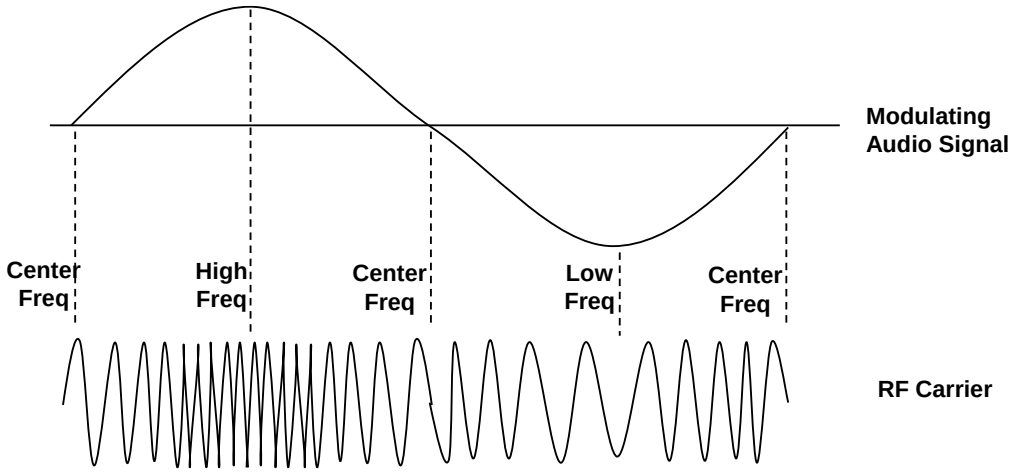
frequency of the modulating signal. This is illustrated below for a commercial AM broadcast station operating at 1,000 KHz while being modulated with a 1 KHz tone.



The same audio information is contained in both sidebands. To conserve bandwidth, the carrier frequency can be suppressed and one of the sidebands can be filtered out, leaving only the upper or lower sideband. This variation is known as single sideband, suppressed carrier amplitude modulation and is common in most High Frequency (HF) radio transceivers operating in the 1 – 30 MHz range.

Frequency Modulation (FM)

With frequency modulation, instead of the varying the *amplitude* of the RF carrier with the modulating signal, the *frequency* of the RF carrier is varied with the modulating single. When the RF carrier is not modulated, the frequency of the radio is at its assigned RF center channel frequency. When it is modulated, the frequency of the RF channel is varied in accordance with the modulating signal as shown below.



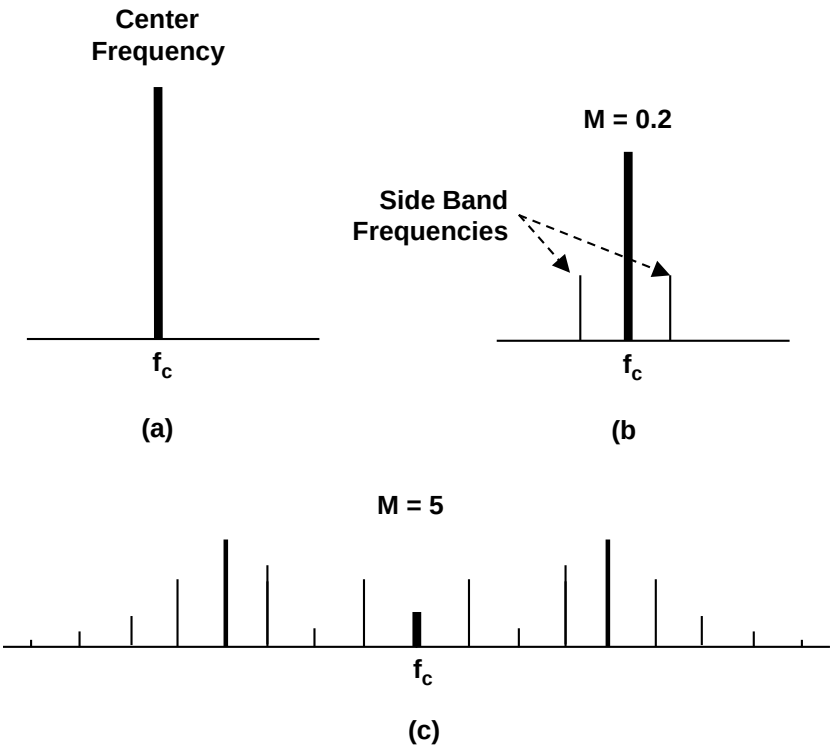
As the audio signal begins to go positive, the frequency of the RF carrier increases. When the audio signal reaches peak positive, the RF carrier is at its highest frequency. Since frequency is the number of cycles pers second, this is depicted as the cycles “bunching up” on the figure. Now the amplitude of the audio signal begins to swing back toward zero. Notice in the figure above that as this happens the RF carrier frequency begins to go down, as depicted by the cycles spreading out. When the amplitude of the audio signal reaches zero, the RF carrier is back at the center frequency. Now as the audio signal begins to

swing negative, the RF carrier frequency continues to decrease and the cycles spread out even more. When the audio signal is at its peak negative amplitude, the RF carrier is at its lowest frequency. As the audio signal negative amplitude begins to swing back to zero volts, the RF carrier frequency begins to increase again. When the audio signal reaches zero volts, the RF carrier is back to its center frequency. This change in the RF carrier frequency is called deviation. The *amplitude* of the audio signal controls how much the RF carrier changes or deviates from the center frequency. The higher the amplitude, the more frequency change or deviation of the RF carrier frequency. The *rate* of deviation (how fast the RF frequency varies) is equal to the frequency of the modulating audio signal.

The degree of FM modulation is defined by the modulation index, M.

M = Deviation of RF Carrier / Frequency of Modulating Signal

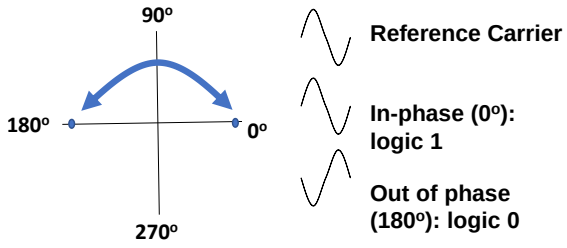
To understand why the modulation index is important, we must first understand what happens to the energy in the RF carrier when it is frequency modulated. The figure below shows how the energy in the RF carrier is distributed as the RF carrier goes from no modulation to various level of modulation. In (a), below, there is no modulation and all the energy is concentrated at the center frequency. This is denoted by the single line in (a) at the center frequency, fc. The height of the line indicates the relative amount of energy in the RF carrier. The higher the line, the more energy. Now assume the RF carrier is frequency modulated with a modulation index, M, of 0.2. Figure (b), shows two sidebands, above and below the center frequency. The energy from the center frequency carrier is partially re-distributed to the sidebands. This is why the center frequency is depicted with a lower amplitude in (b) than in (a). Figure (c) shows the sidebands at M = 5. Several observations can be made by examining this figure. As the modulation index goes up, the number of sidebands goes up and the frequency range over which they are spread increases. This range of frequencies is called the bandwidth of that channel. Also, the power in the outlying sidebands tends to go down, as depicted by the shorter lines, meaning they provide less and less value in conveying information from transmitter to receiver. The FCC governs how much bandwidth can be allocated for a given radio service. For example, commercial FM broadcast is authorized to operate in the 88 to 108 MHz VHF (very high frequency) band. The FCC allows each FM station a maximum carrier deviation of +/- 75 KHz with a maximum audio or modulating frequency of 15 kHz. Thus, the maximum allowable modulation index, M, is 75 KHz / 15 KHz = 5.



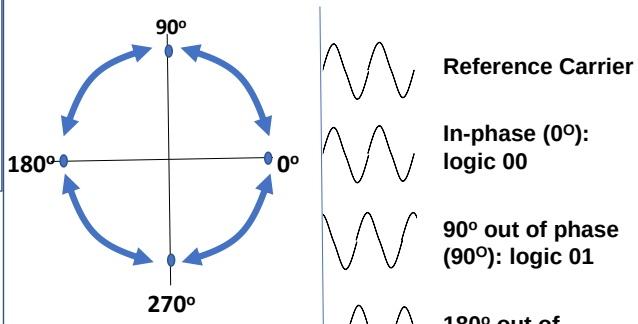
Digital Modulation

Digital circuits have only two conditions: on or off. One condition is defined as a logic "1" and the other a logic "0". Phase Shift Keying (PSK) is common means of digital modulation which varies the *phase* of the carrier signal to convey whether a 1 or 0 was sent. Two common types of PSK are binary phase shift keying (BPSK) and quadrature phase shift keying (QPSK). A single logic value (1 or 0) is a bit. BPSK sends one bit at a time. QPSK sends two bits at a time. The panels below summarizes both forms of digital modulation.

BPSK Modulation: A logic "1" input does not change the phase of the carrier; a logic "0" input shifts the carrier phase by 180 degrees with respect to the reference carrier



QPSK Modulation: A logic "00" input does not change the phase of the carrier; a logic "01" input shifts the carrier phase by 90 degrees with respect to the reference carrier, a logic "11" shifts the carrier phase by 180 degrees with respect to the reference carrier, and a logic "10" shifts the carrier phase by 270 degrees with respect to the reference carrier, as shown below



It should be noted that the association of data bits with a specific phase shift in the examples above is strictly arbitrary and is done for illustrative purposes only.

In amateur radio PSK31 digital modulation, the "carrier" is an audio tone that can be transmitted on the FM transceiver just as a voice signal would be. In order to detect phase shifts, the phase of the audio carrier must be synchronized on both the transmit and receive ends.

