1. Semiconductors

1.AMPLIFIERS AND ACTIVE DEVICES

Amplifiers

The practical benefit of active devices is their *amplifying* ability. Whether the device in question be voltage-controlled or current-controlled, the amount of power required of the controlling signal is typically far less than the amount of power available in the controlled current. In other words, an active device doesn't just allow electricity to control electricity; it allows a *small* amount of electricity to control a *large* amount of electricity.

Amplifier gain

Because amplifiers have the ability to increase the magnitude of an input signal, it is useful to be able to rate an amplifier's amplifying ability in terms of an output/input ratio. The technical term for an amplifier's output/input magnitude ratio is *gain*. As a ratio of equal units (power out / power in, voltage out / voltage in, or current out / current in), gain is naturally a unitless measurement. Mathematically, gain is symbolized by the capital letter "A".

For example, if an amplifier takes in an AC voltage signal measuring 2 volts RMS and outputs an AC voltage of 30 volts RMS, it has an AC voltage gain of 30 divided by 2, or 15:

$$A_V = \frac{V_{output}}{V_{input}}$$

$$A_V = \frac{30\ V}{2\ V}$$

$$A_V = 15$$

Correspondingly, if we know the gain of an amplifier and the magnitude of the input signal, we can calculate the magnitude of the output. For example, if an amplifier with an AC current gain of 3.5 is given an AC input signal of 28 mA RMS, the output will be 3.5 times 28 mA, or 98 mA:

$$I_{output} = (A_V)(V_{input})$$

$$I_{output} = (3.5)(28\ mA)$$

$$I_{output} = 98\ mA$$

In the last two examples I specifically identified the gains and signal magnitudes in terms of "AC." This was intentional, and illustrates an important concept: electronic amplifiers often respond differently to AC and DC input signals, and may amplify them to different extents.

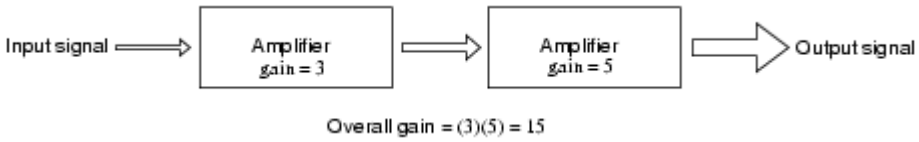
Another way of saying this is that amplifiers often amplify *changes* or *variations* in input signal magnitude (AC) at a different ratio than *steady* input signal magnitudes (DC). The specific reasons for this are too complex to explain at this time, but the fact of the matter is worth mentioning. If gain calculations are to be carried out, it must first be understood what type of signals and gains are being dealt with, AC or DC.

Electrical amplifier gains may be expressed in terms of voltage, current, and/or power, in both AC and DC. A summary of gain definitions is as follows. The triangle-shaped "delta" symbol (Δ) represents *change* in mathematics, so "$\Delta V_{output} / \Delta V_{input}$" means "change in output voltage divided by change in input voltage," or more simply, "AC output voltage divided by AC input voltage":

	DC gains	AC gains
Voltage	$A_V = \frac{V_{output}}{V_{input}}$	$A_V = \frac{\Delta V_{output}}{\Delta V_{input}}$
Current	$A_I = \frac{I_{output}}{I_{input}}$	$A_I = \frac{\Delta I_{output}}{\Delta I_{input}}$
Power	$A_P = \frac{P_{output}}{P_{input}}$	$A_P = \frac{(\Delta V_{output})(\Delta I_{output})}{(\Delta V_{input})(\Delta I_{input})}$
	$A_P = (A_V)(A_I)$	

Δ = *"change in . . ."*

If multiple amplifiers are staged, their respective gains form an overall gain equal to the product (multiplication) of the individual gains:

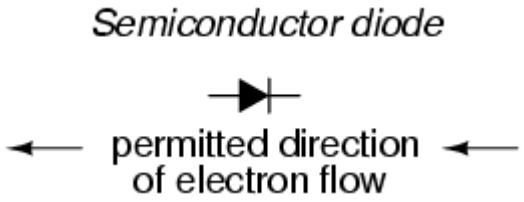


Overall gain = (3)(5) = 15

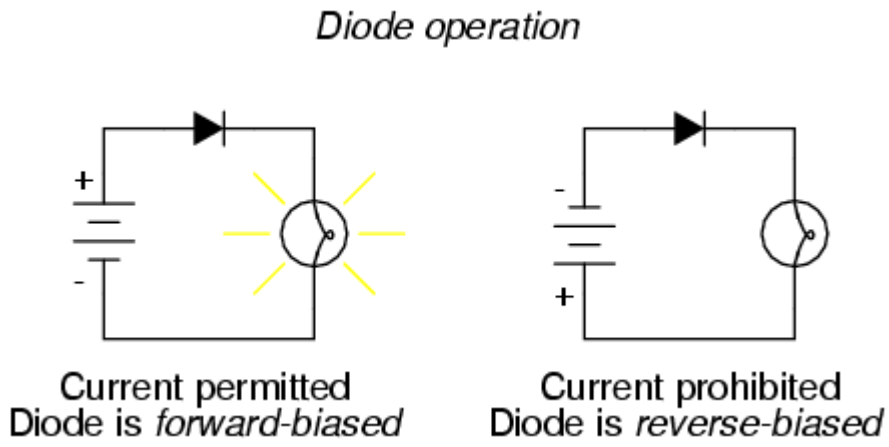
2.DIODES AND RECTIFIERS

Introduction

A *diode* is an electrical device allowing current to move through it in one direction with far greater ease than in the other. The most common type of diode in modern circuit design is the *semiconductor* diode, although other diode technologies exist. Semiconductor diodes are symbolized in schematic diagrams as such:



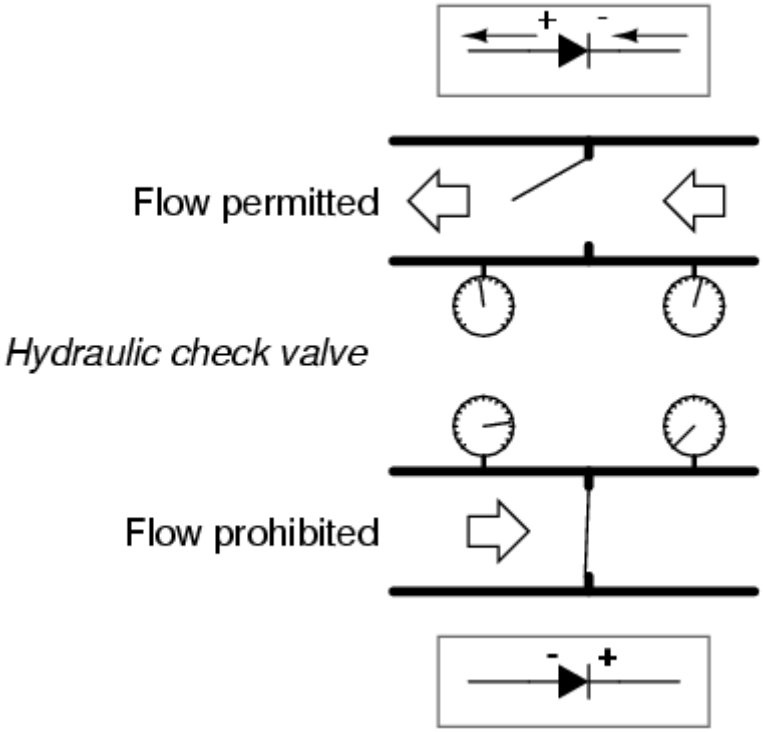
When placed in a simple battery-lamp circuit, the diode will either allow or prevent current through the lamp, depending on the polarity of the applied voltage:



When the polarity of the battery is such that electrons are allowed to flow through the diode, the diode is said to be *forward-biased*. Conversely, when the battery is "backward" and the diode blocks current, the diode is said to be *reverse-biased*. A diode may be thought of as a kind of switch: "closed" when forward-biased and "open" when reverse-biased.

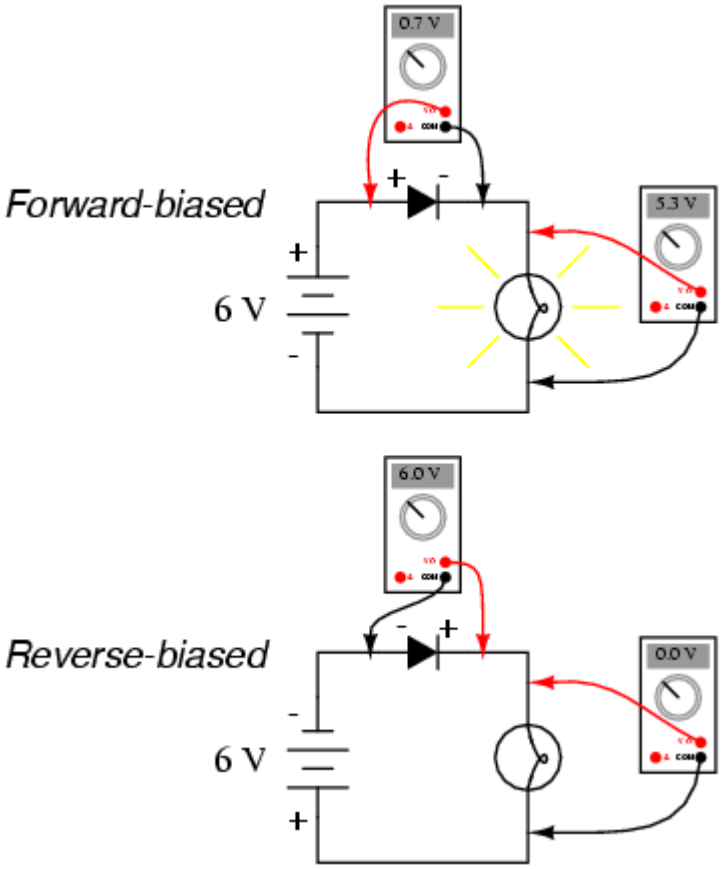
Oddly enough, the direction of the diode symbol's "arrowhead" points *against* the direction of electron flow. This is because the diode symbol was invented by engineers, who predominantly use *conventional flow* notation in their schematics, showing current as a flow of charge from the positive (+) side of the voltage source to the negative (-). This convention holds true for all semiconductor symbols possessing "arrowheads:" the arrow points in the permitted direction of conventional flow, and against the permitted direction of electron flow.

Diode behavior is analogous to the behavior of a hydraulic device called a *check valve*. A check valve allows fluid flow through it in one direction only:



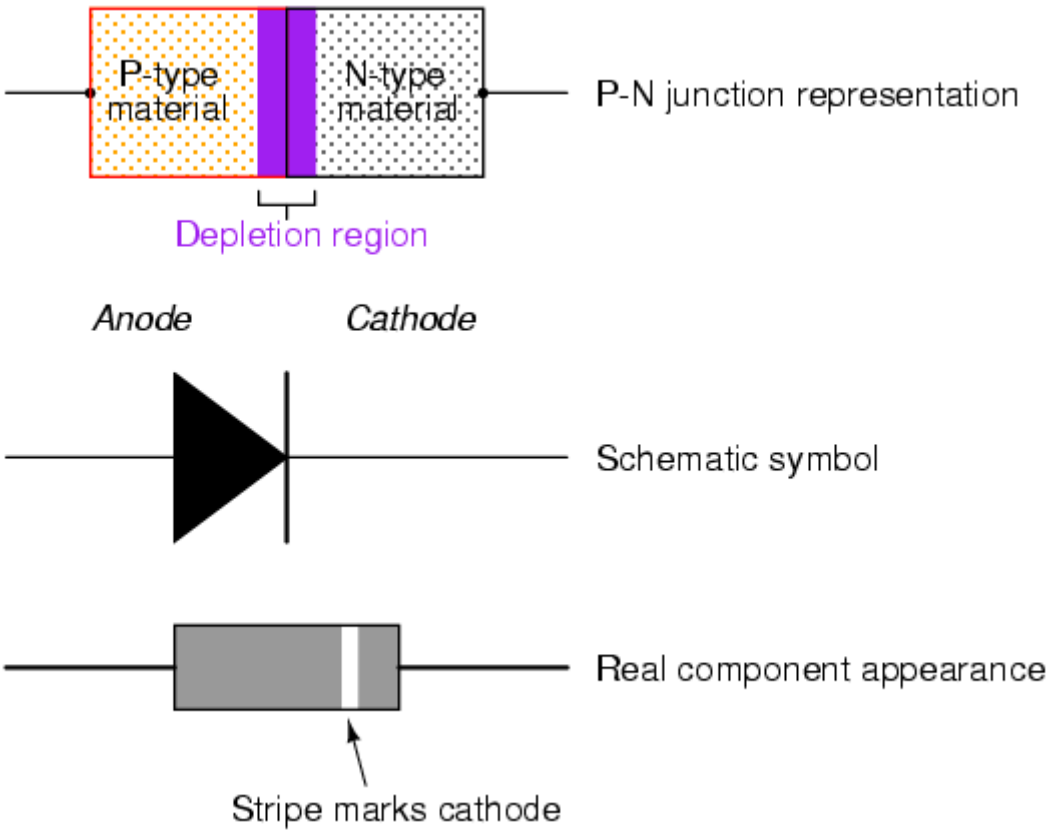
Check valves are essentially pressure-operated devices: they open and allow flow if the pressure across them is of the correct "polarity" to open the gate (in the analogy shown, greater fluid pressure on the right than on the left). If the pressure is of the opposite "polarity," the pressure difference across the check valve will close and hold the gate so that no flow occurs.

Like check valves, diodes are essentially "pressure-" operated (voltage-operated) devices. The essential difference between forward-bias and reverse-bias is the polarity of the voltage dropped across the diode. Let's take a closer look at the simple battery-diode-lamp circuit shown earlier, this time investigating voltage drops across the various components:

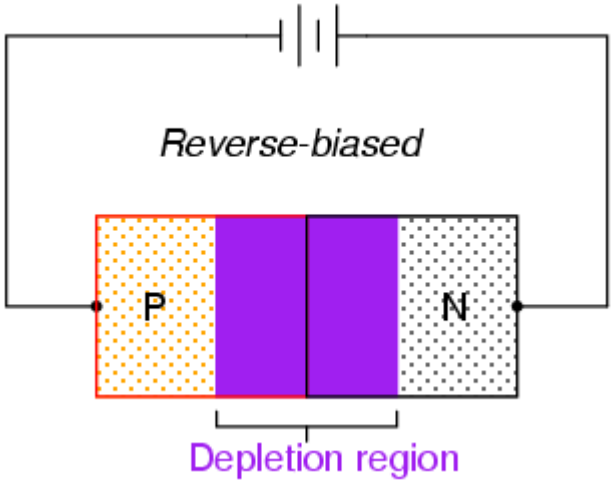


When the diode is forward-biased and conducting current, there is a small voltage dropped across it, leaving most of the battery voltage dropped across the lamp. When the battery's polarity is reversed and the diode becomes reverse-biased, it drops *all* of the battery's voltage and leaves none for the lamp. If we consider the diode to be a sort of self-actuating switch (closed in the forward-bias mode and open in the reverse-bias mode), this behavior makes sense. The most substantial difference here is that the diode drops a lot more voltage when conducting than the average mechanical switch (0.7 volts versus tens of millivolts).

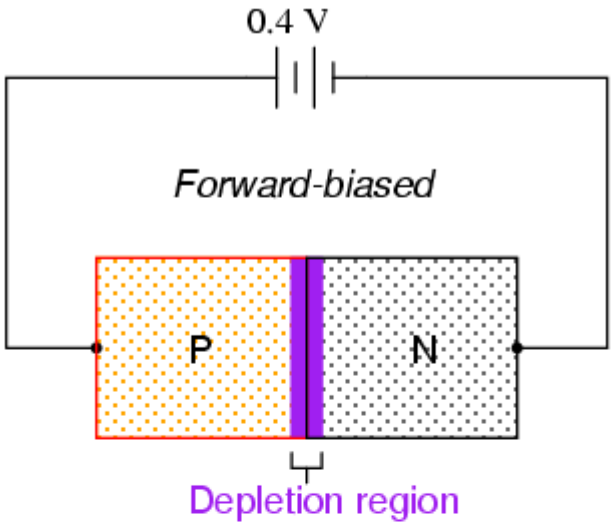
This forward-bias voltage drop exhibited by the diode is due to the action of the depletion region formed by the P-N junction under the influence of an applied voltage. When there is no voltage applied across a semiconductor diode, a thin depletion region exists around the region of the P-N junction, preventing current through it. The depletion region is for the most part devoid of available charge carriers and so acts as an insulator:

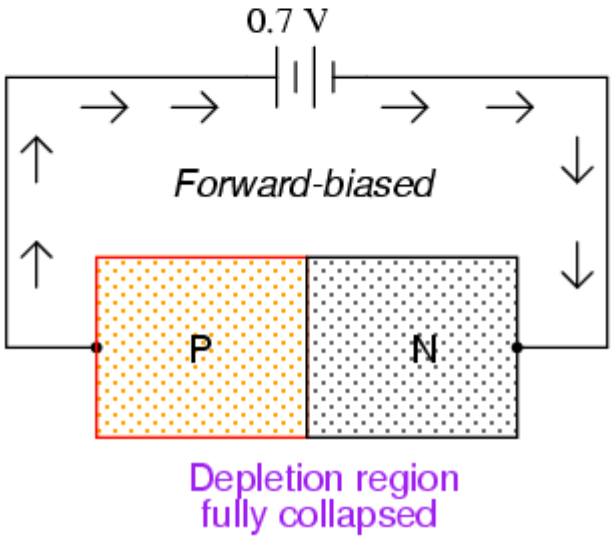


If a reverse-biasing voltage is applied across the P-N junction, this depletion region expands, further resisting any current through it:



Conversely, if a forward-biasing voltage is applied across the P-N junction, the depletion region will collapse and become thinner, so that the diode becomes less resistive to current through it. In order for a sustained current to go through the diode, though, the depletion region must be fully collapsed by the applied voltage. This takes a certain minimum voltage to accomplish, called the *forward voltage*:





For silicon diodes, the typical forward voltage is 0.7 volts, nominal. For germanium diodes, the forward voltage is only 0.3 volts. The chemical constituency of the P-N junction comprising the diode accounts for its nominal forward voltage figure, which is why silicon and germanium diodes have such different forward voltages. Forward voltage drop remains approximately equal for a wide range of diode currents, meaning that diode voltage drop not like that of a resistor or even a normal (closed) switch. For most purposes of circuit analysis, it may be assumed that the voltage drop across a conducting diode remains constant at the nominal figure and is not related to the amount of current going through it.