Normally, the individual bits (1 or 0) are packed together in 8 data bits forming a byte. A ninth bit, called a parity bit, is often added for error detection. For example, if the total number of 1's in the byte is an odd number, the ninth parity bit is set to 1 at the transmitter; otherwise, it is set to 0. This is called odd parity. At the demodulator, the number of 1's in the byte is counted. If the count is an odd number, the demodulator will expect a 1 in the parity bit; if the count is an even number, it will expect a 0 in the parity bit. If these expectations are not meet, it means an error occurred and at least one of the eight data bits in the byte was corrupted during transmission. The receiver can respond to this error by sending a special message back to the sender requesting that the last transmission be repeated. This is sometimes referred to as automatic repeat request (ARQ). Another method uses a "check sum". A block of data bits is processed by a specific algorithm to produce a "check sum". The check sum is appended to the transmitted data block. The receiver will use the received block of data bits to re-generate the "check sum" using the same process as was used by the sender. If the receiver's check sum matches that appended by the sender, then there were no errors in transmission; otherwise, an error occurred during transmission and the receiver can take steps to recover the data such as ARQ.

Forward Error Correction (FEC)

Many digital communication systems use some form of forward error correction. This process adds bits and encodes them, along with the original data bits, in such a manner as to enable the receiver to detect *and* correct errors. Since errors during transmission normally come in bursts, the encoded data at the transmitter is "scrambled" prior to transmission by an interleaver. There are many types of interleavers, but a simple one is a block interleaver. This can be a simple block of memory where the encoded data is loaded in by columns and read out by rows. The process is reversed at the receiver restoring the bits to their original order. The restored bits are then fed to the receivers FEC circuit for error correction. By doing this the "errored" bits presented to the receiver FEC are spread out making recovery easier.

We will look at two popular FEC arrangements: convolutional encoders and turbo coders.

For this discussion, assume the FEC encoder is constraint 3, rate 1/3 convolutional as shown in Figure 4-2. "Constraint 3" means it is a three bit shift register which is often abbreviated as k = 3. The term "rate 1/3" means there are three encoded output bits for each one data bit in. There are two exclusive OR gates (XOR - represented by circles with + sign) and a multiplexer switch or "mux". An XOR gate is a logic device that provides a logic 1 output as long as there are an uneven number of logic 1's on the inputs. Note that all three cells in the shift register start with a logic 0 (flushed condition). In this state, the output of the top XOR is logic 0 and the output of the bottom XOR is also a logic 0. With the mux connected to contact u1, the output u1 is logic 0, the output of the top XOR. When the mux moves to contact u2, output u2 is a logic 0, the output of the bottom XOR. When the mux moves to contact u3, the output is connected to cell1 in the shift register which is a logic 0. The encoded output is therefore 000.

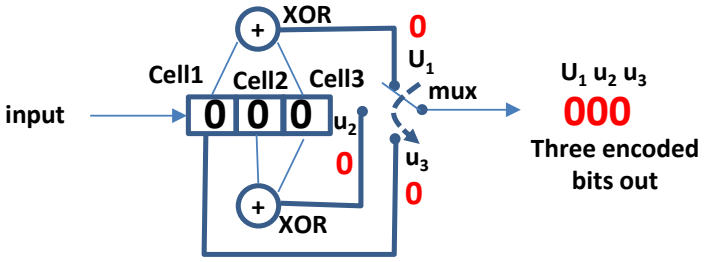


Figure 4-2 Notional Convolutional Encoder for Discussion Purposes

Assume a logic "1" is now applied to the input as shown in figure 4-3. With the next clock pulse (not shown in figure 4-3), the contents of cell2 are shifted into cell 3, the contents of cell 1 are shifted into cell 2 and the input is loaded into cell 1. The state of the shift register is now as shown in figure 4-3. The output of the top XOR will be a logic 1, the output of the bottom XOR logic 0, and the contents of cell 1 will be a logic 1. As the mux switch moves from contact u1 to u2 to u3, the output will be 101, as shown.

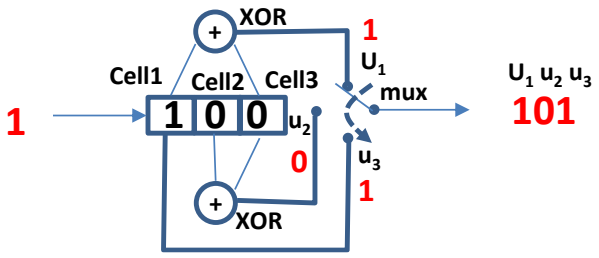


Figure 4-3 Convolutional Encoder State

Since this is a three bit shift register, there are $2^3 = 8$ possible state combinations. Figure 4-4 shows the output of all eight possible states.

Truth Table

Shift Register Content	Encoded Bits
0 0 0	000
0 0 1	110
0 1 0	010
0 1 1	100
1 0 0	101
1 0 1	011
1 1 0	111
1 1 1	001

Figure 4-4 Truth Table

The state machine for this encoder is illustrated in figure 4-5. Starting at state 000 (flush) a logic 1 input will shift the state register content from 000 to 100. Instead of calling it the “content of the shift register”, we will refer to it as the “state”. The logic 1 input to state 000 is represented by the solid line from state 000 down to state 100. This indicates that a logic 1 input at state 000 will advance the state machine to state 100. By inspection of the truth table, in figure 4-4, it can be seen that the output (the encoded bits) of state 000 is 000 while the output of state 100 is 101. This (101) is what is sent to the remote receiver. While in state 100, a logic input of 0 (indicated by dotted line) will move the state machine to state 010 with an output of 010. A logic input of 1 (solid line) while in state 100 will move the state machine to state 110 with an output of 111. This same process can be continued to build the state diagram in figure 4-5.

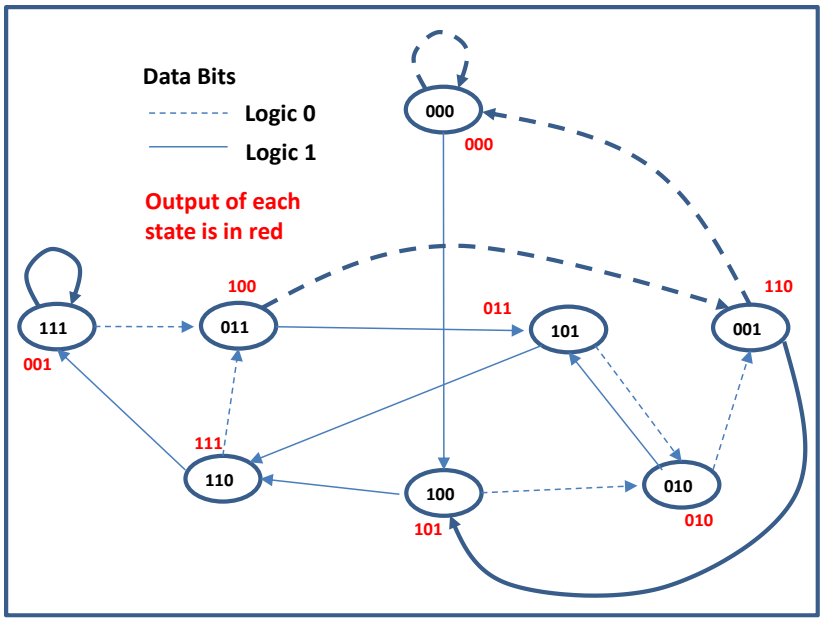


Figure 4-5 Encoder State Machine

So far, a fictitious convolutional encoder has been described for instructional purposes. There are many "real" configurations in use. Two examples are shown below:

Rate ½, constraint 7 with code vectors of 1001111 and 1101101

Rate 1/3, constraint 7 with code vectors of 1001111, 1010111, and 1101101.

The code vectors tell us what cells in the shift register are connected to the XOR gates. Refer to figure 4-6 which shows the "top" half of the rate ½ constraint 7 encoder mentioned above with the code vector of 1001111.

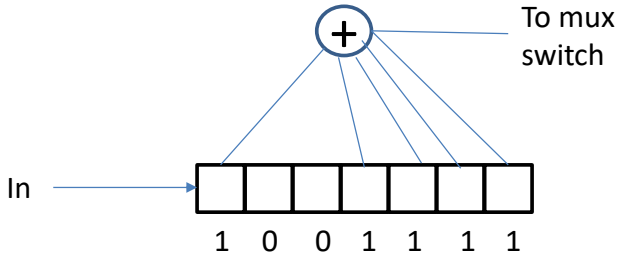


Figure 4-6 Code Vectors

Each cell in the shift register that has a "1" associated with it from the code vector is connected to the XOR gate; the ones with a "0" are not. It should be noted that the code vector is *not* the data bits in the register but is a means of indicating which cells in the shift register are connected to the XOR gate. Some texts show the code vector as its octal equivalent (117_8 in this case).

The complementary component at the remote receiver is the FEC decoder. In this case, it is a Trellis decoder. As will be seen, the Trellis decoder has a state machine which mimics the k=3, rate 1/3 convolutional encoder on the transmit side. While the convolutional encoder converts data bits into a sequence of encoded bits that become symbols, the trellis decoder does just the opposite – it converts a sequence of encoded bits, recovered from the received symbols, into data bits. It also recovers the data bits even when some of the symbols are errored in transmission. Figure 4-7 shows how the trellis decoder state machine follows the convolutional encoder state machine.

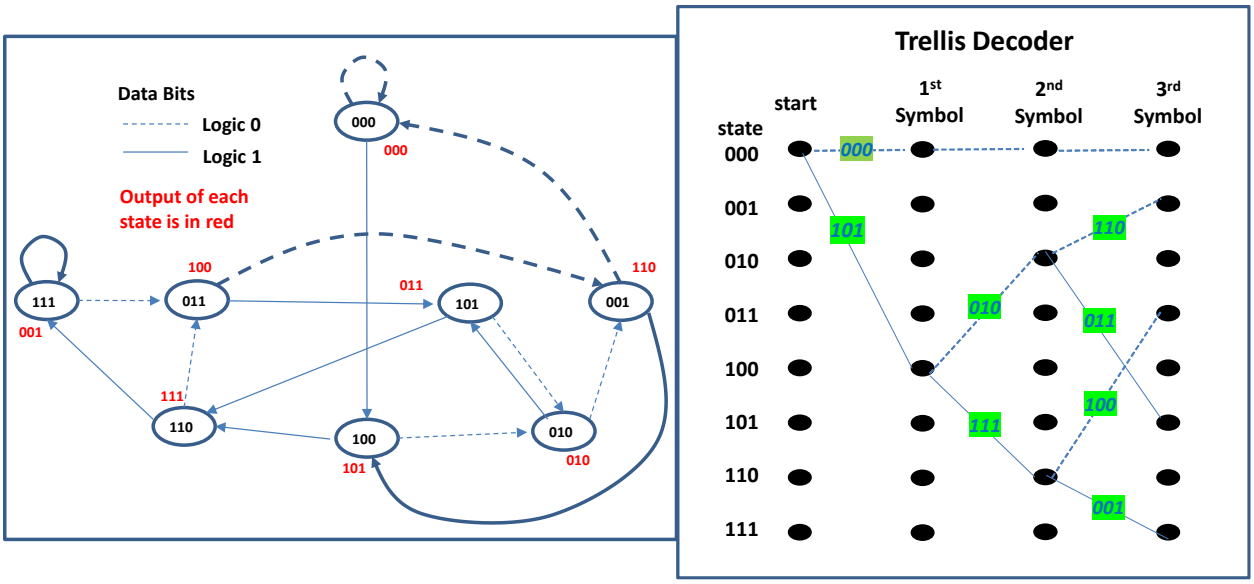


Figure 4-7 Encoder State Machine vs Decoder State Machine

The convolutional encoder state machine is on the left of figure 4-7 and the associated Trellis decoder state machine is on the right. Assume both start at a 000 state. Note on the Trellis Decoder diagram that the decoder states are shown in the left hand column. As with the encoder, we start at state 000.

Assume a logic 1 is applied to the encoder input. As shown in the encoder state machine, this will cause the state to move from 000 to 100. In state 100, it will generate an encoded bit stream of 101 as shown in red under the state bubble 100. 101 is then transmitted as the first symbol. This will be received at the decoder. Note that there are two branches from state 000 on the decoder which is shown highlighted in green. One is labeled 000, the other 101. The received encoded bits in the first symbol will determine which branch is taken. In this case the encoded bits for the first symbol are 101 so the solid branch going down to the bullet associated with state 100 under the 1st symbol column is taken. Note that this branch is a solid line, indicating it is associated with a logic 1. Note that the decoder state transitions match those of the encoder.

Back to the encoder, assume the next data bit for transmission is a logic 1 also. The encoder will move from state 100 to state 110. In state 110 the encoder will generate an encoded bit stream of 111. This is transmitted as the second symbol. At the decoder, the bullet associated with state 100 in the first symbol column also has two branches – one labelled 010, the other 111. Since a 111 is received, the decoder will

take the solid branch to state 110 in the second symbol column. Note again that the solid branch is associated with a logic 1. Note also that the decoder state transitions match those of the encoder.

Returning to the encoder for the third symbol, assume the next data bit for transmission is also a logic 1. The encoder will move from state 110 to state 111, generate an encoded bit stream of 001, which is transmitted as the third symbol. Once received, this will cause the decoder to branch from state 110 to state 111 for the third symbol, again on a solid line representing a logic 1.

From this exercise, it can be seen that the state machines for the encoder and decoder match.

Now we will repeat this exercise only this time it is assumed that, for whatever reason, the channel between transmitter and receiver is noisy at the instant the first symbol (101) is transmitted causing a 111 to be received instead of a 101. Consider figure 4-8, the Trellis decoder after receipt of the errored 111, shown in red at the top of the diagram. From state 000, only two branches are allowed: one for a received bit stream of 000, the other for a 101, neither of which was received. The decoder then calculates the Hamming distance associated with each branch. The Hamming distance is simply the number of bits in the branch labels that are different from the actual received bit stream. So for the dotted line branch requiring a 000, there are three bits that are different from the received 111. For the solid line branch requiring a 101, there is one bit that is different from the received 111 (the middle bit). The Hamming distance is annotated on the decoder state machine for each branch as shown on figure 4-8.

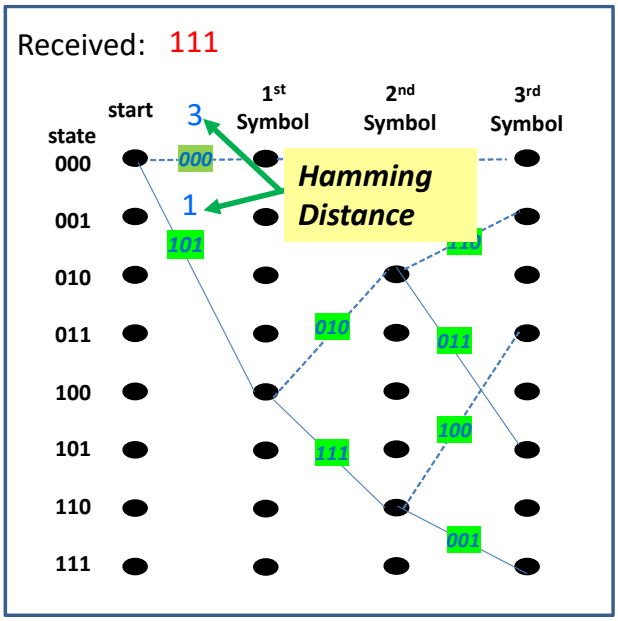


Figure 4-8 Hamming Distance, First Symbol

In figure 4-9, the Hamming distance for the second and third received symbols has been calculated and annotated on the diagram. On the right hand side of the diagram, the sum of all the Hamming distances from each bullet in the third symbol column back to state 000 is shown. The path with the lowest score is the recovered path. The logic value of each branch on this path represents the recovered data: solid line

= logic 1, dotted line = logic 0. This lowest score path is summarized in figure 4-10. As can be seen, the recovered data is 111, exactly what was intended, despite the error in transmission.

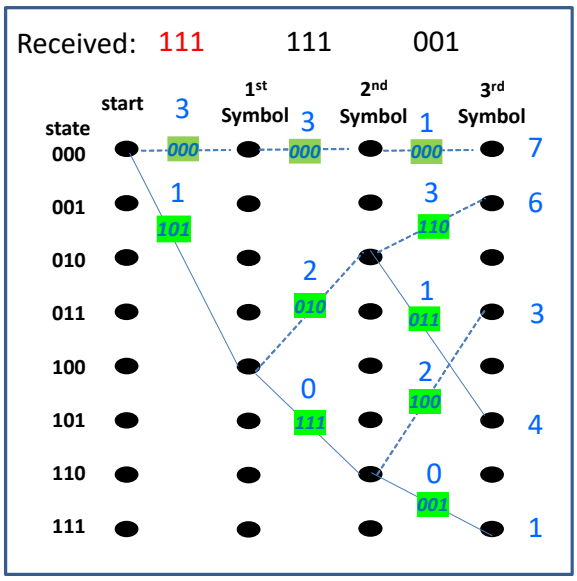


Figure 4-9 Hamming Distance, All Symbols

Symbol	From State	To State	Hamming Distance	Line Style	Line Logic Value
3rd	111	110	0	solid	1
2nd	110	100	0	solid	1
1st	100	000	1	solid	1

Figure 4-10 Recovered Path

At this point it is well to remember the role of the interleaver – it scrambles bits before transmission and restores them to their proper order through the receiver de-interleaver. This spreads out the errored bits making recovery easier.

A popular decoder, first introduced in 1967, is the Vitirbi decoder. It is similar to the trellis decoder except metrics for all branch lines are calculated. These metrics are used to identify any paths that could not possibly be a valid choice. These paths are discarded. If two paths converge on the same state, the path with the best metric is kept and the other is discarded.

Recent years, turbo coding has taken over and with good reason: it can provide performance that reaches the theoretical best performance possible. This section will examine the basic principles of turbo codes (Ruttik, 2007, Sklar 1997). A typical Turbo encoder is shown in figure 4-11:

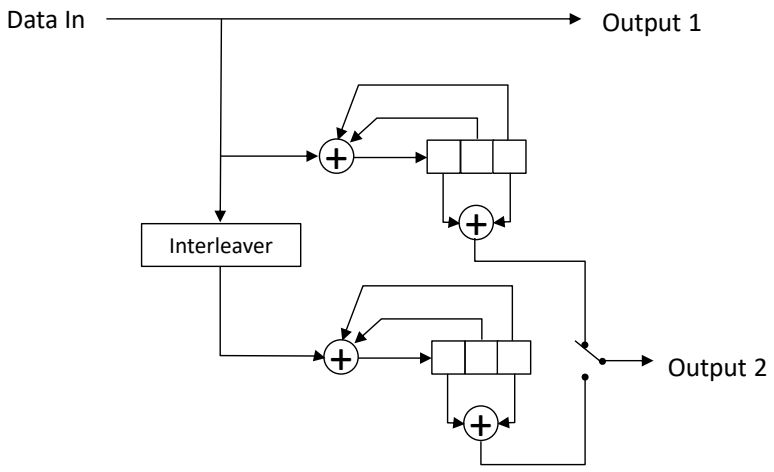


Figure 4-11 Turbo Encoder Example

Those familiar with turbo coders may recognize this as a parallel concatenation of two Recursive Systematic Codes (RSC). While the name is interesting to note, we are concerned with its characteristics for now. First, note that there are two encoders connected by an interleaver. Each encoder consists of a shift register with the output of various cells XOR'd as seen before. Some of the XOR'd signals are feed back into its input. Encoded outputs are taken off as shown in the figure. Note that one output is simply the original data input. In summary, turbo coders are characterized by two or more encoders implementing feedback and with an interleaver between them.