In actuality, things are more complex than this. There is an equation describing the exact current through a diode, given the voltage dropped across the junction, the temperature of the junction, and several physical constants. It is commonly known as the *diode equation*:

$$I_D = I_S (e^{qV_D/NkT} - 1)$$

Where,

I_D = Diode current in amps

I_S = Saturation current in amps (typically 1×10^{-12} amps)

e = Euler's constant (~ 2.718281828)

q = charge of electron (1.6×10^{-19} coulombs)

V_D = Voltage applied across diode in volts

N = "Nonideality" or "emission" coefficient (typically between 1 and 2)

k = Boltzmann's constant (1.38×10^{-23})

T = Junction temperature in degrees Kelvin

The equation kT/q describes the voltage produced within the P-N junction due to the action of temperature, and is called the *thermal voltage*, or V_t of the junction. At room temperature, this is about 26 millivolts. Knowing this, and assuming a "nonideality" coefficient of 1, we may simplify the diode equation and re-write it as such:

$$I_D = I_S (e^{V_D/0.026} - 1)$$

Where,

I_D = Diode current in amps

I_S = Saturation current in amps (typically 1×10^{-12} amps)

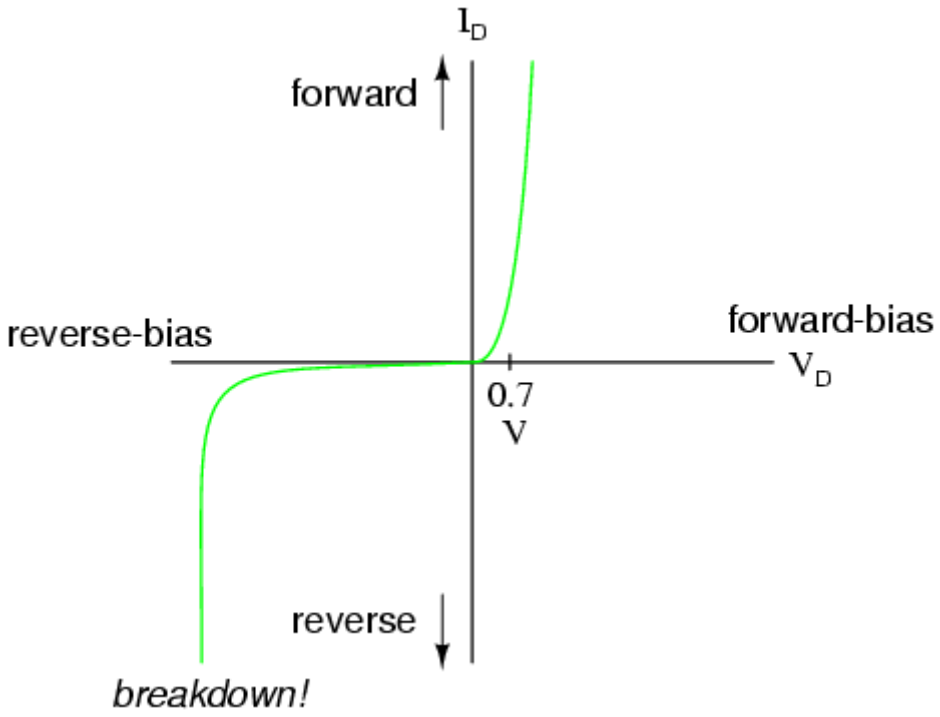
e = Euler's constant (~ 2.718281828)

V_D = Voltage applied across diode in volts

You need not be familiar with the "diode equation" in order to analyze simple diode circuits. Just understand that the voltage dropped across a current-conducting diode *does* change with the amount of current going through it, but that this change is fairly small over a wide range of currents. This is why many textbooks simply say the voltage drop across a conducting, semiconductor diode remains constant at 0.7 volts for silicon and 0.3 volts for germanium. However, some circuits intentionally make use of the P-N junction's inherent exponential current/voltage relationship and thus can only be understood in the context of this equation. Also, since temperature is a factor in the diode equation, a forward-biased P-N junction may also be used as a temperature-sensing device, and thus can only be understood if one has a conceptual grasp on this mathematical relationship.

A reverse-biased diode prevents current from going through it, due to the expanded depletion region. In actuality, a very small amount of current can and does go through a reverse-biased diode, called the *leakage current*, but it can be ignored for most purposes. The ability of a

diode to withstand reverse-bias voltages is limited, like it is for any insulating substance or device. If the applied reverse-bias voltage becomes too great, the diode will experience a condition known as *breakdown*, which is usually destructive. A diode's maximum reverse-bias voltage rating is known as the *Peak Inverse Voltage*, or *PIV*, and may be obtained from the manufacturer. Like forward voltage, the PIV rating of a diode varies with temperature, except that PIV *increases* with increased temperature and *decreases* as the diode becomes cooler -- exactly opposite that of forward voltage.

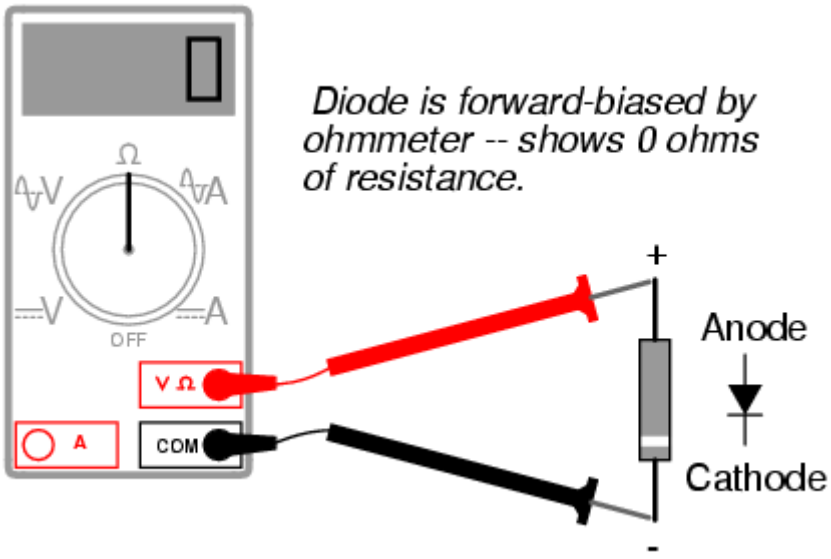


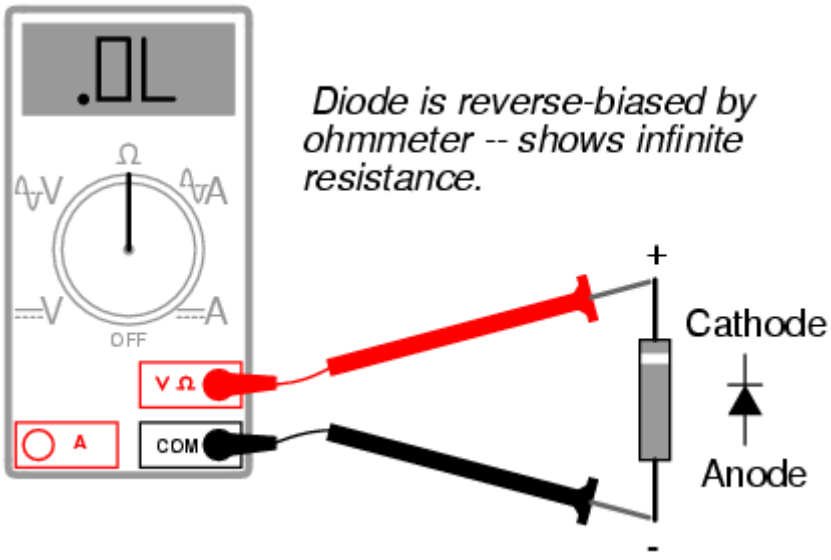
Typically, the PIV rating of a generic "rectifier" diode is at least 50 volts at room temperature. Diodes with PIV ratings in the many thousands of volts are available for modest prices.

- **REVIEW:**
- A *diode* is an electrical component acting as a one-way valve for current.
- When voltage is applied across a diode in such a way that the diode allows current, the diode is said to be *forward-biased*.
- When voltage is applied across a diode in such a way that the diode prohibits current, the diode is said to be *reverse-biased*.
- The voltage dropped across a conducting, forward-biased diode is called the *forward voltage*. Forward voltage for a diode varies only slightly for changes in forward current and temperature, and is fixed principally by the chemical composition of the P-N junction.
- Silicon diodes have a forward voltage of approximately 0.7 volts.
- Germanium diodes have a forward voltage of approximately 0.3 volts.
- The maximum reverse-bias voltage that a diode can withstand without "breaking down" is called the *Peak Inverse Voltage*, or *PIV* rating.

Meter check of a diode

Being able to determine the polarity (cathode versus anode) and basic functionality of a diode is a very important skill for the electronics hobbyist or technician to have. Since we know that a diode is essentially nothing more than a one-way valve for electricity, it makes sense we should be able to verify its one-way nature using a DC (battery-powered) ohmmeter. Connected one way across the diode, the meter should show a very low resistance. Connected the other way across the diode, it should show a very high resistance ("OL" on some digital meter models):



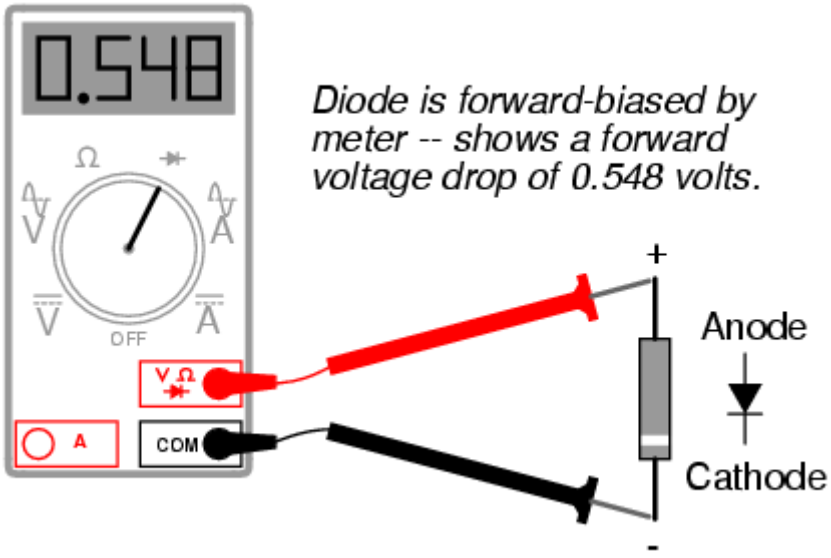


Of course, in order to determine which end of the diode is the cathode and which is the anode, you must know with certainty which test lead of the meter is positive (+) and which is negative (-) when set to the "resistance" or "Ω" function. With most digital multimeters I've seen, the red lead becomes positive and the black lead negative when set to measure resistance, in accordance with standard electronics color-code convention. However, this is

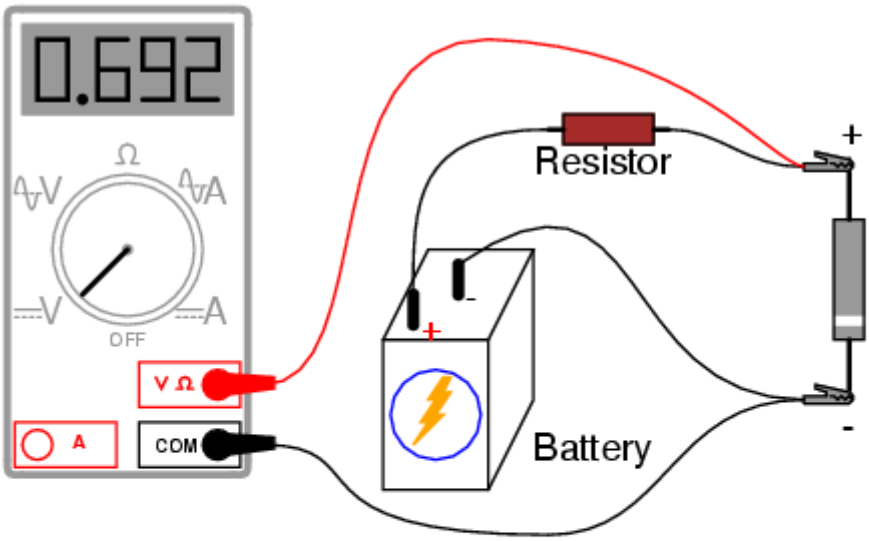
not guaranteed for all meters. Many analog multimeters, for example, actually make their black leads positive (+) and their red leads negative (-) when switched to the "resistance" function, because it is easier to manufacture it that way!

One problem with using an ohmmeter to check a diode is that the readings obtained only have qualitative value, not quantitative. In other words, an ohmmeter only tells you which way the diode conducts; the low-value resistance indication obtained while conducting is useless. If an ohmmeter shows a value of "1.73 ohms" while forward-biasing a diode, that figure of 1.73 Ω doesn't represent any real-world quantity useful to us as technicians or circuit designers. It neither represents the forward voltage drop nor any "bulk" resistance in the semiconductor material of the diode itself, but rather is a figure dependent upon both quantities and will vary substantially with the particular ohmmeter used to take the reading.

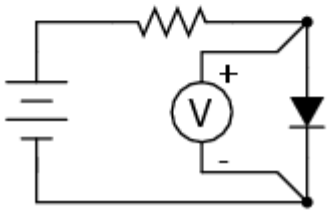
For this reason, some digital multimeter manufacturers equip their meters with a special "diode check" function which displays the actual forward voltage drop of the diode in volts, rather than a "resistance" figure in ohms. These meters work by forcing a small current through the diode and measuring the voltage dropped between the two test leads:



The forward voltage reading obtained with such a meter will typically be less than the "normal" drop of 0.7 volts for silicon and 0.3 volts for germanium, because the current provided by the meter is of trivial proportions. If a multimeter with diode-check function isn't available, or you would like to measure a diode's forward voltage drop at some non-trivial current, the following circuit may be constructed using nothing but a battery, resistor, and a normal voltmeter:



Schematic diagram



Resistor sized to obtain diode current of desired magnitude.

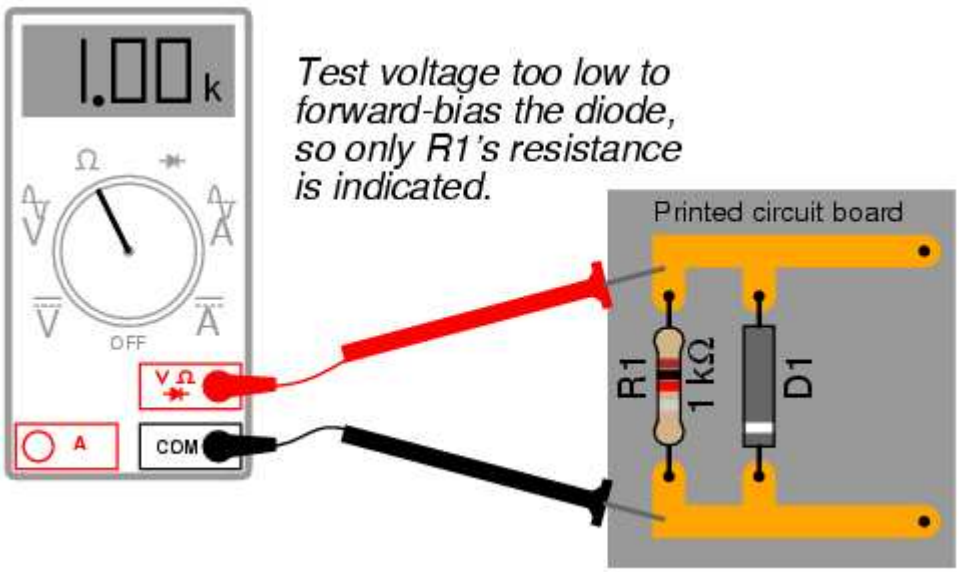
Connecting the diode backwards to this testing circuit will simply result in the voltmeter indicating the full voltage of the battery.

If this circuit were designed so as to provide a constant or nearly constant current through the diode despite changes in forward voltage drop, it could be used as the basis of a temperature-measurement instrument, the voltage measured across the diode being inversely proportional to diode junction temperature. Of course, diode current should be kept to a minimum to avoid self-heating (the diode dissipating substantial amounts of heat energy), which would interfere with temperature measurement.