A block diagram of a notional turbo decoder is shown in figure 4-12.

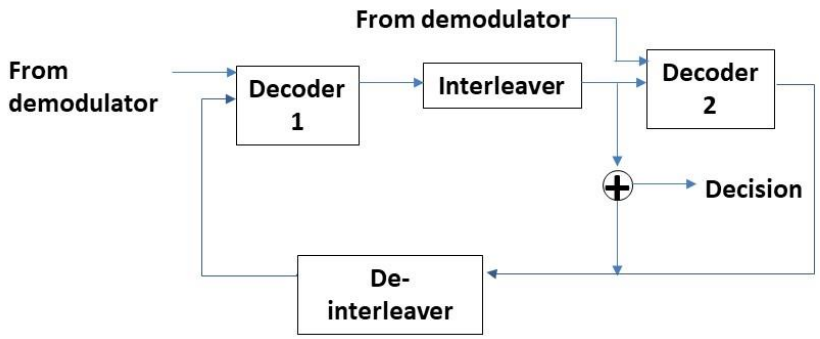


Figure 4-12 Turbo Decoder Block Diagram

As expected, it is the inverse of the turbo coder. There are two decoders, connected by an interleaver with feedback from the output of Decoder 2, through a de-interleaver, back to the input of Decoder 1. It uses the natural log of likelihood ratios to form soft decisions. By way of explanation of log likelihood ratios, refer to figure 4-13.

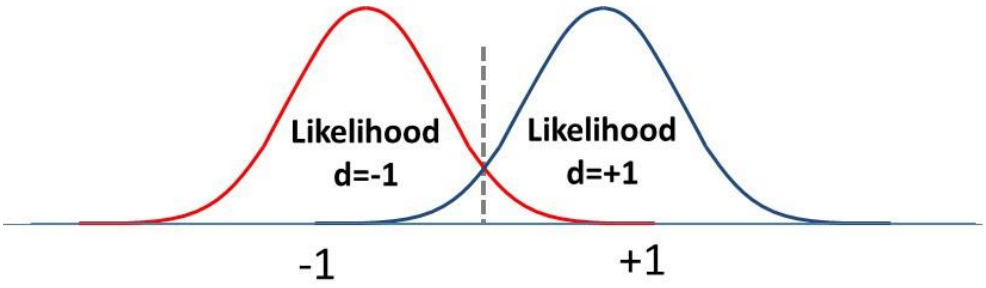


Figure 4-13 Likelihood Ratios

The two curves shown represent the probability distribution function (pdf) of two signals which are generated by encoding a logic “0” data bit (red curve) and a logic “1” data bit (blue curve). In figure 4-13, the traditional “-1” for a logic “0” and +1 for a logic “1” are used. This form is often used in literature and represents a -1 volt or + 1 volt, respectively. One may say that the red curve represents the signal range that could be attributed to the likelihood of the data being a -1 while the blue curve provides the same insight for a signal generated by a +1. Assume that at a given instant, a signal is received as shown by the vertical grey line on figure 4-14.

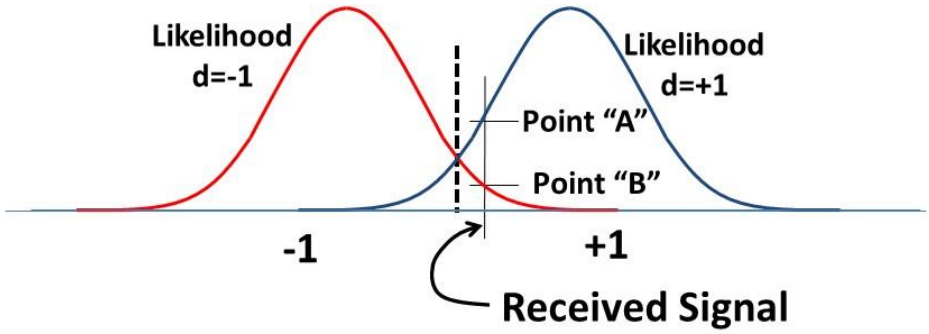


Figure 4-14 Likelihood Ratio Example

The vertical grey line crosses the blue likelihood +1 curve at point “A” and the red likelihood -1 curve at point “B”. These points represent the likelihood that the received signal represents a +1 (blue curve, point “A”) or a -1 (red curve, point “B”). The likelihood of a +1 is greater than the likelihood of a -1. Now assume the likelihood of the received signal being a +1 is 0.18 (point “A”) and the likelihood of it being a -1 is 0.05 (point “B”). The ratio of these two likelihoods is the likelihood ratio and would be 0.18 / 0.05 = 3.6. By taking the naturel logarithm of this ratio we come up with another useful metric, the log-likelihood ratio or LLR. In this case, it would be ~ 1.3.Now we will see how this metric is used in turbo decoding.

Consider the classic example of four data bits in a 2 x 2 array. Another two bits is added for horizontal parity, and another two bits for vertical parity. We will assume that even parity is used. Assume these data bits are converted to electrical signals with +1 Volt representing a logic 1 data bit and -1 Volt a logic 0 data bit. During modulation and transmission these voltages are perturbed by noise such that at the

demodulator output the channel measurements, Lc(x), have been altered. Figure 4-16 illustrates the original data and parity bits as well as the “noised up” channel measurements out of the demodulator.

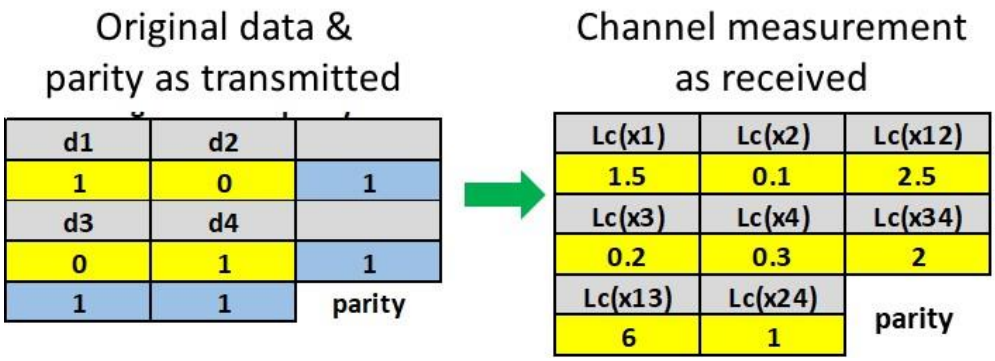


Figure 4-15 Data, Parity, and Channel Measurement

The polarity of Lc(x) on the receive side represents the data: a positive polarity indicates a logic 1 and a negative polarity represents a logic 0. The magnitude represents the confidence in the decision; the higher the number, the higher the confidence.

There are three components that go into turbo decoding. The first is Lc(x), the channel measurement LLR out of the demodulator. This becomes the input to the turbo decoder. The second in L(d) which is the a priori likelihood that the data sent is a logic 1 or 0. The third is Le(d) which is the extrinsic LLR based on extra knowledge gleaned from the decoder. The example that follows will show how these tie together.

In this example, we will walk through the decoding one step at a time. Since the receiver doesn’t know whether a logic 1 or 0 has been sent, initially it must assume a likelihood of 50 - 50 so the initial value of L(d), the a priori knowledge of the state of the data will be 0 for each of the four data bits sent. This is shown in figure 4-16 along with the channel measurements. These are the initial conditions at the receiver showing the first two components: channel measurement LLR (Lc(x)) and the a priori knowledge LLR (L(dx)) where x is 1 - 4 representing the four data bits. .

initial L(d)	
L(d1)	L(d2)
0	0
L(d3)	L(d4)
0	0

channel measurement		
Lc(x1)	Lc(x2)	Lc(x12)
1.5	0.1	2.5
Lc(x3)	Lc(x4)	Lc(x34)
0.2	0.3	2
Lc(x13)	Lc(x24)	
6	1	

Figure 4-16 Initial Conditions

The third component, the extrinsic LLR, is calculated as shown below, starting with the horizontal extrinsic LLR (Leh(dx)).

$$\text{Leh}(\hat{d}_1) = [\text{Lc}(x_2) + \text{L}(d_2)] \boxplus \text{Lc}(x_{12})$$
$$\text{Leh}(\hat{d}_2) = [\text{Lc}(x_1) + \text{L}(d_1)] \boxplus \text{Lc}(x_{12})$$
$$\text{Leh}(\hat{d}_3) = [\text{Lc}(x_4) + \text{L}(d_4)] \boxplus \text{Lc}(x_{34})$$
$$\text{Leh}(\hat{d}_4) = [\text{Lc}(x_3) + \text{L}(d_3)] \boxplus \text{Lc}(x_{34})$$

Where + is normal addition
$\boxplus$ is Log Likelihood addition

Note that there are two forms of addition: normal addition (arithmetic) and log-likelihood addition. The following example clarifies their operation.

$$Leh(\hat{d}_1) = [Lc(x_2) + L(d_2)] \boxplus Lc(x_{12})$$

$$Leh(\hat{d}_1) = \underbrace{[0.1 + 0]}_{\textit{Normal addition}} \boxplus 2.5$$

$$Leh(\hat{d}_1) = \underbrace{[0.1] \boxplus 2.5}_{\textit{Log Likelihood addition}}$$