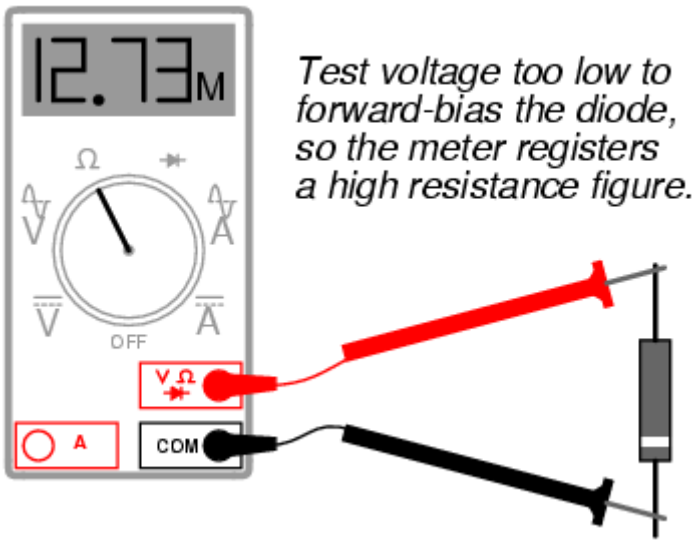
Beware that some digital multimeters equipped with a "diode check" function may output a very low test voltage (less than 0.3 volts) when set to the regular "resistance" (Ω) function: too low to fully collapse the depletion region of a PN junction. The philosophy here is that the "diode check" function is to be used for testing semiconductor devices, and the "resistance" function for anything else. By using a very low test voltage to measure resistance, it is easier for a technician to measure the resistance of non-semiconductor components connected to semiconductor components, since the semiconductor component junctions will not become forward-biased with such low voltages.

Consider the example of a resistor and diode connected in parallel, soldered in place on a printed circuit board (PCB). Normally, one would have to unsolder the resistor from the circuit (disconnect it from all other components) before being able to measure its resistance, otherwise any parallel-connected components would affect the reading obtained. However, using a multimeter that outputs a very low test voltage to the probes in the "resistance" function mode, the diode's PN junction will not have enough voltage impressed across it to

become forward-biased, and as such will pass negligible current. Consequently, the meter "sees" the diode as an open (no continuity), and only registers the resistor's resistance:



If such an ohmmeter were used to test a diode, it would indicate a very high resistance (many mega-ohms) even if connected to the diode in the "correct" (forward-biased) direction:



Reverse voltage strength of a diode is not as easily tested, because exceeding a normal diode's PIV usually results in destruction of the diode. There are special types of diodes, though, which are designed to "break down" in reverse-bias mode without damage (called *Zener diodes*), and they are best tested with the same type of voltage source / resistor / voltmeter circuit, provided that the voltage source is of high enough value to force the diode into its breakdown region. More on this subject in a later section of this chapter.

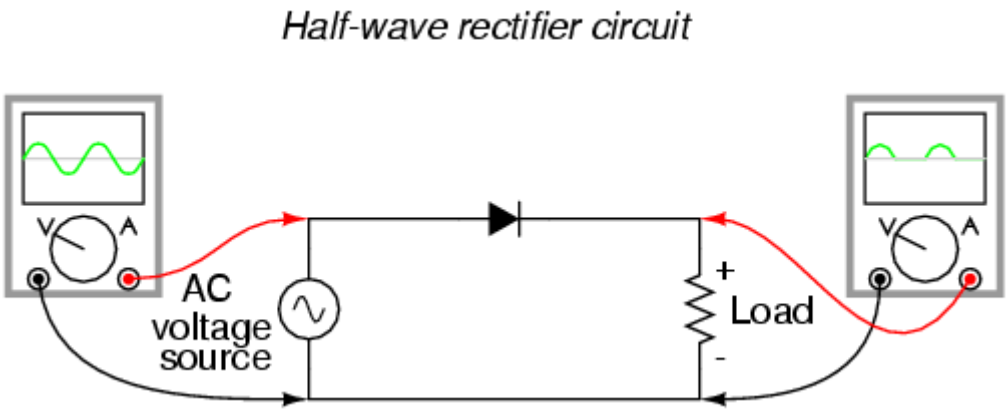
- **REVIEW:**
- An ohmmeter may be used to qualitatively check diode function. There should be low resistance measured one way and very high resistance measured the other way. When using an ohmmeter for this purpose, be sure you know which test lead is positive and

which is negative! The actual polarity may not follow the colors of the leads as you might expect, depending on the particular design of meter.

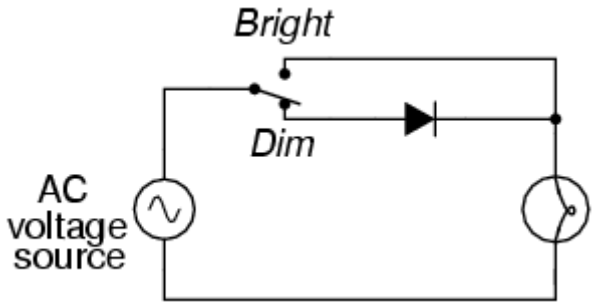
- Some multimeters provide a "diode check" function that displays the actual forward voltage of the diode when it's conducting current. Such meters typically indicate a slightly lower forward voltage than what is "nominal" for a diode, due to the very small amount of current used during the check.

Rectifier circuits

Now we come to the most popular application of the diode: *rectification*. Simply defined, rectification is the conversion of alternating current (AC) to direct current (DC). This almost always involves the use of some device that only allows one-way flow of electrons. As we have seen, this is exactly what a semiconductor diode does. The simplest type of rectifier circuit is the *half-wave* rectifier, so called because it only allows one half of an AC waveform to pass through to the load:



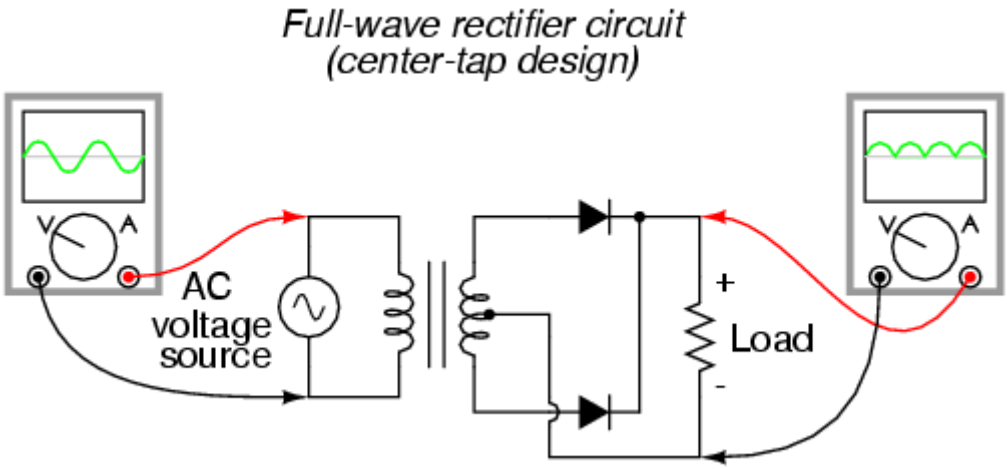
For most power applications, half-wave rectification is insufficient for the task. The harmonic content of the rectifier's output waveform is very large and consequently difficult to filter. Furthermore, AC power source only works to supply power to the load once every half-cycle, meaning that much of its capacity is unused. Half-wave rectification is, however, a very simple way to reduce power to a resistive load. Some two-position lamp dimmer switches apply full AC power to the lamp filament for "full" brightness and then half-wave rectify it for a lesser light output:



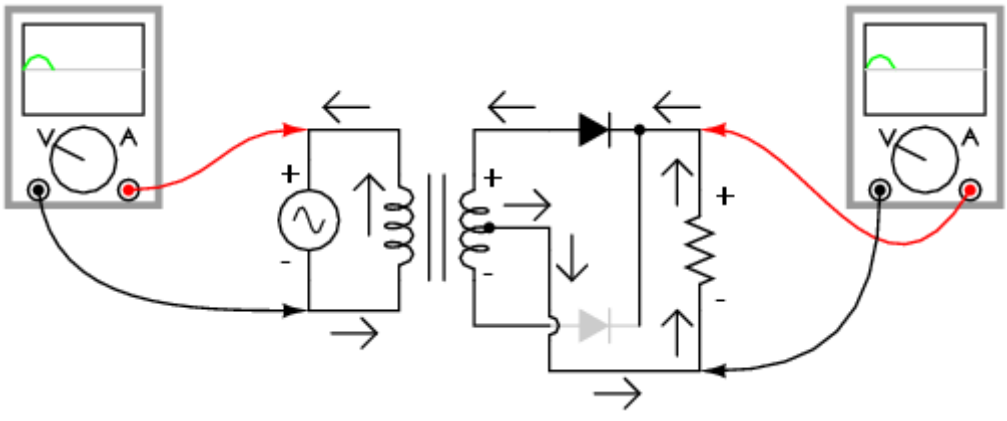
In the "Dim" switch position, the incandescent lamp receives approximately one-half the power it would normally receive operating on full-wave AC. Because the half-wave rectified power pulses far more rapidly than the filament has time to heat up and cool down, the lamp does not blink. Instead, its filament merely operates at a lesser temperature than normal,

providing less light output. This principle of "pulsing" power rapidly to a slow-responding load device in order to control the electrical power sent to it is very common in the world of industrial electronics. Since the controlling device (the diode, in this case) is either fully conducting or fully nonconducting at any given time, it dissipates little heat energy while controlling load power, making this method of power control very energy-efficient. This circuit is perhaps the crudest possible method of pulsing power to a load, but it suffices as a proof-of-concept application.

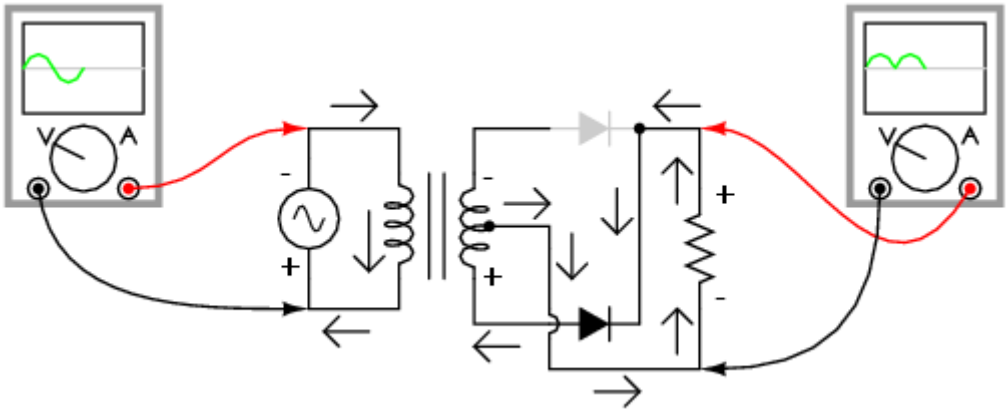
If we need to rectify AC power so as to obtain the full use of *both* half-cycles of the sine wave, a different rectifier circuit configuration must be used. Such a circuit is called a *full-wave* rectifier. One type of full-wave rectifier, called the *center-tap* design, uses a transformer with a center-tapped secondary winding and two diodes, like this:



This circuit's operation is easily understood one half-cycle at a time. Consider the first half-cycle, when the source voltage polarity is positive (+) on top and negative (-) on bottom. At this time, only the top diode is conducting; the bottom diode is blocking current, and the load "sees" the first half of the sine wave, positive on top and negative on bottom. Only the top half of the transformer's secondary winding carries current during this half-cycle:

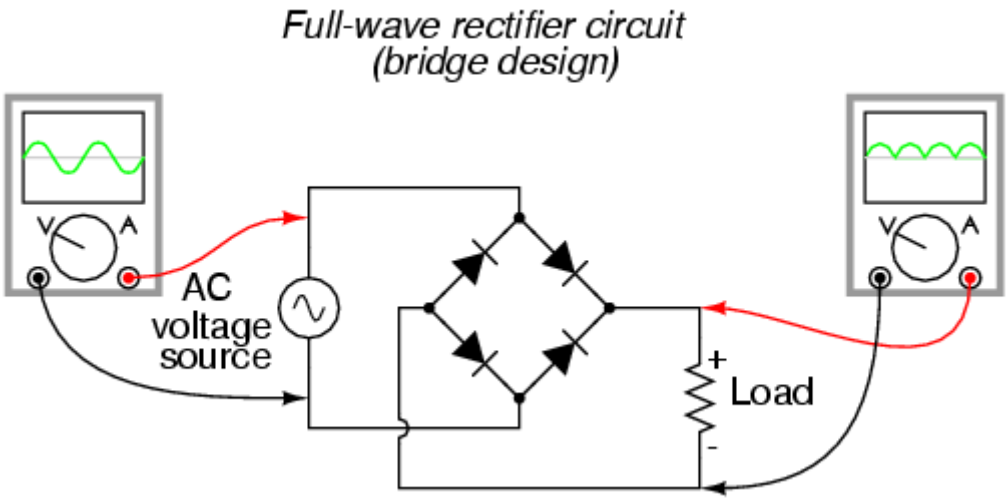


During the next half-cycle, the AC polarity reverses. Now, the other diode and the other half of the transformer's secondary winding carry current while the portions of the circuit formerly carrying current during the last half-cycle sit idle. The load still "sees" half of a sine wave, of the same polarity as before: positive on top and negative on bottom:

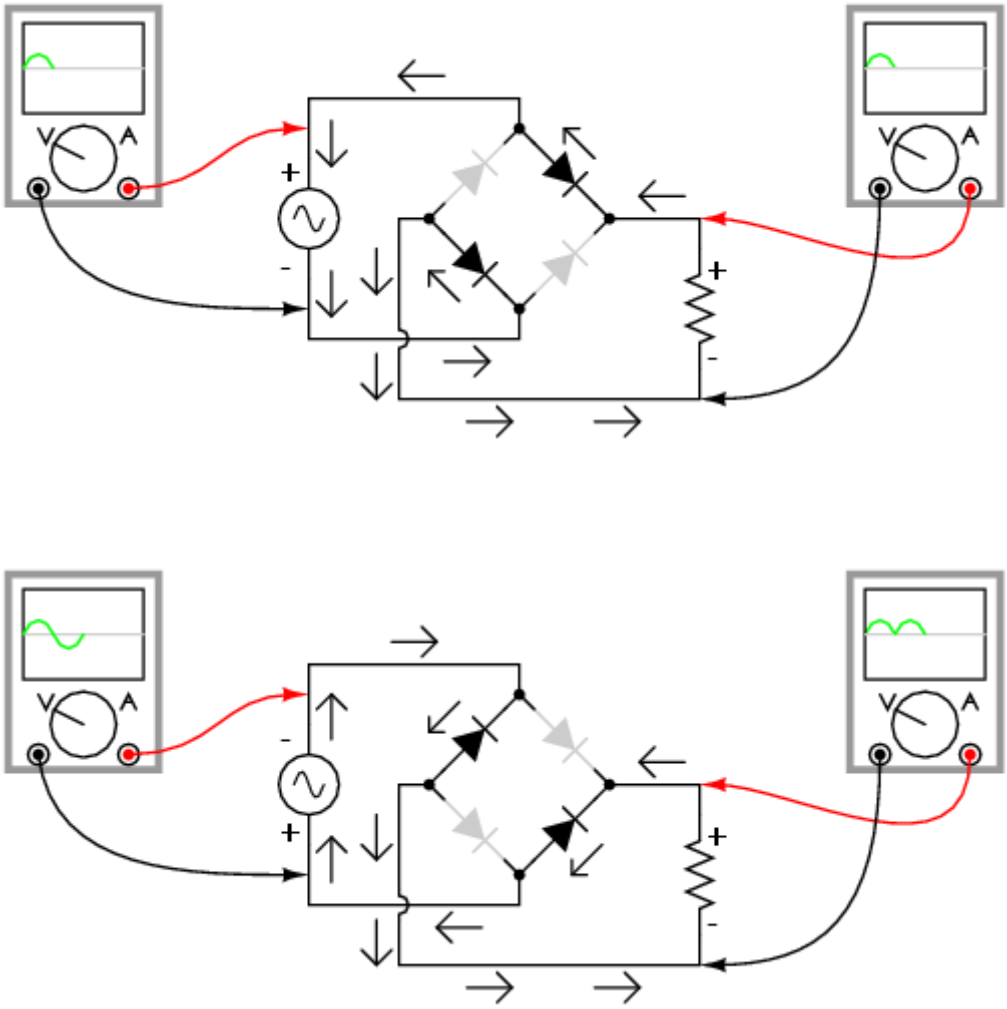


One disadvantage of this full-wave rectifier design is the necessity of a transformer with a center-tapped secondary winding. If the circuit in question is one of high power, the size and expense of a suitable transformer is significant. Consequently, the center-tap rectifier design is seen only in low-power applications.