Log Likelihood addition is approximated as follows:

$$Leh(\hat{d}_1) = \underbrace{[0.1]}_{LLR1} \boxplus \underbrace{2.5}_{LLR2}$$

$$LLR1 + LLR2 = -1 * sign(LLR1) * sign(LLR2) * MIN(|LLR1|, |LLR2|)$$

$$Leh(\hat{d}_1) = -1 * (+1) * (+1) * MIN(|0.1|, |2.5|) = -0.1$$

Following this method, the Leh(dx) results are as shown in figure 4-17.

initial L(d)		
L(d1)	L(d2)	
0	0	
L(d3)	L(d4)	
0	0	
channel measurement		
Lc(x1)	Lc(x2)	Lc(x12)
1.5	0.1	2.5
Lc(x3)	Lc(x4)	Lc(x34)
0.2	0.3	2
Lc(x13)	Lc(x24)	
6	1	
Leh		
Leh(d1)	Leh(d2)	
-0.1	-1.5	
Leh(d3)	Leh(d4)	
-0.3	-0.2	

Figure 4-17 Horizontal Extrinsic LLR

Using the following equations and the same math operations as before, the vertical extrinsic LLR, designated Lev(dx), can be determined with one exception: *the initial L(d) is replaced with the Leh(dx)*

values just calculated. Now we will begin to see how the extrinsic LLRs improve decoding and / or confidence.

$$\mathbf{Lev}(\hat{d}_1) = [Lc(x_3) + L(d_3)] \boxplus Lc(x_{13})$$
$$\mathbf{Lev}(\hat{d}_2) = [Lc(x_4) + L(d_4)] \boxplus Lc(x_{24})$$
$$\mathbf{Lev}(\hat{d}_3) = [Lc(x_1) + L(d_1)] \boxplus Lc(x_{13})$$
$$\mathbf{Lev}(\hat{d}_4) = [Lc(x_2) + L(d_2)] \boxplus Lc(x_{24})$$

The results are shown in figure 4-18.

initial L(d)	
L(d1)	L(d2)
0	0
L(d3)	L(d4)
0	0

channel measurement		
Lc(x1)	Lc(x2)	Lc(x12)
1.5	0.1	2.5
Lc(x3)	Lc(x4)	Lc(x34)
0.2	0.3	2
Lc(x13)	Lc(x24)	
6	1	

Leh	
Leh(d1)	Leh(d2)
-0.1	-1.5
Leh(d3)	Leh(d4)
-0.3	-0.2

next L(d)	
L(d1)	L(d2)
-0.1	-1.5
L(d3)	L(d4)
-0.3	-0.2

Lev	
Lev(d1)	Lev(d2)
0.1	-0.1
Lev(d3)	Lev(d4)
-1.4	1

Figure 4-18 Vertical Extrinsic LLR

Note that the "next L(d)" that is used for determining the vertical extrinsic LLR, Lev(dx), is the same as the horizontal extrinsic LLR, Leh(dx), just calculated. A soft decision can now be made using the equation below.

$$L(\hat{d}) = Lc(x) + Leh(\hat{d}) + Lev(\hat{d})$$

where $\hat{L}(d)$ = soft decision LLR

$Lc(x)$ = channel measurement LLR

$Leh(\hat{d})$ = horizontal extrinsic LLR

$Lev(\hat{d})$ = vertical extrinsic LLR

The final results, along with a decode of the soft decision is shown in figure 4-20. This represents the first iteration of the turbo decoder. The decode is just an observation of the soft decision polarities: a + indicates a logic 1 while a – indicates a logic 0. Note that the decode, 1001, matches the original data sent with a confidence as defined by the magnitude of the soft decision.

initial L(d)	
L(d1)	L(d2)
0	0
L(d3)	L(d4)
0	0

channel measurement		
Lc(x1)	Lc(x2)	Lc(x12)
1.5	0.1	2.5
Lc(x3)	Lc(x4)	Lc(x34)
0.2	0.3	2
Lc(x13)	Lc(x24)	
6	1	

Leh	
Leh(d1)	Leh(d2)
-0.1	-1.5
Leh(d3)	Leh(d4)
-0.3	-0.2

next L(d)	
L(d1)	L(d2)
-0.1	-1.5
L(d3)	L(d4)
-0.3	-0.2

Lev	
Lev(d1)	Lev(d2)
0.1	-0.1
Lev(d3)	Lev(d4)
-1.4	1

soft decision	
1.5	-1.5
-1.5	1.1

decode of soft decision	
1	0
0	1

Figure 4-19 First Iteration

Turbo decoders iterate several times to improve decode and / or confidence. Figure 4-20 illustrates a second iteration, along with the first. Note that the starting values of L(d) for this second iteration is the values of Lev(dx) from the first iteration. All other processes are the same. Recall that the data was successfully decoded with the first iteration. The second iteration retains that decode but improves the confidence as indicated by the increased magnitude of the soft decision values.

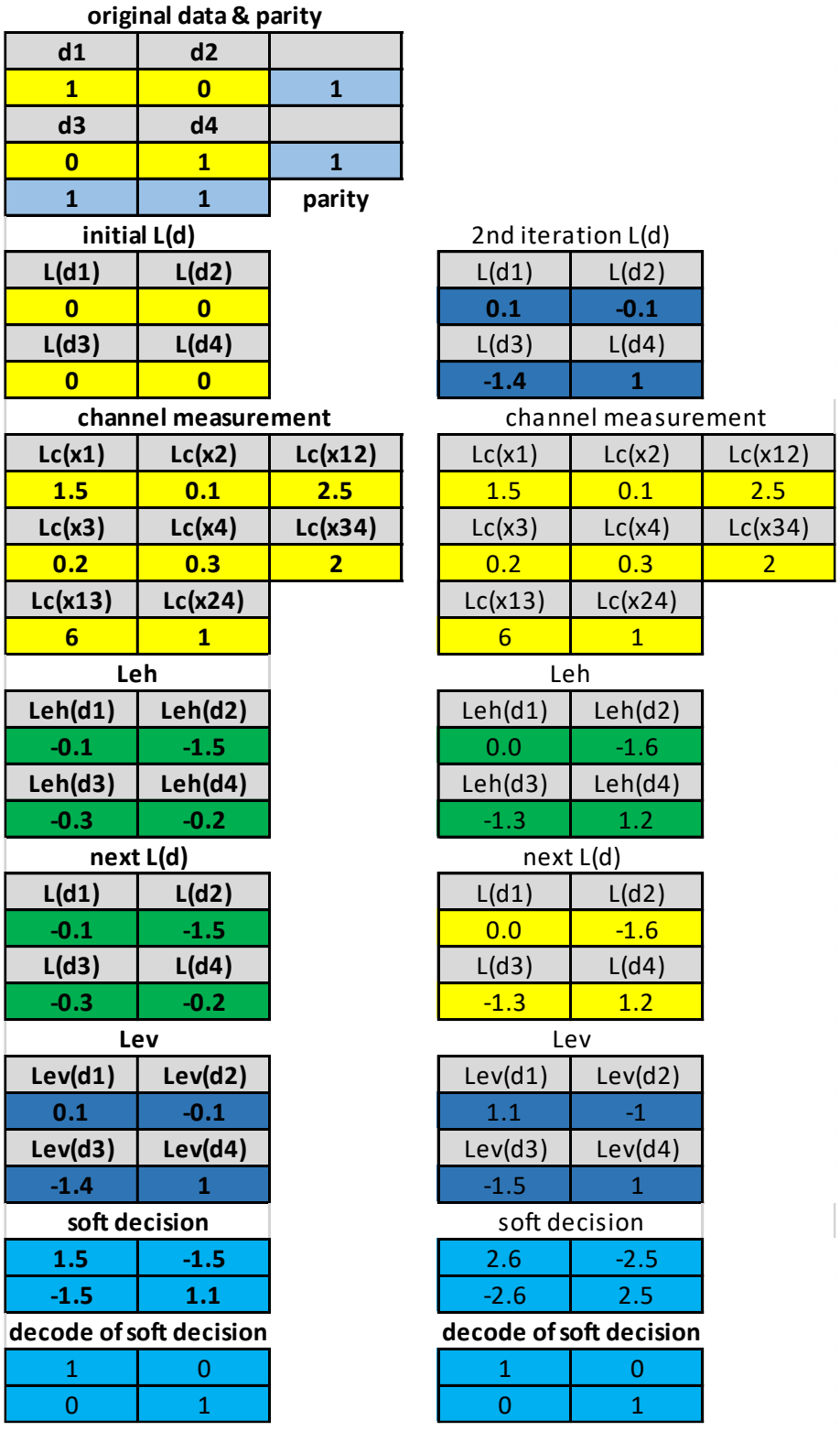


Figure 4-20 Second Iteration

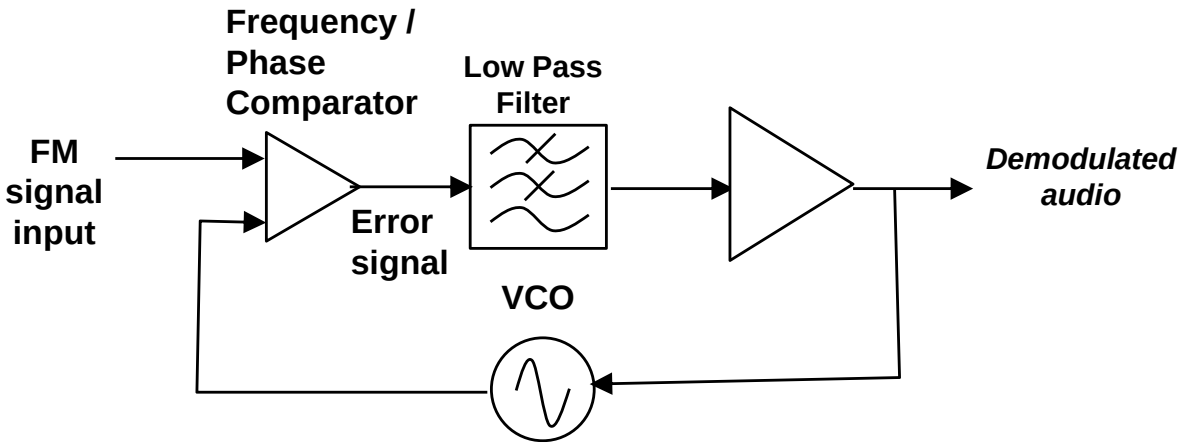
A comparison of turbo coding to other methods is in order. Consider a k = 7, rate ½ convolutional encoder which requires an Eb/No of 4.5 dB to achieve a BER of 10^{-5} in AWGN using soft decision Viterbi decoding. A turbo decoder with one iteration provides comparable performance. However, double the number of iterations to 2 and the required Eb/No drops to about 2.8 dB; increase to 3 iterations and it drops to about 1.8 dB; 5 iterations drops it to about 1 dB! Significant improvements indeed! Turbo decoder BER performance drops so rapidly it is often referred to as the "turbo cliff". It should be noted that this same "cliff effect" is seen with other forms of encoding as well.