Another, more popular full-wave rectifier design exists, and it is built around a four-diode bridge configuration. For obvious reasons, this design is called a *full-wave bridge*:

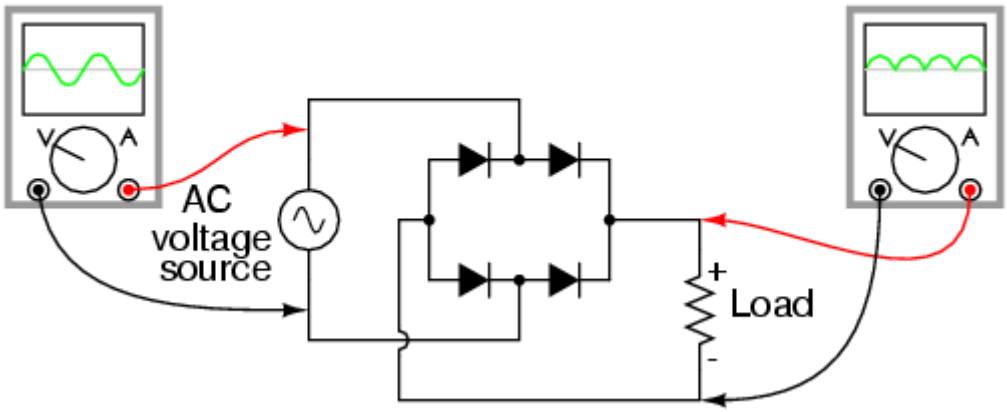


Current directions in the full-wave bridge rectifier circuit are as follows for each half-cycle of the AC waveform:



Remembering the proper layout of diodes in a full-wave bridge rectifier circuit can often be frustrating to the new student of electronics. I've found that an alternative representation of this circuit is easier both to remember and to comprehend. It's the exact same circuit, except all diodes are drawn in a horizontal attitude, all "pointing" the same direction:

Full-wave bridge rectifier circuit
(alternative layout)



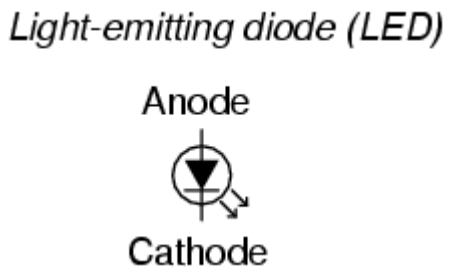
Light-emitting diodes

Diodes, like all semiconductor devices, are governed by the principles described in quantum physics. One of these principles is the emission of specific-frequency radiant energy whenever electrons fall from a higher energy level to a lower energy level. This is the same principle at work in a neon lamp, the characteristic pink-orange glow of ionized neon due to the specific energy transitions of its electrons in the midst of an electric current. The unique color of a neon lamp's glow is due to the fact that it's *neon* gas inside the tube, and not due to the particular amount of current through the tube or voltage between the two electrodes. Neon gas glows pinkish-orange over a wide range of ionizing voltages and currents. Each chemical element has its own "signature" emission of radiant energy when its electrons "jump" between different, quantized energy levels. Hydrogen gas, for example, glows red when ionized; mercury vapor glows blue. This is what makes spectrographic identification of elements possible.

Electrons flowing through a PN junction experience similar transitions in energy level, and emit radiant energy as they do so. The frequency of this radiant energy is determined by the crystal structure of the semiconductor material, and the elements comprising it. Some semiconductor junctions, composed of special chemical combinations, emit radiant energy within the spectrum of visible light as the electrons transition in energy levels. Simply put, these junctions *glow* when forward biased. A diode intentionally designed to glow like a lamp is called a *light-emitting diode*, or *LED*.

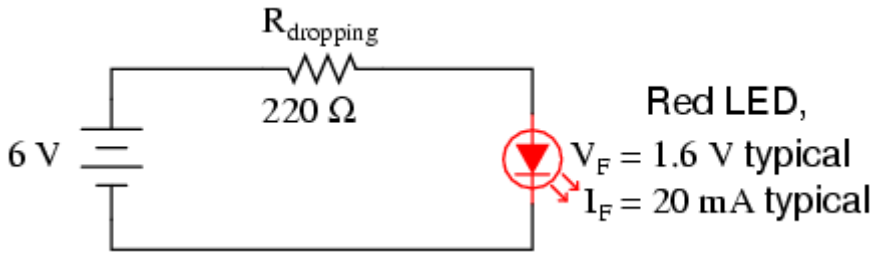
Diodes made from a combination of the elements gallium, arsenic, and phosphorus (called *gallium-arsenide-phosphide*) glow bright red, and are some of the most common LEDs manufactured. By altering the chemical constituency of the PN junction, different colors may be obtained. Some of the currently available colors other than red are green, blue, and infra-red (invisible light at a frequency lower than red). Other colors may be obtained by combining two or more primary-color (red, green, and blue) LEDs together in the same package, sharing the same optical lens. For instance, a yellow LED may be made by merging a red LED with a green LED.

The schematic symbol for an LED is a regular diode shape inside of a circle, with two small arrows pointing away (indicating emitted light):

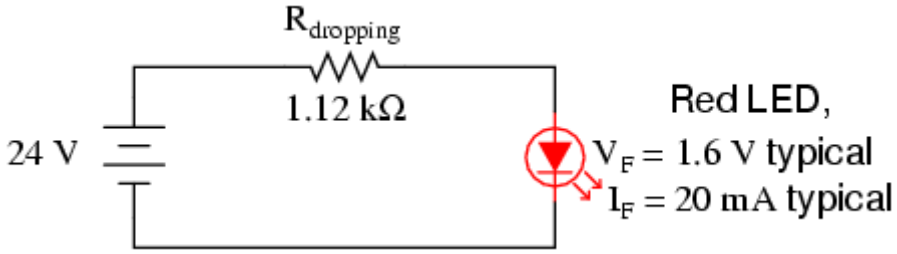


This notation of having two small arrows pointing away from the device is common to the schematic symbols of all light-emitting semiconductor devices. Conversely, if a device is light-*activated* (meaning that incoming light stimulates it), then the symbol will have two small arrows pointing *toward* it. It is interesting to note, though, that LEDs are capable of acting as light-sensing devices: they will generate a small voltage when exposed to light, much like a solar cell on a small scale. This property can be gainfully applied in a variety of light-sensing circuits.

Because LEDs are made of different chemical substances than normal rectifying diodes, their forward voltage drops will be different. Typically, LEDs have much larger forward voltage drops than rectifying diodes, anywhere from about 1.6 volts to over 3 volts, depending on the color. Typical operating current for a standard-sized LED is around 20 mA. When operating an LED from a DC voltage source greater than the LED's forward voltage, a series-connected "dropping" resistor must be included to prevent full source voltage from damaging the LED. Consider this example circuit:



With the LED dropping 1.6 volts, there will be 4.4 volts dropped across the resistor. Sizing the resistor for an LED current of 20 mA is as simple as taking its voltage drop (4.4 volts) and dividing by circuit current (20 mA), in accordance with Ohm's Law (R=E/I). This gives us a figure of 220 Ω. Calculating power dissipation for this resistor, we take its voltage drop and multiply by its current (P=IE), and end up with 88 mW, well within the rating of a 1/8 watt resistor. Higher battery voltages will require larger-value dropping resistors, and possibly higher-power rating resistors as well. Consider this example for a supply voltage of 24 volts:

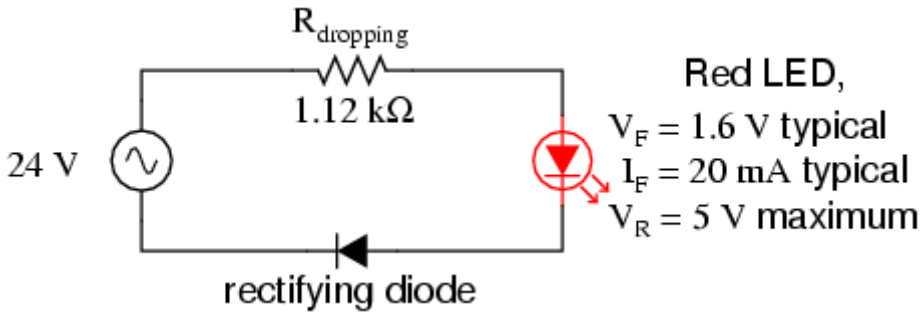


Here, the dropping resistor must be increased to a size of 1.12 kΩ in order to drop 22.4 volts at 20 mA so that the LED still receives only 1.6 volts. This also makes for a higher resistor power dissipation: 448 mW, nearly one-half a watt of power! Obviously, a resistor rated for 1/8 watt power dissipation or even 1/4 watt dissipation will overheat if used here.

Dropping resistor values need not be precise for LED circuits. Suppose we were to use a 1 kΩ resistor instead of a 1.12 kΩ resistor in the circuit shown above. The result would be a slightly greater circuit current and LED voltage drop, resulting in a brighter light from the LED and slightly reduced service life. A dropping resistor with too much resistance (say, 1.5 kΩ instead of 1.12 kΩ) will result in less circuit current, less LED voltage, and a dimmer light. LEDs are quite tolerant of variation in applied power, so you need not strive for perfection in sizing the dropping resistor.

Also because of their unique chemical makeup, LEDs have much, much lower peak-inverse voltage (PIV) ratings than ordinary rectifying diodes. A typical LED might only be rated at 5 volts in reverse-bias mode. Therefore, when using alternating current to power an LED, you

should connect a protective rectifying diode in series with the LED to prevent reverse breakdown every other half-cycle:

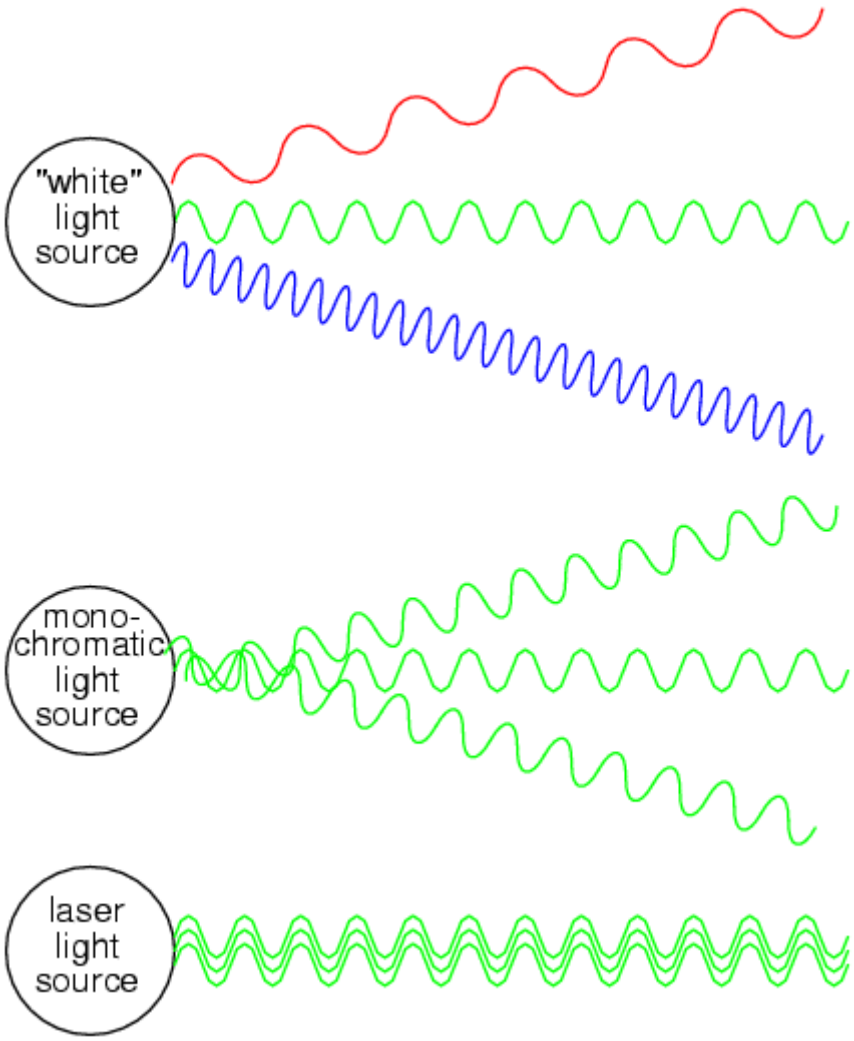


As lamps, LEDs are superior to incandescent bulbs in many ways. First and foremost is efficiency: LEDs output far more light power per watt than an incandescent lamp. This is a significant advantage if the circuit in question is battery-powered, efficiency translating to longer battery life. Second is the fact that LEDs are far more reliable, having a much greater service life than an incandescent lamp. This advantage is primarily due to the fact that LEDs are "cold" devices: they operate at much cooler temperatures than an incandescent lamp with a white-hot metal filament, susceptible to breakage from mechanical and thermal shock. Third is the high speed at which LEDs may be turned on and off. This advantage is also due to the "cold" operation of LEDs: they don't have to overcome thermal inertia in transitioning from off to on or vice versa. For this reason, LEDs are used to transmit digital (on/off) information as pulses of light, conducted in empty space or through fiber-optic cable, at very high rates of speed (millions of pulses per second).

One major disadvantage of using LEDs as sources of illumination is their monochromatic (single-color) emission. No one wants to read a book under the light of a red, green, or blue LED. However, if used in combination, LED colors may be mixed for a more broad-spectrum glow.

Laser diodes

The *laser diode* is a further development upon the regular light-emitting diode, or LED. The term "laser" itself is actually an acronym, despite the fact it's often written in lower-case letters. "Laser" stands for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation, and refers to another strange quantum process whereby characteristic light emitted by electrons transitioning from high-level to low-level energy states in a material stimulate other electrons in a substance to make similar "jumps," the result being a synchronized output of light from the material. This synchronization extends to the actual *phase* of the emitted light, so that all light waves emitted from a "lasing" material are not just the same frequency (color), but also the same phase as each other, so that they reinforce one another and are able to travel in a very tightly-confined, nondispersing beam. This is why laser light stays so remarkably focused over long distances: each and every light wave coming from the laser is in step with each other:



Incandescent lamps produce "white" (mixed-frequency, or mixed-color) light. Regular LEDs produce monochromatic light: same frequency (color), but different phases, resulting in similar beam dispersion. Laser LEDs produce *coherent light*: light that is both monochromatic (single-color) and monophasic (single-phase), resulting in precise beam confinement.

Laser light finds wide application in the modern world: everything from surveying, where a straight and nondispersing light beam is very useful for precise sighting of measurement markers, to the reading and writing of optical disks, where only the narrowness of a focused laser beam is able to resolve the microscopic "pits" in the disk's surface comprising the binary 1's and 0's of digital information.

Some laser diodes require special high-power "pulsing" circuits to deliver large quantities of voltage and current in short bursts. Other laser diodes may be operated continuously at lower power. In the latter case, laser action occurs only within a certain range of diode current, necessitating some form of current-regulator circuit. As laser diodes age, their power requirements may change (more current required for less output power), but it should be remembered that low-power laser diodes, like LEDs, are fairly long-lived devices, with typical service lives in the tens of thousands of hours.

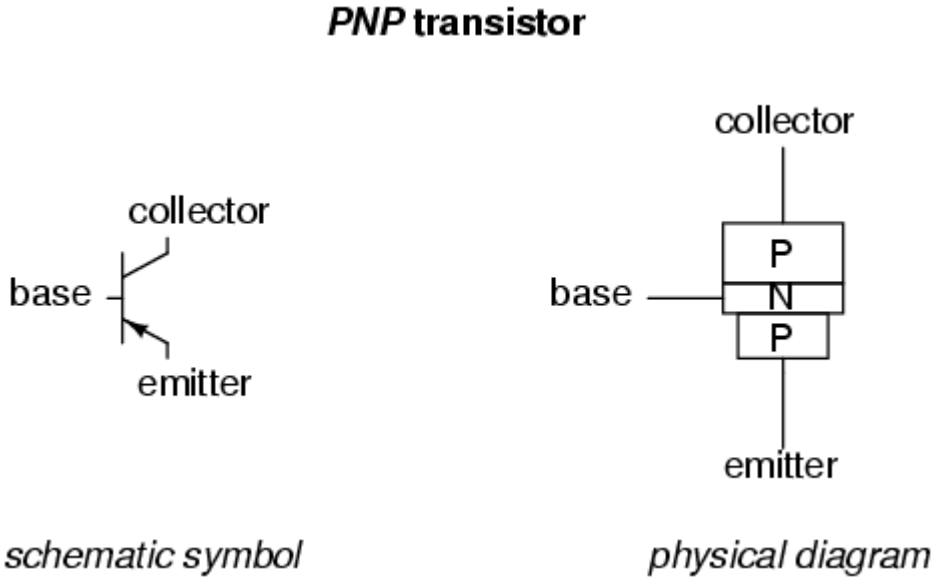
3.BIPOLAR TRANSISTORS

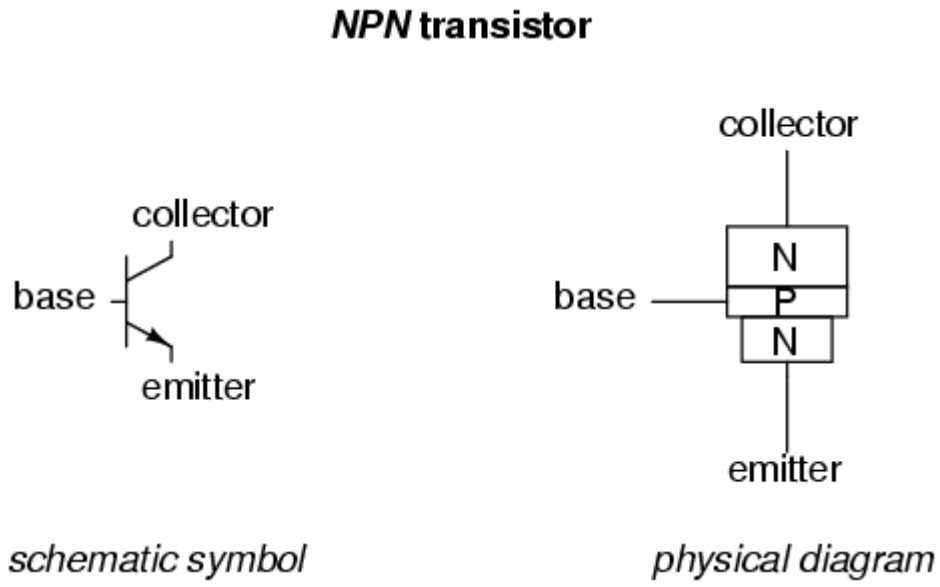
Introduction

The invention of the bipolar transistor in 1948 ushered in a revolution in electronics. Technical feats previously requiring relatively large, mechanically fragile, power-hungry vacuum tubes were suddenly achievable with tiny, mechanically rugged, power-thrifty specks of crystalline silicon. This revolution made possible the design and manufacture of lightweight, inexpensive electronic devices that we now take for granted. Understanding how transistors function is of paramount importance to anyone interested in understanding modern electronics.

My intent here is to focus as exclusively as possible on the practical function and application of bipolar transistors, rather than to explore the quantum world of semiconductor theory. Discussions of holes and electrons are better left to another chapter in my opinion. Here I want to explore how to *use* these components, not analyze their intimate internal details. I don't mean to downplay the importance of understanding semiconductor physics, but sometimes an intense focus on solid-state physics detracts from understanding these devices' functions on a component level. In taking this approach, however, I assume that the reader possesses a certain minimum knowledge of semiconductors: the difference between "P" and "N" doped semiconductors, the functional characteristics of a PN (diode) junction, and the meanings of the terms "reverse biased" and "forward biased." If these concepts are unclear to you, it is best to refer to earlier chapters in this book before proceeding with this one.

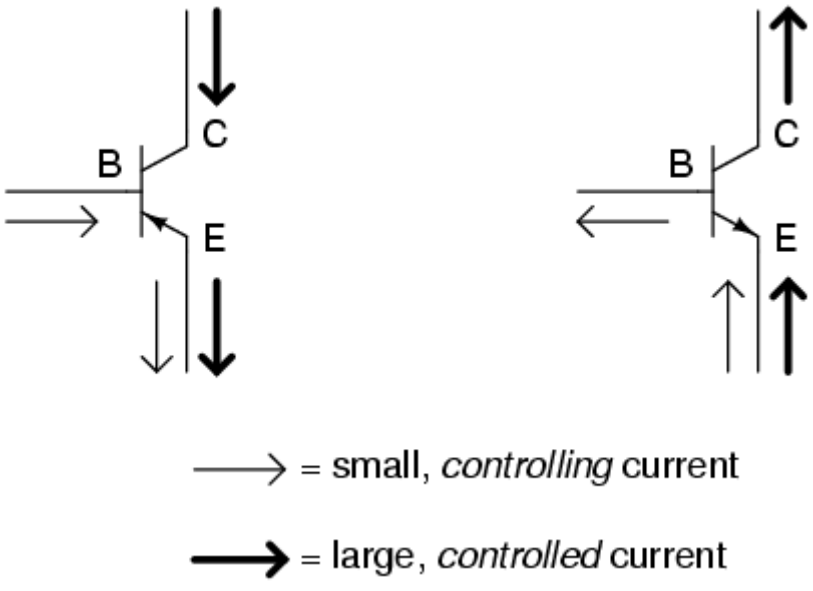
A bipolar transistor consists of a three-layer "sandwich" of doped (extrinsic) semiconductor materials, either P-N-P or N-P-N. Each layer forming the transistor has a specific name, and each layer is provided with a wire contact for connection to a circuit. Shown here are schematic symbols and physical diagrams of these two transistor types:





The only functional difference between a PNP transistor and an NPN transistor is the proper biasing (polarity) of the junctions when operating. For any given state of operation, the current directions and voltage polarities for each type of transistor are exactly opposite each other.

Bipolar transistors work as current-controlled current *regulators*. In other words, they restrict the amount of current that can go through them according to a smaller, controlling current. The main current that is *controlled* goes from collector to emitter, or from emitter to collector, depending on the type of transistor it is (PNP or NPN, respectively). The small current that *controls* the main current goes from base to emitter, or from emitter to base, once again depending on the type of transistor it is (PNP or NPN, respectively). According to the confusing standards of semiconductor symbology, the arrow always points *against* the direction of electron flow:



Bipolar transistors are called *bi*polar because the main flow of electrons through them takes place in *two* types of semiconductor material: P and N, as the main current goes from emitter

to collector (or vice versa). In other words, two types of charge carriers -- electrons and holes -- comprise this main current through the transistor.

As you can see, the *controlling* current and the *controlled* current always mesh together through the emitter wire, and their electrons always flow *against* the direction of the transistor's arrow. This is the first and foremost rule in the use of transistors: all currents must be going in the proper directions for the device to work as a current regulator. The small, controlling current is usually referred to simply as the *base current* because it is the only current that goes through the base wire of the transistor. Conversely, the large, controlled current is referred to as the *collector current* because it is the only current that goes through the collector wire. The emitter current is the sum of the base and collector currents, in compliance with Kirchhoff's Current Law.

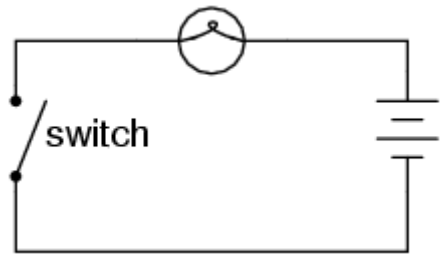
If there is no current through the base of the transistor, it shuts off like an open switch and prevents current through the collector. If there is a base current, then the transistor turns on like a closed switch and allows a proportional amount of current through the collector. Collector current is primarily limited by the base current, regardless of the amount of voltage available to push it. The next section will explore in more detail the use of bipolar transistors as switching elements.

- **REVIEW:**
- Bipolar transistors are so named because the controlled current must go through *two* types of semiconductor material: P and N. The current consists of both electron and hole flow, in different parts of the transistor.
- Bipolar transistors consist of either a P-N-P or an N-P-N semiconductor "sandwich" structure.
- The three leads of a bipolar transistor are called the *Emitter*, *Base*, and *Collector*.
- Transistors function as current regulators by allowing a small current to *control* a larger current. The amount of current allowed between collector and emitter is primarily determined by the amount of current moving between base and emitter.
- In order for a transistor to properly function as a current regulator, the controlling (base) current and the controlled (collector) currents must be going in the proper directions: meshing additively at the emitter and going *against* the emitter arrow symbol.

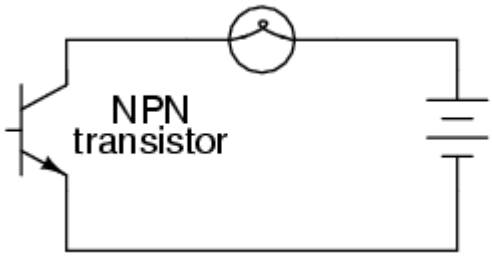
The transistor as a switch

Because a transistor's collector current is proportionally limited by its base current, it can be used as a sort of current-controlled switch. A relatively small flow of electrons sent through the base of the transistor has the ability to exert control over a much larger flow of electrons through the collector.

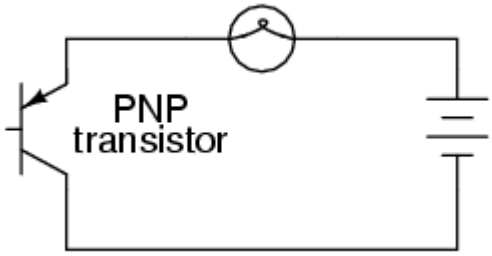
Suppose we had a lamp that we wanted to turn on and off by means of a switch. Such a circuit would be extremely simple:



For the sake of illustration, let's insert a transistor in place of the switch to show how it can control the flow of electrons through the lamp. Remember that the controlled current through a transistor must go between collector and emitter. Since it's the current through the lamp that we want to control, we must position the collector and emitter of our transistor where the two contacts of the switch are now. We must also make sure that the lamp's current will move *against* the direction of the emitter arrow symbol to ensure that the transistor's junction bias will be correct:

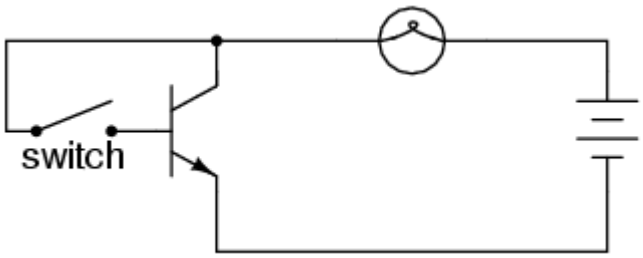


In this example I happened to choose an NPN transistor. A PNP transistor could also have been chosen for the job, and its application would look like this:

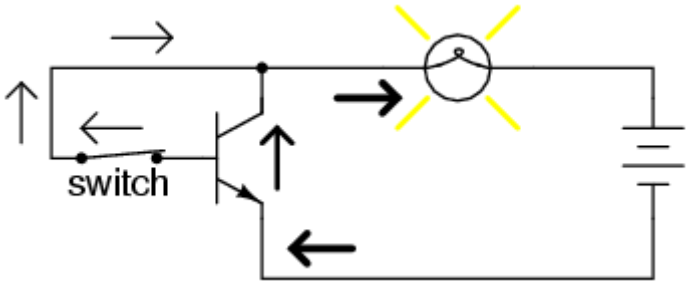


The choice between NPN and PNP is really arbitrary. All that matters is that the proper current directions are maintained for the sake of correct junction biasing (electron flow going *against* the transistor symbol's arrow).

Going back to the NPN transistor in our example circuit, we are faced with the need to add something more so that we can have base current. Without a connection to the base wire of the transistor, base current will be zero, and the transistor cannot turn on, resulting in a lamp that is always off. Remember that for an NPN transistor, base current must consist of electrons flowing from emitter to base (against the emitter arrow symbol, just like the lamp current). Perhaps the simplest thing to do would be to connect a switch between the base and collector wires of the transistor like this:

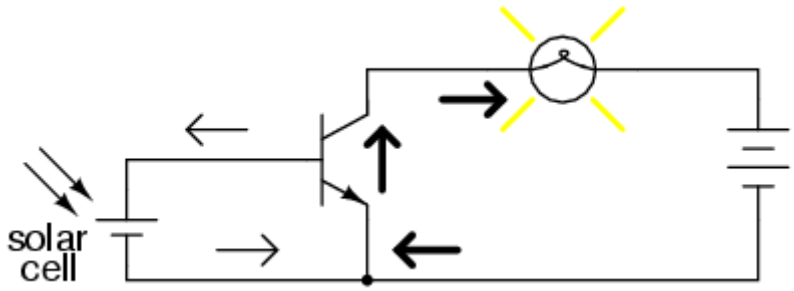


If the switch is open, the base wire of the transistor will be left "floating" (not connected to anything) and there will be no current through it. In this state, the transistor is said to be *cutoff*. If the switch is closed, however, electrons will be able to flow from the emitter through to the base of the transistor, through the switch and up to the left side of the lamp, back to the positive side of the battery. This base current will enable a much larger flow of electrons from the emitter through to the collector, thus lighting up the lamp. In this state of maximum circuit current, the transistor is said to be *saturated*.

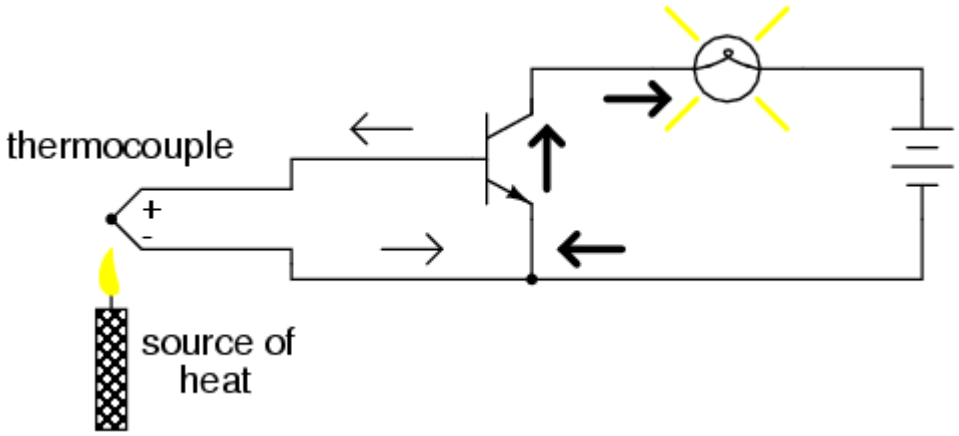


Of course, it may seem pointless to use a transistor in this capacity to control the lamp. After all, we're still using a switch in the circuit, aren't we? If we're still using a switch to control the lamp -- if only indirectly -- then what's the point of having a transistor to control the current? Why not just go back to our original circuit and use the switch directly to control the lamp current?

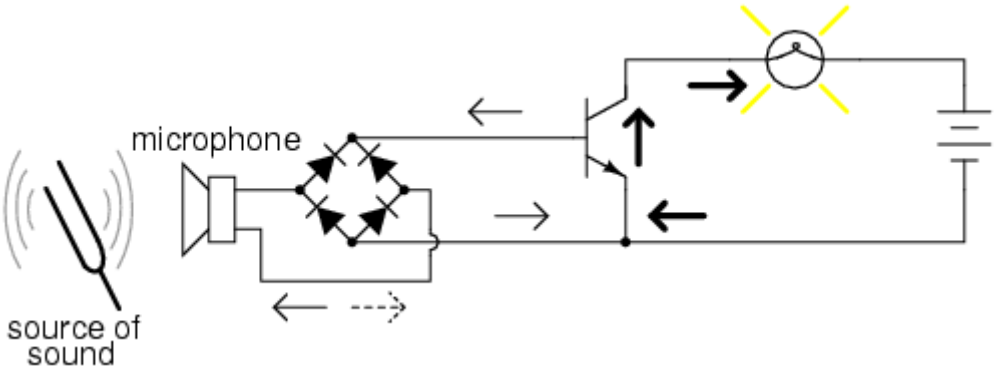
There are a couple of points to be made here, actually. First is the fact that when used in this manner, the switch contacts need only handle what little base current is necessary to turn the transistor on, while the transistor itself handles the majority of the lamp's current. This may be an important advantage if the switch has a low current rating: a small switch may be used to control a relatively high-current load. Perhaps more importantly, though, is the fact that the current-controlling behavior of the transistor enables us to use something completely different to turn the lamp on or off. Consider this example, where a solar cell is used to control the transistor, which in turn controls the lamp:



Or, we could use a thermocouple to provide the necessary base current to turn the transistor on:



Even a microphone of sufficient voltage and current output could be used to turn the transistor on, provided its output is rectified from AC to DC so that the emitter-base PN junction within the transistor will always be forward-biased:

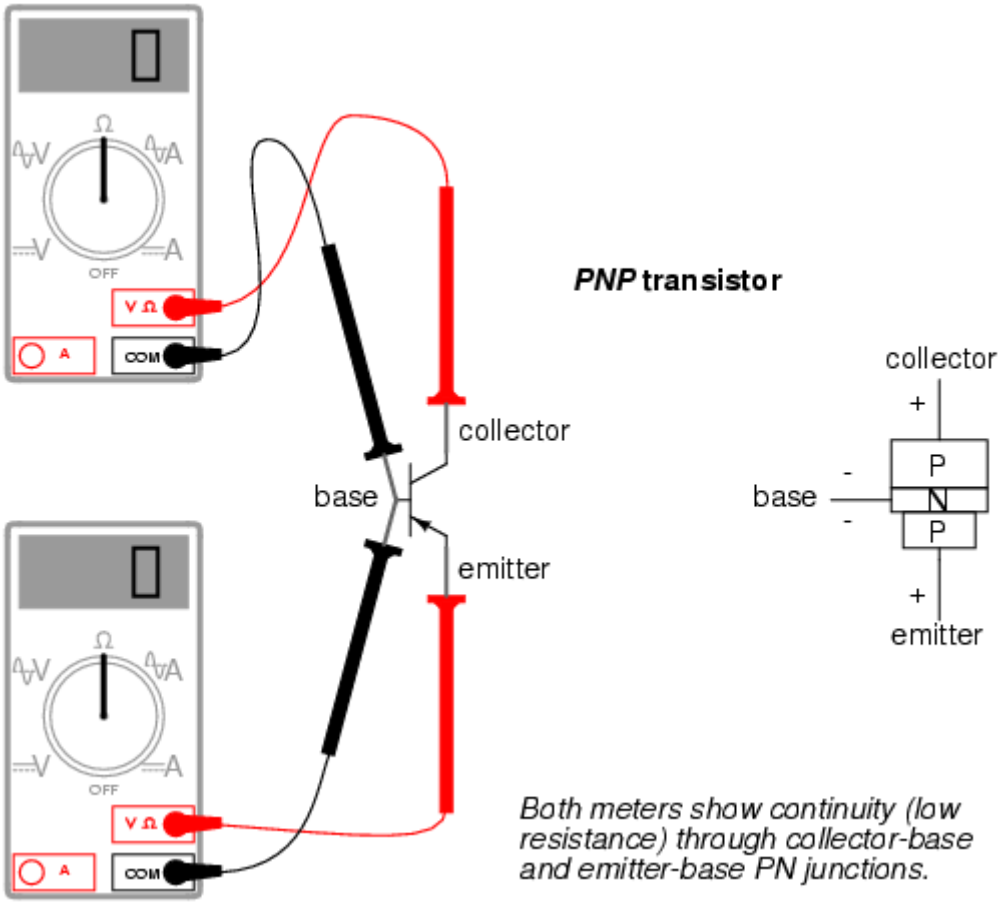


The point should be quite apparent by now: *any* sufficient source of DC current may be used to turn the transistor on, and that source of current need only be a fraction of the amount of current needed to energize the lamp. Here we see the transistor functioning not only as a switch, but as a true *amplifier*: using a relatively low-power signal to *control* a relatively large amount of power. Please note that the actual power for lighting up the lamp comes from the battery to the right of the schematic. It is not as though the small signal current from the solar cell, thermocouple, or microphone is being magically transformed into a greater amount of power. Rather, those small power sources are simply *controlling* the battery's power to light up the lamp.