AM Demodulation

AM can be demodulated with a diode followed by a capacitor – the diode rectifies the AM intermediate frequency and the capacitor filters it, recovering the original modulating signal. This is then passed on to the audio amplifiers.

Frequency Demodulation

There are many kinds of FM demodulators, one of the classics being the Foster-Seeley discriminator. In modern radios, that has been replaced by using phased lock loops technology. Consider the PLL arrangement below (see the section on PLL's for further detail).



The frequency modulated (FM) signal input is applied to one input of the frequency / phase comparator. The other input is from the VCO which is set to the same carrier frequency as that of the FM signal. Since the FM signal is varying in frequency at an audio rate, the error signal out of the frequency / phase comparator will also be at the same audio rate, thus effectively demodulating the received signal. The low pass filter blocks any of the carrier frequency allowing only the lower frequency audio to pass through. Finally, the audio signal is amplified and sent out as the demodulated audio.

BPSK / QPSK Demodulation

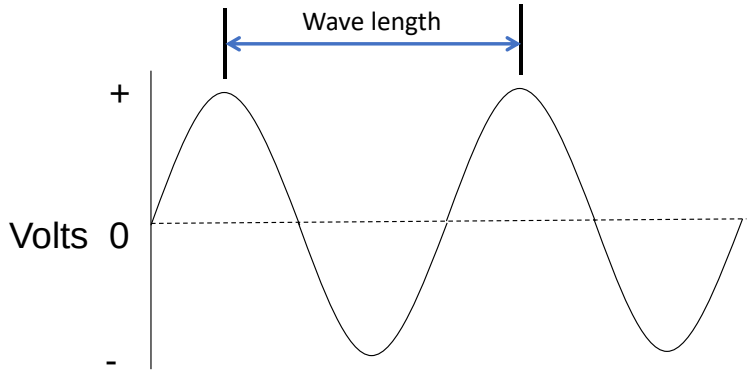
The basic concept is to determine the phase shift of the received signal. One form (there are others) is called coherent demodulation which means as part of the demodulation process, the receiver must recover the "at rest" (no modulation) phase of the carrier to use as a reference point in its demodulation.

This allows the receiver to have a reference phase to compare to the received phase to determine how much phase shift has occurred. For BPSK, the receiver would need to determine if a 180-degree phase shift occurred or not. For QPSK, the receiver would need to determine if phase shifts of 90, 180, 270 degrees or no phase shift has occurred. Knowing what modulating bit patterns are associated with each detected phase shift allows the receiver to recover the transmitted data.

Chapter 18: Antennas and RF Propagation

Wavelength

Electromagnetic fields are like AC signals - they rise in amplitude to some peak positive value, decrease back to zero, continue decreasing to a peak negative amplitude, and, finally, increase back zero. The distance between any two similar points on the AC signal – say from peak positive to the next peak positive – is the wavelength, abbreviated by the Greek symbol lambda, λ**.**



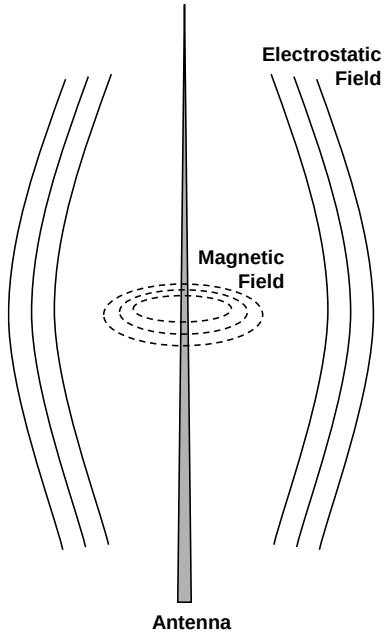
Wavelength is related to frequency as follows: $\lambda = c / f$ where λ = wavelength, in meters in this case since C is in meters/sec, C = speed of light (approximately 300,000,000 meters per second), and f = frequency, in Hz

Now we can tie wavelength and frequency together. Suppose the frequency of the RF carrier were 150 MHz. The wavelength would be: $\lambda = c/ f$ = (300E6 m/s) / (150E6 Hz) = 2 meters

Wave length is an important factor in antennas.

Antenna

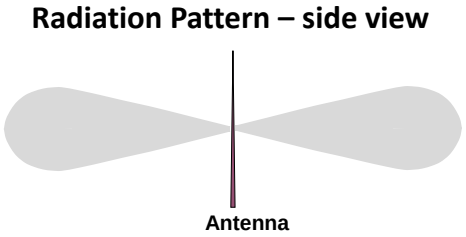
An antenna is a conductor that can radiate and receive RF signals. When an RF carrier is applied to an antenna, two fields develop around it. One field is called the electrostatic field which is like the force field between the two plates of a charged capacitor. The lines of force for this field are parallel to the antenna, as shown below.

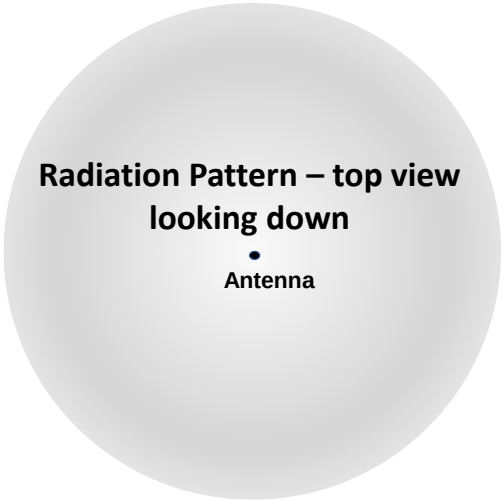


The current flowing through the antenna sets up a magnetic field around the antenna. This is also shown in the same figure. Notice that the electrostatic and magnetic fields are at right angles, or 90 degrees to each other. These two fields combine to make the electromagnetic field which is radiated out of the antenna. An antenna acts like a series resonant RLC circuit. It can be tuned to operate at specific frequencies depending on its length.

Radiation Pattern and Antenna Gain

The energy radiated from antennas creates specific patterns or areas over which the radiated energy is spread. This is called a radiation pattern. For example, the antenna depicted above would have an omnidirectional radiation pattern, meaning it would radiate energy 360 degrees around it, with null, or area of no radiation, directly above it, as shown below.





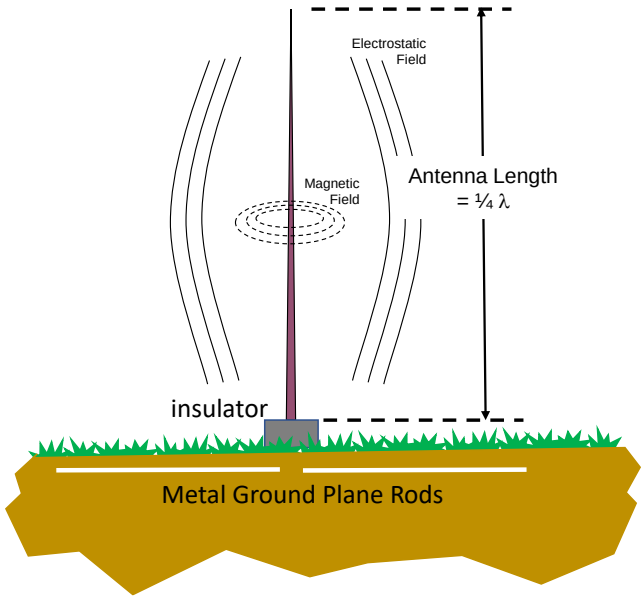
The figure above offers two views – one from the side of the antenna and one from the top of the antenna looking straight down towards ground. In the top view, one can see the omnidirectional pattern with RF energy radiating out all around the antenna in a circular pattern. In the side view, the RF energy is confined to a figure eight pattern rather than a circle. This radiation pattern is often referred to as a doughnut pattern.

Antennas can introduce signal gain into the radio system. They do this by focusing energy over a smaller area much like a flash light whose lens focus the light into a narrow beam. It is this ability of an antenna to "focus" the RF energy that gives it "gain". The figure eight pattern in the side view above does this. Typical omnidirectional "whip" antennas on 2-meter hand held transceivers can provide around 2 dB of gain this way.

Antenna Dimensions

Antenna length is based on the wave length of the frequency that will be used to transmit and receive. For example, if we physically cut an antenna so that it appeared to be 2 meters long at 150 MHz, the antenna would pick up or receive RF signals at 150 MHz. It would act like a series resonant RLC circuit tuned to 150 MHz. In fact, it wouldn't have to be a wavelength long. It could be some multiple of a wavelength such as twice as long or half as long. To keep the antenna small, the antenna is generally cut to be shorter.