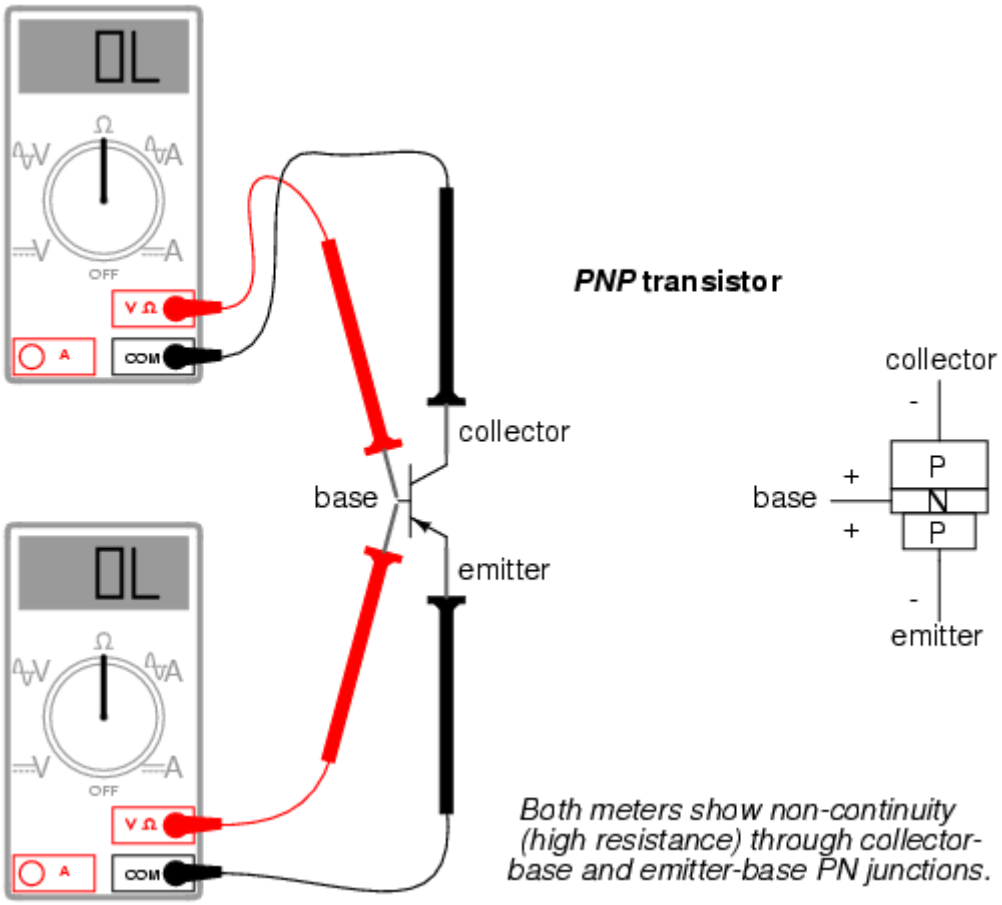
- **REVIEW:**
- Transistors may be used as switching elements to control DC power to a load. The switched (controlled) current goes between emitter and collector, while the controlling current goes between emitter and base.
- When a transistor has zero current through it, it is said to be in a state of *cutoff* (fully nonconducting).
- When a transistor has maximum current through it, it is said to be in a state of *saturation* (fully conducting).

Meter check of a transistor

Bipolar transistors are constructed of a three-layer semiconductor "sandwich," either PNP or NPN. As such, they register as two diodes connected back-to-back when tested with a multimeter's "resistance" or "diode check" functions:



Here I'm assuming the use of a multimeter with only a single continuity range (resistance) function to check the PN junctions. Some multimeters are equipped with two separate continuity check functions: resistance and "diode check," each with its own purpose. If your meter has a designated "diode check" function, use that rather than the "resistance" range, and the meter will display the actual forward voltage of the PN junction and not just whether or not it conducts current.

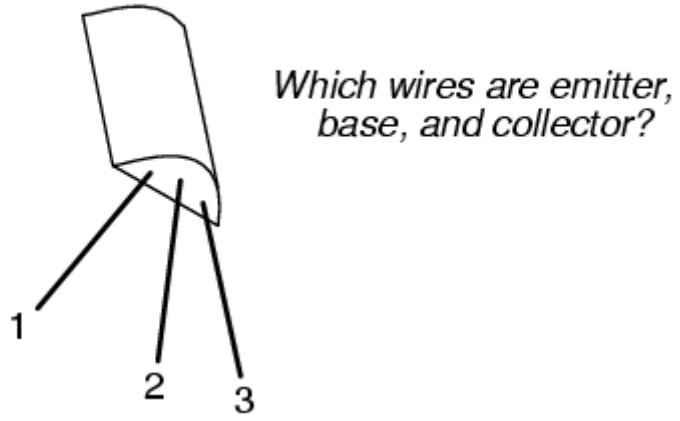


Meter readings will be exactly opposite, of course, for an NPN transistor, with both PN junctions facing the other way. If a multimeter with a "diode check" function is used in this test, it will be found that the emitter-base junction possesses a slightly greater forward voltage drop than the collector-base junction. This forward voltage difference is due to the disparity in doping concentration between the emitter and collector regions of the transistor: the emitter is a much more heavily doped piece of semiconductor material than the collector, causing its junction with the base to produce a higher forward voltage drop.

Knowing this, it becomes possible to determine which wire is which on an unmarked transistor. This is important because transistor packaging, unfortunately, is not standardized. All bipolar transistors have three wires, of course, but the positions of the three wires on the actual physical package are not arranged in any universal, standardized order.

Suppose a technician finds a bipolar transistor and proceeds to measure continuity with a multimeter set in the "diode check" mode. Measuring between pairs of wires and recording the values displayed by the meter, the technician obtains the following data:

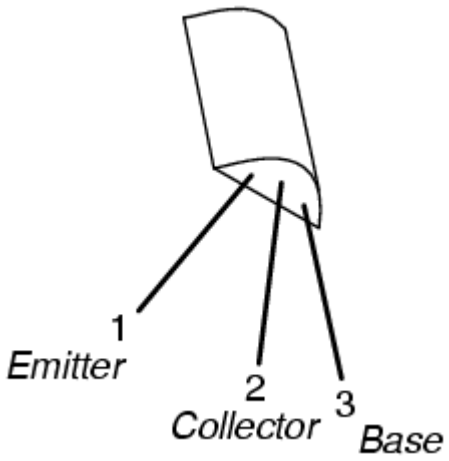
Unknown bipolar transistor



- Meter touching wire 1 (+) and 2 (-): "OL"
- Meter touching wire 1 (-) and 2 (+): "OL"
- Meter touching wire 1 (+) and 3 (-): 0.655 volts
- Meter touching wire 1 (-) and 3 (+): "OL"
- Meter touching wire 2 (+) and 3 (-): 0.621 volts
- Meter touching wire 2 (-) and 3 (+): "OL"

The only combinations of test points giving conducting meter readings are wires 1 and 3 (red test lead on 1 and black test lead on 3), and wires 2 and 3 (red test lead on 2 and black test lead on 3). These two readings *must* indicate forward biasing of the emitter-to-base junction (0.655 volts) and the collector-to-base junction (0.621 volts).

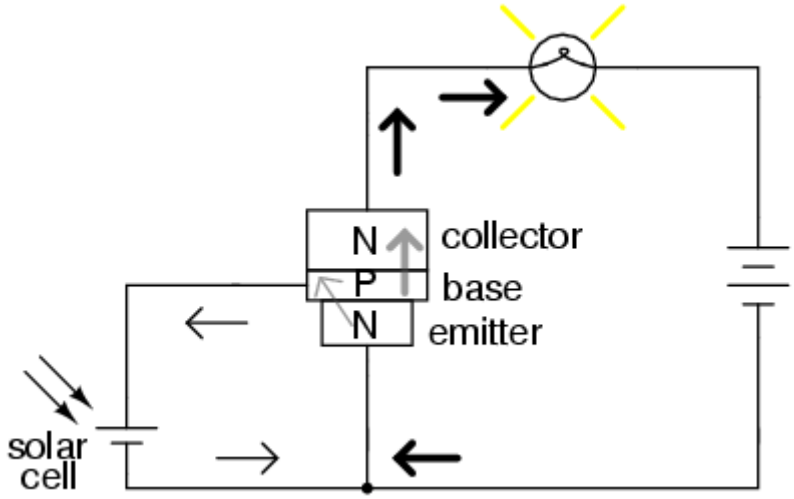
Now we look for the one wire common to both sets of conductive readings. It must be the base connection of the transistor, because the base is the only layer of the three-layer device common to both sets of PN junctions (emitter-base and collector-base). In this example, that wire is number 3, being common to both the 1-3 and the 2-3 test point combinations. In both those sets of meter readings, the *black* (-) meter test lead was touching wire 3, which tells us that the base of this transistor is made of N-type semiconductor material (black = negative). Thus, the transistor is an PNP type with base on wire 3, emitter on wire 1 and collector on wire 2:



Please note that the base wire in this example is *not* the middle lead of the transistor, as one might expect from the three-layer "sandwich" model of a bipolar transistor. This is quite often the case, and tends to confuse new students of electronics. The only way to be sure which lead is which is by a meter check, or by referencing the manufacturer's "data sheet" documentation on that particular part number of transistor.

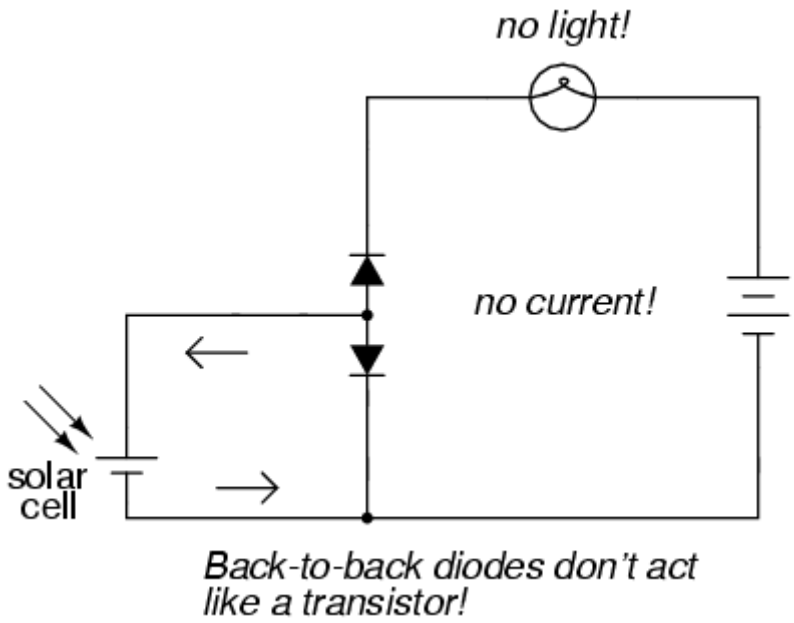
Knowing that a bipolar transistor behaves as two back-to-back diodes when tested with a conductivity meter is helpful for identifying an unknown transistor purely by meter readings. It is also helpful for a quick functional check of the transistor. If the technician were to measure continuity in any more than two or any less than two of the six test lead combinations, he or she would immediately know that the transistor was defective (or else that it *wasn't* a bipolar transistor but rather something else -- a distinct possibility if no part numbers can be referenced for sure identification!). However, the "two diode" model of the transistor fails to explain how or why it acts as an amplifying device.

To better illustrate this paradox, let's examine one of the transistor switch circuits using the physical diagram rather than the schematic symbol to represent the transistor. This way the two PN junctions will be easier to see:



A grey-colored diagonal arrow shows the direction of electron flow through the emitter-base junction. This part makes sense, since the electrons are flowing from the N-type emitter to the P-type base: the junction is obviously forward-biased. However, the base-collector junction is another matter entirely. Notice how the grey-colored thick arrow is pointing in the direction of electron flow (upwards) from base to collector. With the base made of P-type material and the collector of N-type material, this direction of electron flow is clearly backwards to the direction normally associated with a PN junction! A normal PN junction wouldn't permit this "backward" direction of flow, at least not without offering significant opposition. However, when the transistor is saturated, there is very little opposition to electrons all the way from emitter to collector, as evidenced by the lamp's illumination!

Clearly then, something is going on here that defies the simple "two-diode" explanatory model of the bipolar transistor. When I was first learning about transistor operation, I tried to construct my own transistor from two back-to-back diodes, like this:



My circuit didn't work, and I was mystified. However useful the "two diode" description of a transistor might be for testing purposes, it doesn't explain how a transistor can behave as a controlled switch.

What happens in a transistor is this: the reverse bias of the base-collector junction prevents collector current when the transistor is in cutoff mode (that is, when there is no base current). However, when the base-emitter junction is forward biased by the controlling signal, the normally-blocking action of the base-collector junction is overridden and current is permitted through the collector, despite the fact that electrons are going the "wrong way" through that PN junction. This action is dependent on the quantum physics of semiconductor junctions, and can only take place when the two junctions are properly spaced and the doping concentrations of the three layers are properly proportioned. Two diodes wired in series fail to meet these criteria, and so the top diode can never "turn on" when it is reversed biased, no matter how much current goes through the bottom diode in the base wire loop.

That doping concentrations play a crucial part in the special abilities of the transistor is further evidenced by the fact that collector and emitter are not interchangeable. If the transistor is merely viewed as two back-to-back PN junctions, or merely as a plain N-P-N or P-N-P sandwich of materials, it may seem as though either end of the transistor could serve as collector or emitter. This, however, is not true. If connected "backwards" in a circuit, a base-collector current will fail to control current between collector and emitter. Despite the fact that both the emitter and collector layers of a bipolar transistor are of the same doping *type* (either N or P), they are definitely not identical!

So, current through the emitter-base junction allows current through the reverse-biased base-collector junction. The action of base current can be thought of as "opening a gate" for current through the collector. More specifically, any given amount of emitter-to-base current *permits a limited amount* of base-to-collector current. For every electron that passes through the emitter-base junction and on through the base wire, there is allowed a certain, restricted number of electrons to pass through the base-collector junction and no more.

In the next section, this current-limiting behavior of the transistor will be investigated in more detail.

- **REVIEW:**
- Tested with a multimeter in the "resistance" or "diode check" modes, a transistor behaves like two back-to-back PN (diode) junctions.
- The emitter-base PN junction has a slightly greater forward voltage drop than the collector-base PN junction, due to more concentrated doping of the emitter semiconductor layer.
- The reverse-biased base-collector junction normally blocks any current from going through the transistor between emitter and collector. However, that junction begins to conduct if current is drawn through the base wire. Base current can be thought of as "opening a gate" for a certain, limited amount of current through the collector.

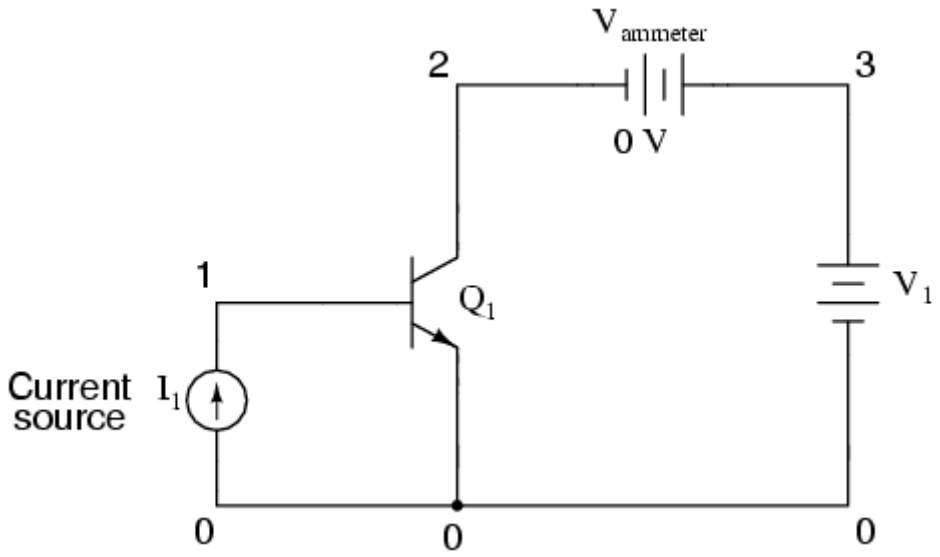
Active mode operation

When a transistor is in the fully-off state (like an open switch), it is said to be *cutoff*. Conversely, when it is fully conductive between emitter and collector (passing as much current through the collector as the collector power supply and load will allow), it is said to be *saturated*. These are the two modes of operation explored thus far in using the transistor as a switch.

However, bipolar transistors don't have to be restricted to these two extreme modes of operation. As we learned in the previous section, base current "opens a gate" for a limited amount of current through the collector. If this limit for the controlled current is greater than zero but less than the maximum allowed by the power supply and load circuit, the transistor will "throttle" the collector current in a mode somewhere between cutoff and saturation. This mode of operation is called the *active* mode.