There are two basic types of antennas. One is the Marconi antenna, which is cut to appear as a quarter wavelength (λ / 4) at the desired frequency. A Marconi antenna must have a good ground plane underneath it, insulated from the antenna itself. This ground plane is a conductive plane that forms a mirror quarter-wavelength "section" underneath the antenna. It consists of heavy metal rods buried in the ground radiating out from the antenna base, in a 360-degree pattern around the antenna base. This provides good conductivity to the ground. The antenna is feed between the base of the antenna (at the insulator) and the ground plane. The figure below provides a cross sectional view.



On a handheld portable FM transceiver the ground plane is usually the metal case. The second antenna type is a Hertz antenna. The Hertz antenna is cut to appear as a half wavelength at the desired frequency and is driven from one end.

Antenna Polarity

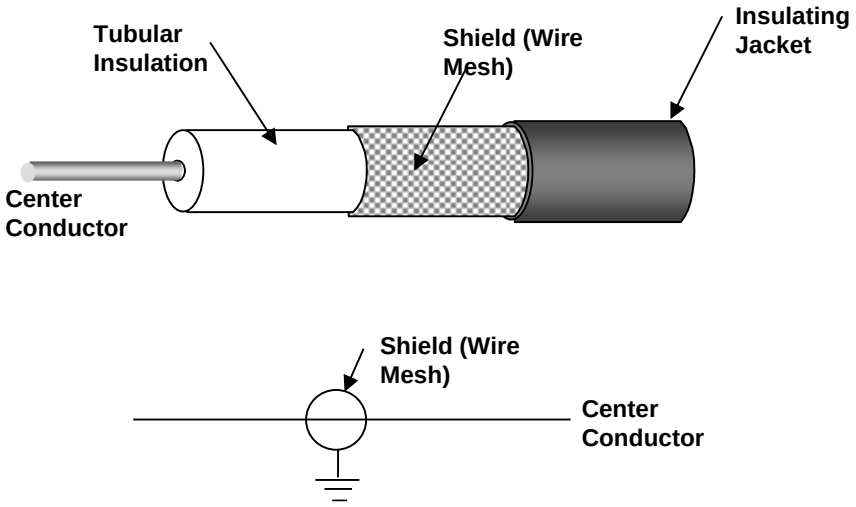
The orientation of the antenna with respect to local ground determines the antenna polarity. If the antenna is oriented perpendicular to the ground, as shown in the figure above, it is vertically polarized. If the antenna is orientated parallel to the ground, it is horizontally polarized. AM and FM commercial broadcast use vertical polarization. VHF two-way communication systems are usually vertically polarized. Shortwave broadcast and amateur radio HF (1 – 30 MHz range) are usually horizontally polarized. This enables them to launch RF waves up to the ionosphere as sky waves. The important thing to remember is that the polarization of the transmitting and receiving antennas must be the same for best reception.

Signal Bandwidth

All of these various forms of modulation discussed earlier will occupy a specific range or band of frequencies. Band width is the range of frequencies over which the modulation spreads. For example, SSB voice modulation spreads over 3 KHz for a 3 KHz bandwidth. The bandwidth for a 2-meter FM transceiver is around 10 to 15 KHz. Analog fast scan television on the amateur radio 70-centimeter band can be as high as 6 MHz. The terms 2-meter band and 70-centimeter band refer to the approximate wavelength associated with that band. The antennas, although tuned to specific frequencies, must have sufficient bandwidth to support the required number of channels and their modulation signal bandwidth.

Transmission Line

Many times an antenna is not located at the same place as the transmitter / receiver equipment. A transmission line is use get the RF carrier power from the equipment to the antenna. Transmission line can take many forms but one of the most common is a coaxial cable or coax. This transmission line is a two-wire conductor especially designed to carry RF energy with minimum loss. A coax cable construction is illustrated below along with its schematic symbol.



The coax has a center conductor surrounded by a tubular insulator. This tubular insulator is covered by a wire mesh called a shield, which serves as the second conductor. The shield is covered with an insulating jacket. Coax is often used with a Marconi antenna. The center conductor is connected to the bottom of the antenna which is insulated from ground and the shield is connected to the ground plane.

Standing Wave Ratio (SWR)

Transmission lines and antennas both have impedance (see lesson on impedance). The impedance of the transmission line is called the characteristic or surge impedance. Recall that maximum power transfer occurs when source and load impedances are equal. Therefore, the radio has to match the impedance of the transmission line and the transmission line has to match the impedance of the antenna. In this way, maximum power is transferred from the radio to the antenna. The power traveling down the transmission line from the radio to the antenna is called the incident power. If the impedances are not equal, some of the power is reflected and travels back to the radio. This is called the reflected power. The interaction of these two waves creates a standing wave on the transmission line. If we measured the maximum and minimum voltage of the standing wave, we could calculate the standing wave ratio, abbreviated SWR. SWR = E_{MAX} / E_{MIN}. The SWR tells us how well the impedances of the radio, transmission line, and antenna are matched. If they are matched perfectly, SWR = 1:1 (usually this is stated as SWR = 1). In the real world, such perfection is not possible, but an SWR of 1:1.15 (SWR = 1.15) is typical for a good system. The higher the SWR, the worse impedance match we have. In transmitting, this means less radiated power. In receiving, it means less received power is coupled from antenna to radio.

Decibels

Radio work often uses “decibels”, abbreviated “dB”. It is the logarithm or LOG of the ratio of two signal levels multiplied by 10 if we are dealing with power or by 20 if we are dealing with voltage or current. For those not familiar with logarithms, see the Math Appendix.

dB = 10 LOG (Power 1, in Watts / Power 2, in Watts) for power

dB = 20 LOG (Voltage 1, in Volts / Voltage 2, in Volts) for voltages

Two standard units of measure are dBW meaning the power level with respect to one Watt and dBm meaning the power level with respect to one milliwatt. For example, a 5-Watt transmitter would provide 10 LOG (5 Watts / 1 Watt) = ~ 7 dBW or 10 LOG (5 Watts / 1 milliwatt) = ~ 37 dBm.

There are a few rules of thumb: every 3 dB increase in power means the power has doubled; conversely, every 3 dB decrease in power means the power has cut in half. A 10 dB increase is power means the power has increased by ten times. A 10 dB decrease in power means the power has decreased to 1/10 of what it was originally. With these rules in mind, it is easy to make some rough estimates. For example, if a transmitter’s power is changed from 12 Watts to 3 Watts, this would represent a 6 dB loss, determined as follows. Cutting 12 Watts in half to 6 Watts would be one, 3 dB loss. Cutting the 6 Watts in half again to 3 Watts would be another 3 dB loss for a total of 6 dB. RF losses are also often measured in dB. For example, if a coaxial cable had 1.5 dB of loss at a given frequency, the power will be reduced to 0.707 of its original value.

RF Propagation

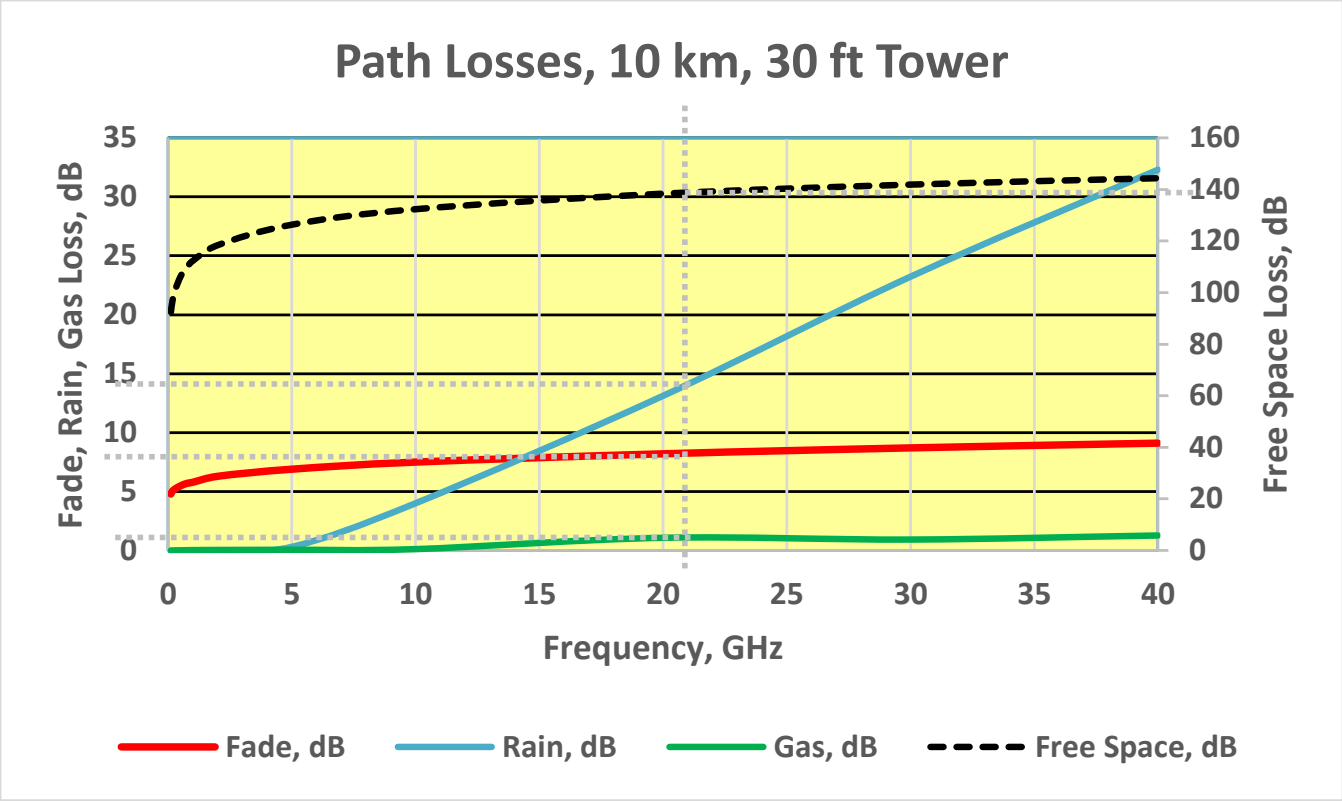
There are three types of RF waves: ground waves, direct waves, and sky waves. Ground waves hug the surface of the Earth and occur in the KHz range such as commercial AM broadcast. Direct waves are line of sight, that is, the travel in a "straight" line (sort of, they do bend some with the Earth's curvature). Commercial FM broadcast and VHF radios use direct waves. Sky waves are those that bounce off the ionosphere and are reflected back to Earth hundreds of kilometers away. This includes commercial short-wave broadcast and HF frequencies (2 – 30 MHz).