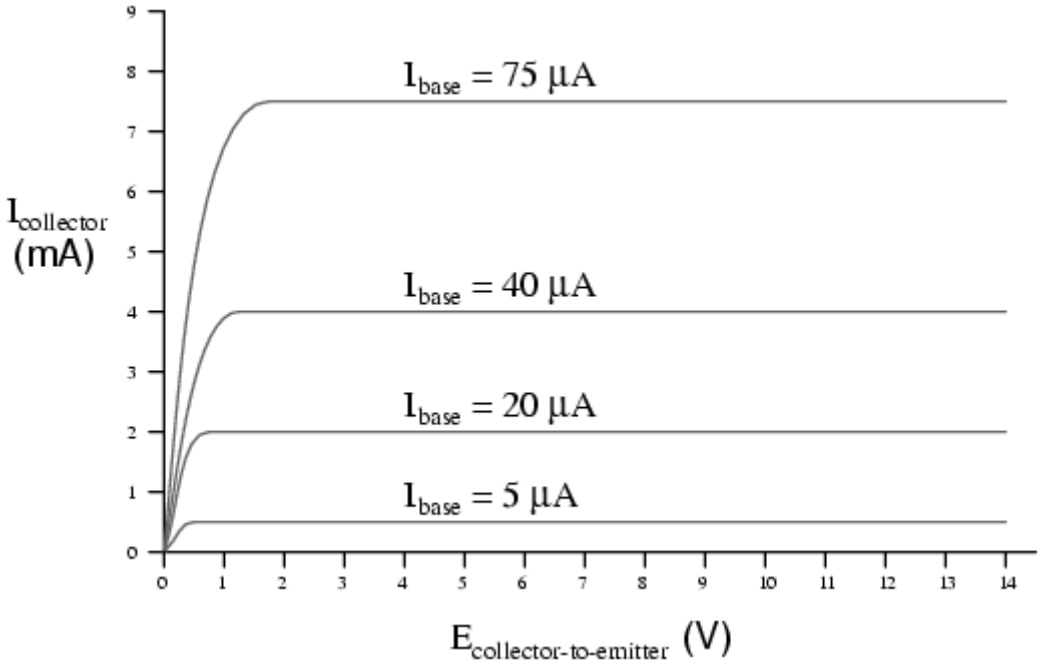
An automotive analogy for transistor operation is as follows: *cutoff* is the condition where there is no motive force generated by the mechanical parts of the car to make it move. In cutoff mode, the brake is engaged (zero base current), preventing motion (collector current). *Active mode* is when the automobile is cruising at a constant, controlled speed (constant, controlled collector current) as dictated by the driver. *Saturation* is when the automobile is driving up a steep hill that prevents it from going as fast as the driver would wish. In other words, a "saturated" automobile is one where the accelerator pedal is pushed all the way down (base current calling for more collector current than can be provided by the power supply/load circuit).

I'll set up a circuit for SPICE simulation to demonstrate what happens when a transistor is in its active mode of operation:



"Q" is the standard letter designation for a transistor in a schematic diagram, just as "R" is for resistor and "C" is for capacitor. In this circuit, we have an NPN transistor powered by a battery (V_1) and controlled by current through a *current source* (I_1). A current source is a device that outputs a specific amount of current, generating as much or as little voltage as necessary across its terminals to ensure that exact amount of current through it. Current sources are notoriously difficult to find in nature (unlike voltage sources, which by contrast attempt to maintain a constant voltage, outputting as much or as little current in the fulfillment of that task), but can be simulated with a small collection of electronic components. As we are about to see, transistors themselves tend to mimic the constant-current behavior of a current source in their ability to *regulate* current at a fixed value.

Often it is useful to superimpose several collector current/voltage graphs for different base currents on the same graph. A collection of curves like this -- one curve plotted for each distinct level of base current -- for a particular transistor is called the transistor's *characteristic curves*:



Each curve on the graph reflects the collector current of the transistor, plotted over a range of collector-to-emitter voltages, for a given amount of base current. Since a transistor tends to act as a current regulator, limiting collector current to a proportion set by the base current, it is useful to express this proportion as a standard transistor performance measure. Specifically, the ratio of collector current to base current is known as the *Beta* ratio (symbolized by the Greek letter β):

$$\beta = \frac{I_{collector}}{I_{base}}$$

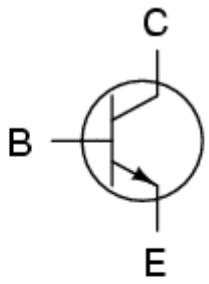
β *is also known as* h_{fe}

Sometimes the β ratio is designated as "h_{fe}," a label used in a branch of mathematical semiconductor analysis known as "hybrid parameters" which strives to achieve very precise predictions of transistor performance with detailed equations. Hybrid parameter variables are many, but they are all labeled with the general letter "h" and a specific subscript. The variable "h_{fe}" is just another (standardized) way of expressing the ratio of collector current to base current, and is interchangeable with "β." Like all ratios, β is unitless.

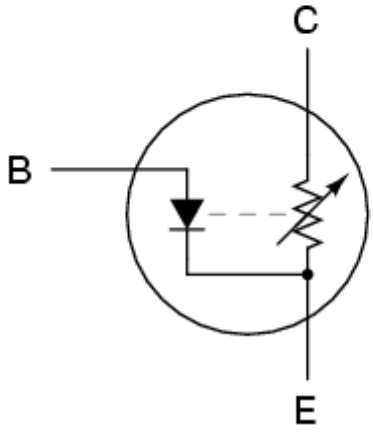
β for any transistor is determined by its design: it cannot be altered after manufacture. However, there are so many physical variables impacting β that it is rare to have two transistors of the same design exactly match. If a circuit design relies on equal β ratios between multiple transistors, "matched sets" of transistors may be purchased at extra cost. However, it is generally considered bad design practice to engineer circuits with such dependencies.

It would be nice if the β of a transistor remained stable for all operating conditions, but this is not true in real life. For an actual transistor, the β ratio may vary by a factor of over 3 within its operating current limits. For example, a transistor with advertised β of 50 may actually test with I_c/I_b ratios as low as 30 and as high as 100, depending on the amount of collector current, the transistor's temperature, and frequency of amplified signal, among other factors. For tutorial purposes it is adequate to assume a constant β for any given transistor (which is what SPICE tends to do in a simulation), but just realize that real life is not that simple!

Sometimes it is helpful for comprehension to "model" complex electronic components with a collection of simpler, better-understood components. The following is a popular model shown in many introductory electronics texts:

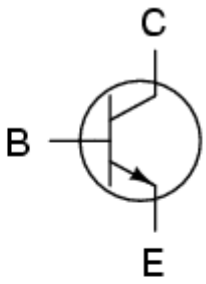


NPN diode-rheostat model

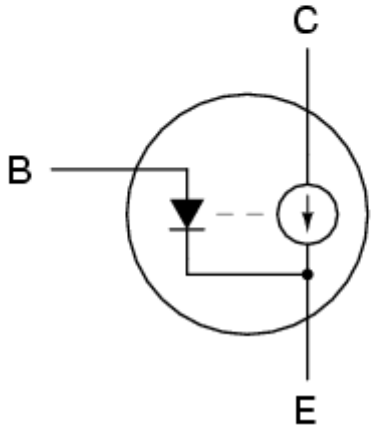


This model casts the transistor as a combination of diode and rheostat (variable resistor). Current through the base-emitter diode controls the resistance of the collector-emitter rheostat (as implied by the dashed line connecting the two components), thus controlling collector current. An NPN transistor is modeled in the figure shown, but a PNP transistor would be only slightly different (only the base-emitter diode would be reversed). This model succeeds in illustrating the basic concept of transistor amplification: how the base current signal can exert control over the collector current. However, I personally don't like this model because it tends to miscommunicate the notion of a set amount of collector-emitter resistance for a given amount of base current. If this were true, the transistor wouldn't *regulate* collector current at all like the characteristic curves show. Instead of the collector current curves flattening out after their brief rise as the collector-emitter voltage increases, the collector current would be directly proportional to collector-emitter voltage, rising steadily in a straight line on the graph.

A better transistor model, often seen in more advanced textbooks, is this:

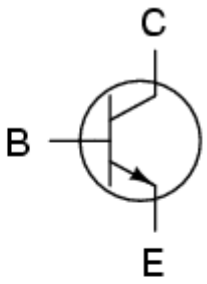


NPN diode-current source model

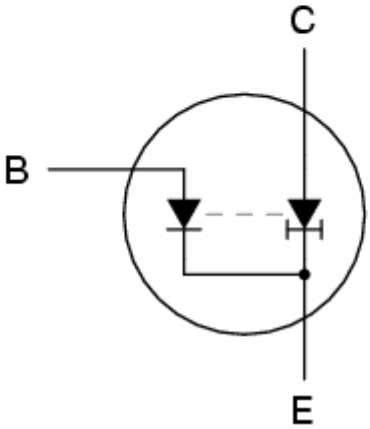


It casts the transistor as a combination of diode and current source, the output of the current source being set at a multiple (β ratio) of the base current. This model is far more accurate in depicting the true input/output characteristics of a transistor: base current establishes a certain amount of collector *current*, rather than a certain amount of collector-emitter *resistance* as the first model implies. Also, this model is favored when performing network analysis on transistor circuits, the current source being a well-understood theoretical component. Unfortunately, using a current source to model the transistor's current-controlling behavior can be misleading: in no way will the transistor ever act as a *source* of electrical energy, which the current source symbol implies is a possibility.

My own personal suggestion for a transistor model substitutes a constant-current diode for the current source:



NPN diode-regulating diode model



Since no diode ever acts as a *source* of electrical energy, this analogy escapes the false implication of the current source model as a source of power, while depicting the transistor's constant-current behavior better than the rheostat model. Another way to describe the constant-current diode's action would be to refer to it as a *current regulator*, so this transistor illustration of mine might also be described as a *diode-current regulator* model. The greatest disadvantage I see to this model is the relative obscurity of constant-current diodes. Many people may be unfamiliar with their symbology or even of their existence, unlike either rheostats or current sources, which are commonly known.

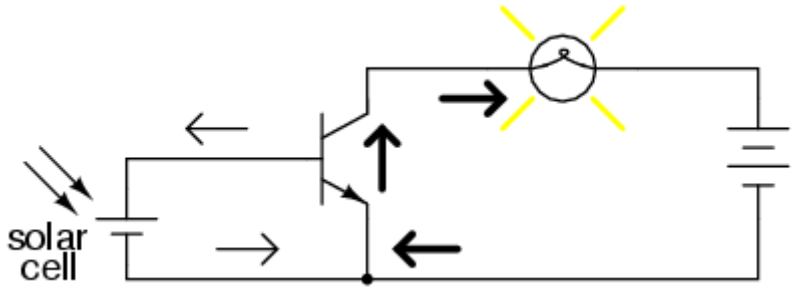
- **REVIEW:**
- A transistor is said to be in its *active* mode if it is operating somewhere between fully on (saturated) and fully off (cutoff).
- Base current tends to regulate collector current. By *regulate*, we mean that no more collector current may exist than what is allowed by the base current.
- The ratio between collector current and base current is called "Beta" (β) or "h_{fe}".
- β ratios are different for every transistor, and they tend to change for different operating conditions.

The common-emitter amplifier

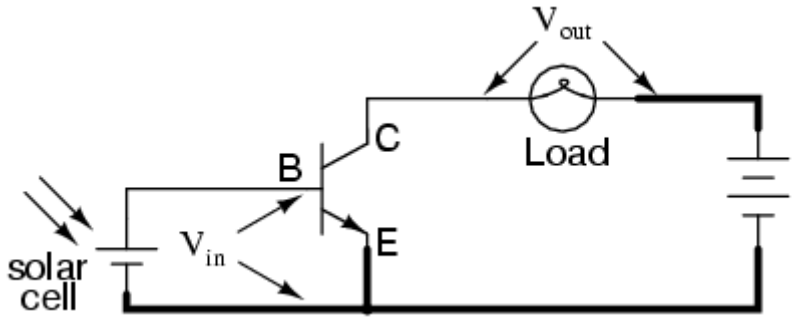
At the beginning of this chapter we saw how transistors could be used as switches, operating in either their "saturation" or "cutoff" modes. In the last section we saw how transistors behave within their "active" modes, between the far limits of saturation and cutoff. Because

transistors are able to control current in an analog (infinitely divisible) fashion, they find use as amplifiers for analog signals.

One of the simpler transistor amplifier circuits to study is the one used previously for illustrating the transistor's switching ability:

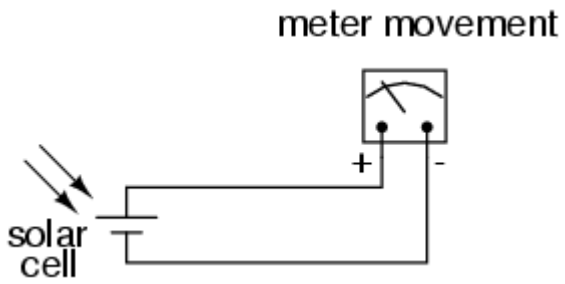


It is called the *common-emitter* configuration because (ignoring the power supply battery) both the signal source and the load share the emitter lead as a common connection point. This is not the only way in which a transistor may be used as an amplifier, as we will see in later sections of this chapter:



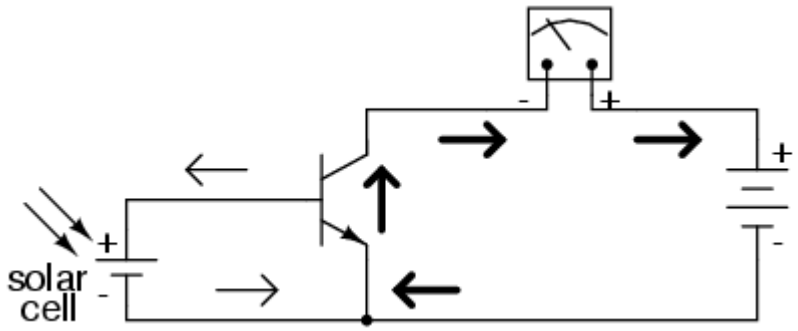
Before, this circuit was shown to illustrate how a relatively small current from a solar cell could be used to saturate a transistor, resulting in the illumination of a lamp. Knowing now that transistors are able to "throttle" their collector currents according to the amount of base current supplied by an input signal source, we should be able to see that the brightness of the lamp in this circuit is controllable by the solar cell's light exposure. When there is just a little light shone on the solar cell, the lamp will glow dimly. The lamp's brightness will steadily increase as more light falls on the solar cell.

Suppose that we were interested in using the solar cell as a light intensity instrument. We want to be able to measure the intensity of incident light with the solar cell by using its output current to drive a meter movement. It is possible to directly connect a meter movement to a solar cell for this purpose. In fact, the simplest light-exposure meters for photography work are designed like this:



While this approach might work for moderate light intensity measurements, it would not work as well for low light intensity measurements. Because the solar cell has to supply the meter movement's power needs, the system is necessarily limited in its sensitivity. Supposing that our need here is to measure very low-level light intensities, we are pressed to find another solution.

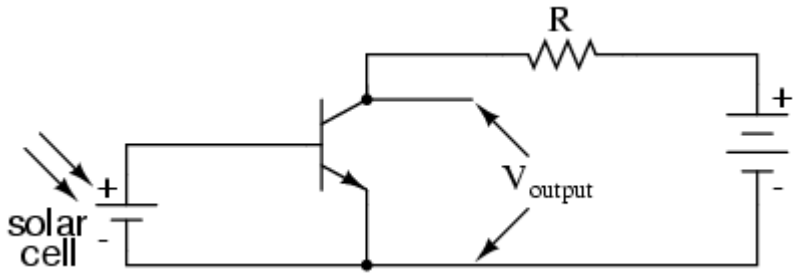
Perhaps the most direct solution to this measurement problem is to use a transistor to *amplify* the solar cell's current so that more meter movement needle deflection may be obtained for less incident light. Consider this approach:



Current through the meter movement in this circuit will be β times the solar cell current. With a transistor β of 100, this represents a substantial increase in measurement sensitivity. It is prudent to point out that the additional power to move the meter needle comes from the battery on the far right of the circuit, not the solar cell itself. All the solar cell's current does is *control* battery current to the meter to provide a greater meter reading than the solar cell could provide unaided.

Because the transistor is a current-regulating device, and because meter movement indications are based on the amount of current through their movement coils, meter indication in this circuit should depend only on the amount of current from the solar cell, not on the amount of voltage provided by the battery. This means the accuracy of the circuit will be independent of battery condition, a significant feature! All that is required of the battery is a certain minimum voltage and current output ability to be able to drive the meter full-scale if needed.

Another way in which the common-emitter configuration may be used is to produce an output *voltage* derived from the input signal, rather than a specific output *current*. Let's replace the meter movement with a plain resistor and measure voltage between collector and emitter:



With the solar cell darkened (no current), the transistor will be in cutoff mode and behave as an open switch between collector and emitter. This will produce maximum voltage drop between collector and emitter for maximum V_{output}, equal to the full voltage of the battery.

At full power (maximum light exposure), the solar cell will drive the transistor into saturation mode, making it behave like a closed switch between collector and emitter. The result will be minimum voltage drop between collector and emitter, or almost zero output voltage. In actuality, a saturated transistor can never achieve zero voltage drop between collector and emitter due to the two PN junctions through which collector current must travel. However, this "collector-emitter saturation voltage" will be fairly low, around several tenths of a volt, depending on the specific transistor used.