The greatest source of loss, at all frequencies, is free space loss. Free space loss is the natural decrease in signal strength as the RF energy spreads out over a greater and greater area as the distance from the source increases. The equation for free space loss is as follows –

Loss, dB = 32.4 + 20 LOG (f) + 20 LOG (path)

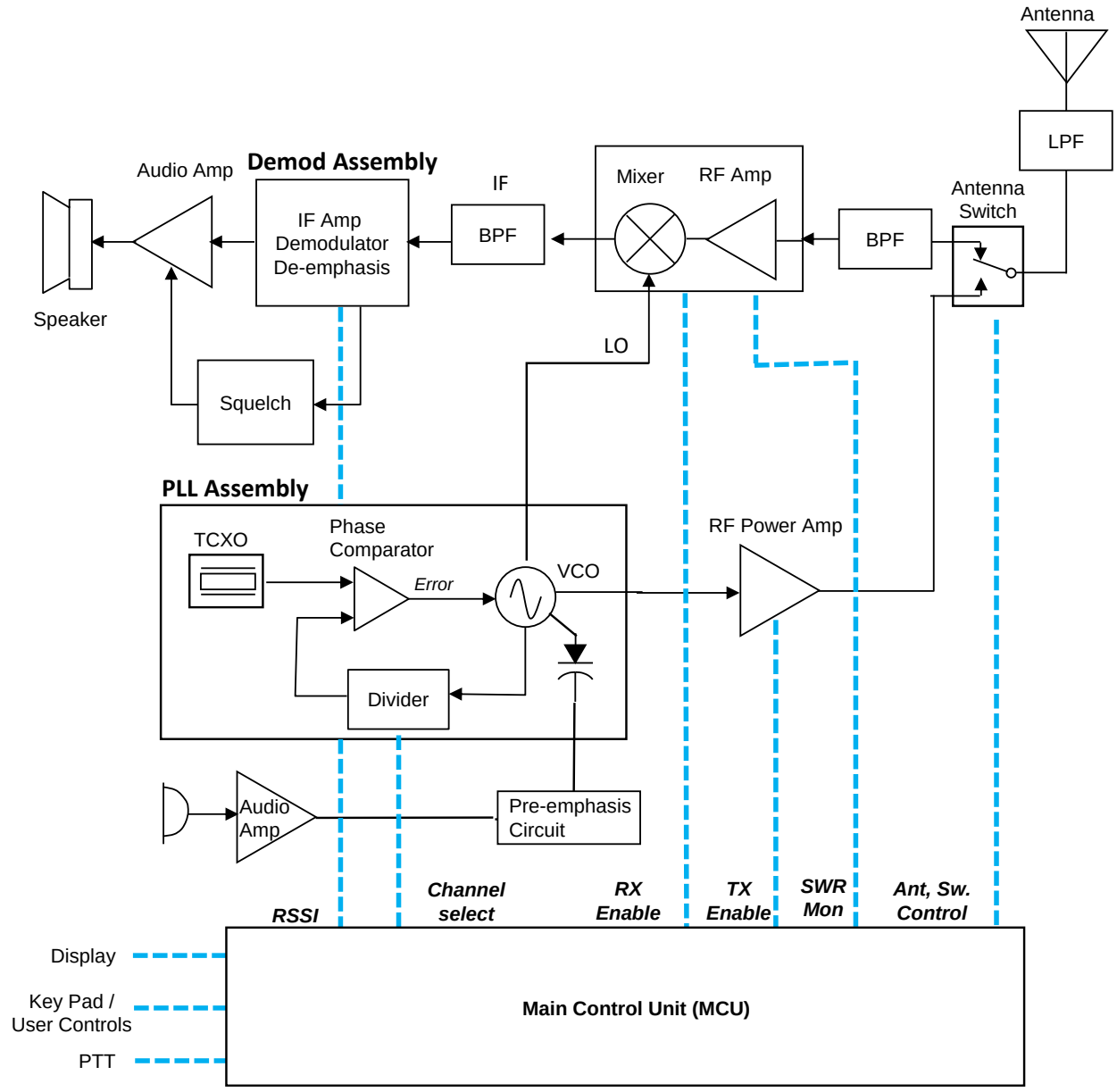
Where f = frequency, in ***MHz***
Path = distance, in ***km***

Students not familiar with logarithms (log) should refer to the Math Appendix. With paths engineered to avoid obstructions and destructive reflections, other losses include fade (due to atmospheric thermal gradients), rain, and gas. An example is provided in the graph below, where losses are calculated for a 10 km path in the SE United States with an antenna on a 30-foot tower at both transmit and receive ends, no path obstructions such as hills or buildings, no destructive reflections, and severe rain conditions. The losses for rain, gas, and fade are read from the left-hand vertical axis and the free space loss is read from the right-hand vertical axis of the chart. For example, what are the losses at 21 GHz? Find 21 GHz on the horizontal frequency axis, marked by a dotted gray line running vertically at the 21 GHz mark. Follow the vertical gray dotted line until it intercepts the gas attenuation line (green line). From this point, follow the horizontal gray dotted line to the left and you will see it intercepts the vertical axis at about 1 dB. This means this path would have 1 dB of gas loss at 21 GHz. Using this same approach, you will find that the fade loss (red line) is about 8 dB and rain loss (blue line) is just under 15 dB. When you reach the free space loss (black dotted line), you go to the *right-hand* scale and read about 140 dB loss. Thus, it can be seen that free space loss dominates the other losses by over 100 dB!


Path Losses, 10 km, 30 ft Tower
Fade, Rain, Gas Loss, dB
Free Space Loss, dB
Frequency, GHz
0
5
10
15
20
25
30
35
40
0
20
40
60
80
100
120
140
160
Fade, dB
Rain, dB
Gas, dB
Free Space, dB

Chapter 19: FM Transceiver Block Diagram

In this section, we will review the block diagram of a typical 2-meter FM hand held transceiver, to see how a basic radio works, putting together many of the components already discussed. Throughout the discussion, refer to the block diagram below.



The Main Control Unit (MCU) is a microcontroller that is the brains of the radio. It interfaces with the display, key pad, users' controls, and push to talk (PTT) switch. The MCU translates user commands to the specific set of signals required to operate the radio such as channel selection, receive enable, transmit enable, antenna switch control, etc. It also provides feedback to the user via the display. For example, the

receiver section provides a received signal strength indication (RSSI). The MCU takes this information and provides the user a visual indication on the display of the received signal strength. The MCU also provides protective measures. For example, the transmit RF power amplifier provides an indication of SWR to the MCU. MCU monitors the status of the SWR and if it gets too high, indicating a poor impedance match, it can take measures to protect the RF power amplifier such as terminating transmission or reducing transmit power. The MCU also monitors the battery status, providing a visual indication of battery charge on the display.

The radio requires many signals at different frequencies to operate: for example, the transmit frequency and one or more local oscillator (LO) frequencies for the mixers in the receiver. In this example, the PLL Assembly is the source of all of these for both transmit and receive. A stable reference source is provided by the temperature-controlled crystal oscillator (TCXO). This provides the reference input to the PLL / VCO to generate the channel transmit frequency. In the receive mode, the PLL / VCO provides the LO signal for the mixer. If a second, but lower LO frequency is needed for a second mixer, it is often derived directly from the TCXO. In this example, frequency modulation, via a varactor diode, is included in the PLL Assembly.

The transmit path starts with the microphone and push to talk (PTT) switch. The user presses the PTT switch and speaks into the microphone. The microphone converts the audio into an electrical signal which is amplified and processed, applying the necessary pre-emphasis. This signal is then applied to the PLL Assembly varactor diode where it frequency modulates the RF carrier directly. The modulated RF carrier is amplified in the power amplifier and passed through the antenna switch and low pass filter (LPF) to the antenna for radiation. The MCU makes sure the antenna switch is set to allow the transmit RF signal through.

The receive processing begins with the antenna intercepting the RF signal and passing it through the low pass filter (LPF), the antenna switch, and the band pass filter (BPF) to the RF amplifier and then to the first – and in this case, only – mixer. The MCU makes sure the antenna switch is set to allow the receive RF signal through. The output of the mixer is the RF signal, the LO signal, the sum of the two, and the difference of the two. The band pass filter selects the difference of the two as the intermediate frequency and passes it on to the Demod Assembly. There, the IF is amplified by an IF amplifier, demodulated to recover the audio, and processed to reverse the effects of the transmitter pre-emphasis. The recovered audio is passed to the audio amplifier which drives the speaker. Parallel to this is the squelch circuit. If there is no received RF signal or if it is too weak, the squelch circuit will mute the speaker so the user is not subject to tedious back ground noise.