For light exposure levels somewhere between zero and maximum solar cell output, the transistor will be in its active mode, and the output voltage will be somewhere between zero and full battery voltage. An important quality to note here about the common-emitter configuration is that the output voltage is *inversely proportional* to the input signal strength. That is, the output voltage decreases as the input signal increases. For this reason, the common-emitter amplifier configuration is referred to as an *inverting* amplifier.

Feedback

If some percentage of an amplifier's output signal is connected to the input, so that the amplifier amplifies part of its own output signal, we have what is known as *feedback*. Feedback comes in two varieties: *positive* (also called *regenerative*), and *negative* (also called *degenerative*). Positive feedback reinforces the direction of an amplifier's output voltage change, while negative feedback does just the opposite.

A familiar example of feedback happens in public-address ("PA") systems where someone holds the microphone too close to a speaker: a high-pitched "whine" or "howl" ensues, because the audio amplifier system is detecting and amplifying its own noise. Specifically, this is an example of *positive* or *regenerative* feedback, as any sound detected by the microphone is amplified and turned into a louder sound by the speaker, which is then detected by the microphone again, and so on . . . the result being a noise of steadily increasing volume until the system becomes "saturated" and cannot produce any more volume.

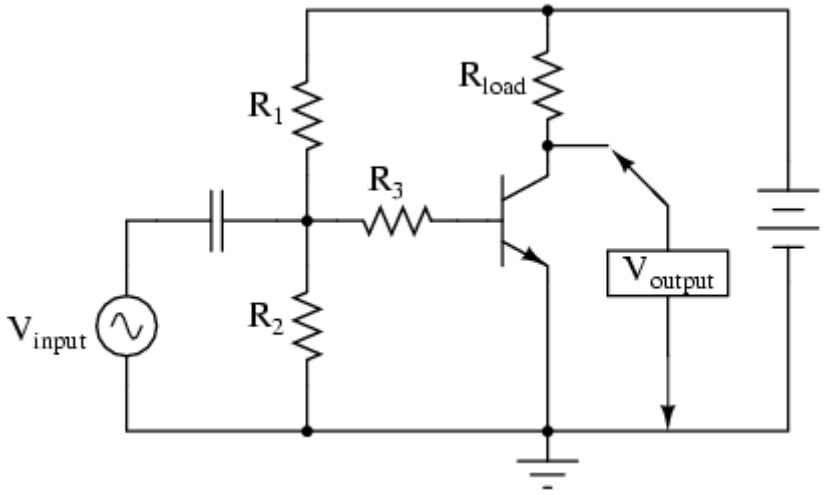
One might wonder what possible benefit feedback is to an amplifier circuit, given such an annoying example as PA system "howl." If we introduce positive, or regenerative, feedback into an amplifier circuit, it has the tendency of creating and sustaining oscillations, the frequency of which determined by the values of components handling the feedback signal from output to input. This is one way to make an *oscillator* circuit to produce AC from a DC

power supply. Oscillators are very useful circuits, and so feedback has a definite, practical application for us.

Negative feedback, on the other hand, has a "dampening" effect on an amplifier: if the output signal happens to increase in magnitude, the feedback signal introduces a decreasing influence into the input of the amplifier, thus opposing the change in output signal. While positive feedback drives an amplifier circuit toward a point of instability (oscillations), negative feedback drives it the opposite direction: toward a point of stability.

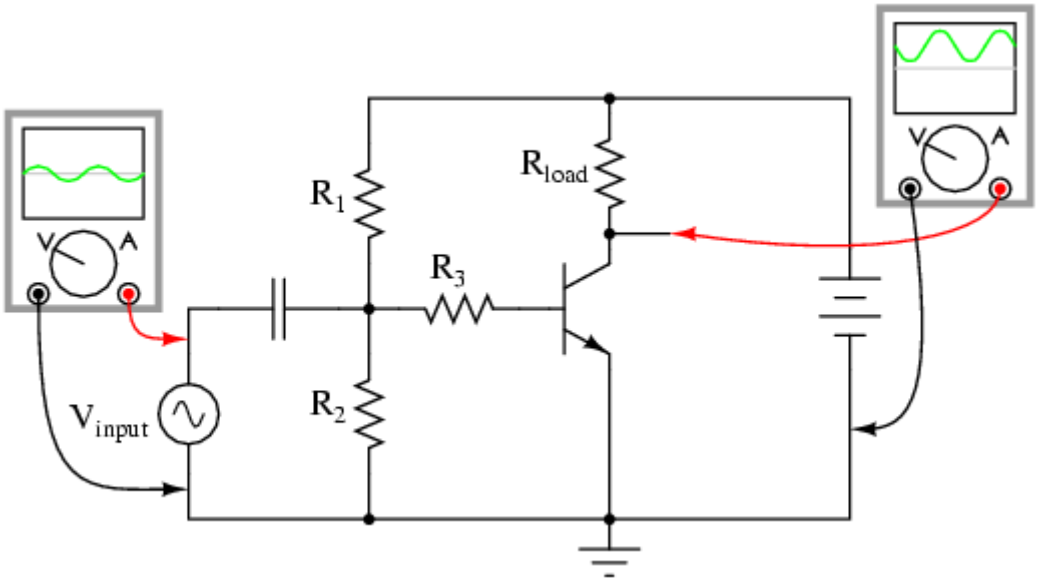
An amplifier circuit equipped with some amount of negative feedback is not only more stable, but it tends to distort the input waveform to a lesser degree and is generally capable of amplifying a wider range of frequencies. The tradeoff for these advantages (there just *has* to be a disadvantage to negative feedback, right?) is decreased gain. If a portion of an amplifier's output signal is "fed back" to the input in such a way as to oppose any changes in the output, it will require a greater input signal amplitude to drive the amplifier's output to the same amplitude as before. This constitutes a decreased gain. However, the advantages of stability, lower distortion, and greater bandwidth are worth the tradeoff in reduced gain for many applications.

Let's examine a simple amplifier circuit and see how we might introduce negative feedback into it:

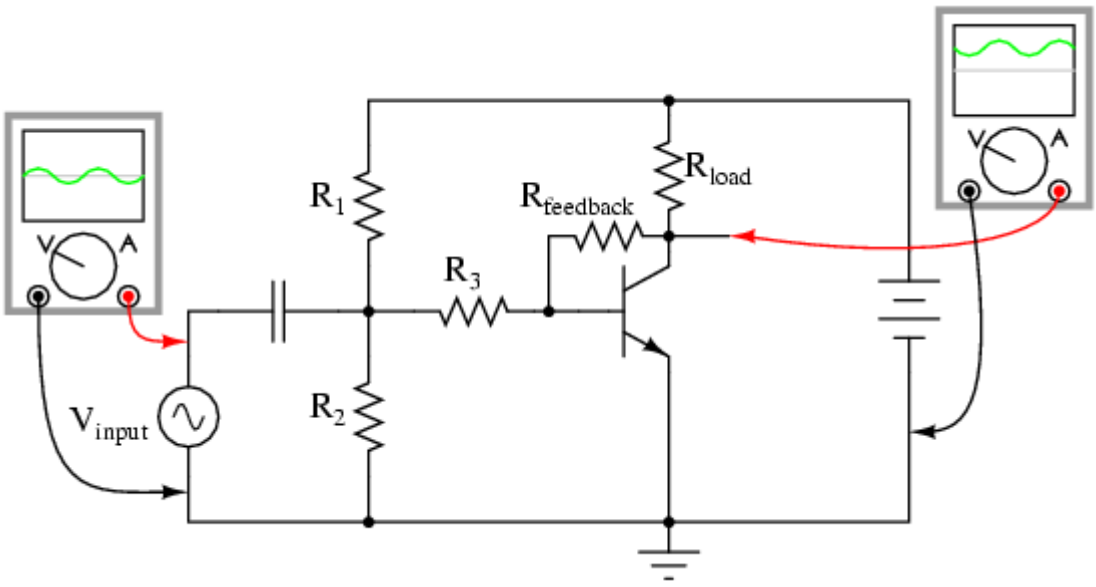


The amplifier configuration shown here is a common-emitter, with a resistor bias network formed by R_1 and R_2. The capacitor couples V_{input} to the amplifier so that the signal source doesn't have a DC voltage imposed on it by the R_1/R_2 divider network. Resistor R_3 serves the purpose of controlling voltage gain. We could omit if for maximum voltage gain, but since base resistors like this are common in common-emitter amplifier circuits, we'll keep it in this schematic.

Like all common-emitter amplifiers, this one *inverts* the input signal as it is amplified. In other words, a positive-going input voltage causes the output voltage to decrease, or go in the direction of negative, and vice versa. If we were to examine the waveforms with oscilloscopes, it would look something like this:

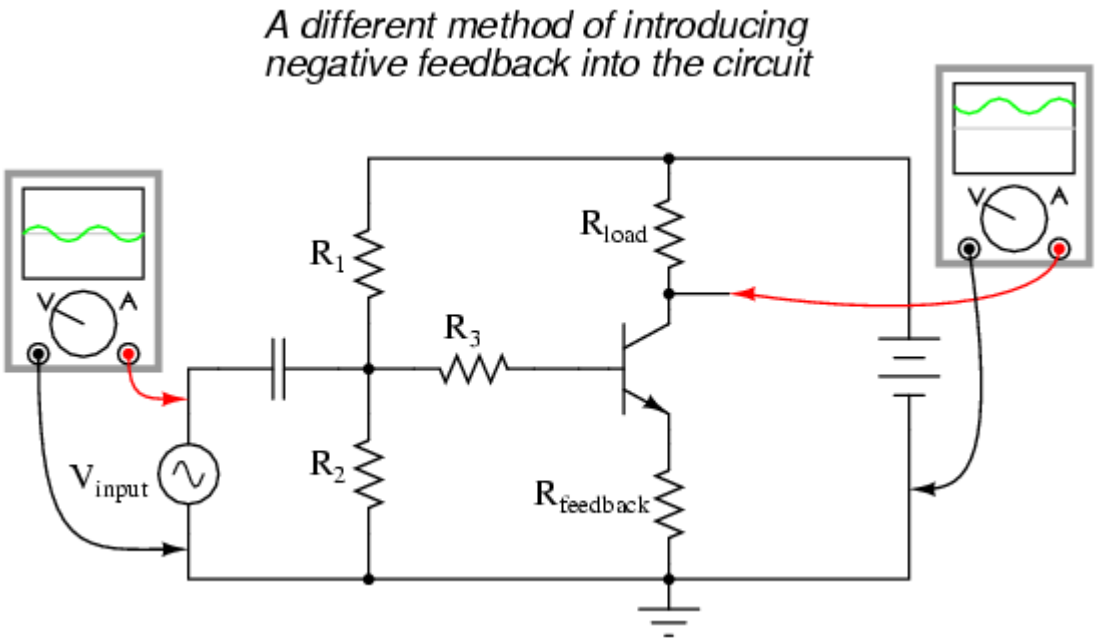


Because the output is an inverted, or mirror-image, reproduction of the input signal, any connection between the output (collector) wire and the input (base) wire of the transistor will result in *negative* feedback:

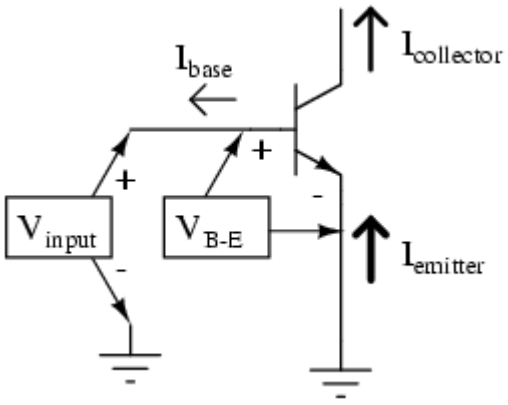


The resistances of R_1, R_2, R_3, and $R_{feedback}$ function together as a signal-mixing network so that the voltage seen at the base of the transistor (in reference to ground) is a weighted average of the input voltage and the feedback voltage, resulting in signal of reduced amplitude going into the transistor. As a result, the amplifier circuit will have reduced voltage gain, but improved linearity (reduced distortion) and increased bandwidth.

A resistor connecting collector to base is not the only way to introduce negative feedback into this amplifier circuit, though. Another method, although more difficult to understand at first, involves the placement of a resistor between the transistor's emitter terminal and circuit ground, like this:

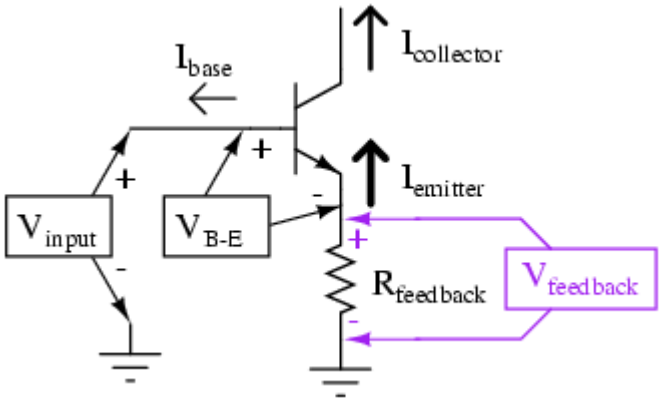


This new feedback resistor drops voltage proportional to the emitter current through the transistor, and it does so in such a way as to oppose the input signal's influence on the base-emitter junction of the transistor. Let's take a closer look at the emitter-base junction and see what difference this new resistor makes:



With no feedback resistor connecting the emitter to ground, whatever level of input signal (V_{input}) makes it through the coupling capacitor and $R_1/R_2/R_3$ resistor network will be impressed directly across the base-emitter junction as the transistor's input voltage (V_{B-E}). In other words, with no feedback resistor, V_{B-E} equals V_{input}. Therefore, if V_{input} increases by 100 mV, then V_{B-E} likewise increases by 100 mV: a change in one is the same as a change in the other, since the two voltages are equal to each other.

Now let's consider the effects of inserting a resistor ($R_{feedback}$) between the transistor's emitter lead and ground:



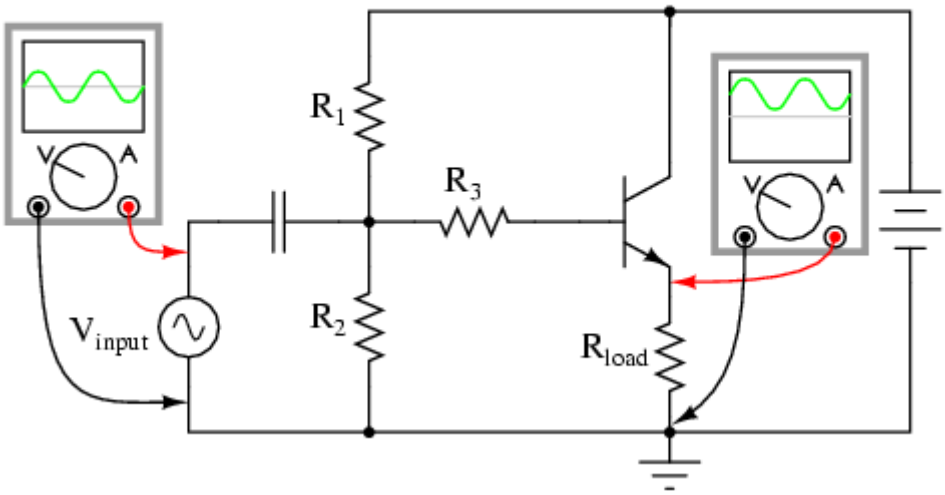
Note how the voltage dropped across $R_{feedback}$ adds with V_{B-E} to equal V_{input}. With $R_{feedback}$ in the V_{input} -- V_{B-E} loop, V_{B-E} will no longer be equal to V_{input}. We know that $R_{feedback}$ will drop a voltage proportional to emitter current, which is in turn controlled by the base current, which is in turn controlled by the voltage dropped across the base-emitter junction of the transistor (V_{B-E}). Thus, if V_{input} were to increase in a positive direction, it would increase V_{B-E}, causing more base current, causing more collector (load) current, causing more emitter current, and causing more feedback voltage to be dropped across $R_{feedback}$. This increase of voltage drop across the feedback resistor, though, *subtracts* from V_{input} to reduce the V_{B-E}, so that the actual voltage increase for V_{B-E} will be less than the voltage increase of V_{input}. No longer will a 100 mV increase in V_{input} result in a full 100 mV increase for V_{B-E}, because the two voltages are *not* equal to each other.

Consequently, the input voltage has less control over the transistor than before, and the voltage gain for the amplifier is reduced: just what we expected from negative feedback.

In practical common-emitter circuits, negative feedback isn't just a luxury; it's a necessity for stable operation. In a perfect world, we could build and operate a common-emitter transistor amplifier with no negative feedback, and have the full amplitude of V_{input} impressed across the transistor's base-emitter junction. This would give us a large voltage gain. Unfortunately, though, the relationship between base-emitter voltage and base-emitter current changes with temperature, as predicted by the "diode equation." As the transistor heats up, there will be less of a forward voltage drop across the base-emitter junction for any given current. This causes a problem for us, as the R_1/R_2 voltage divider network is designed to provide the correct quiescent current through the base of the transistor so that it will operate in whatever class of operation we desire (in this example, I've shown the amplifier working in class-A mode). If the transistor's voltage/current relationship changes with temperature, the amount of DC bias voltage necessary for the desired class of operation will change. In this case, a hot transistor will draw more bias current for the same amount of bias voltage, making it heat up even more, drawing even more bias current. The result, if unchecked, is called *thermal runaway*.

Common-collector amplifiers, however, do not suffer from thermal runaway. Why is this? The answer has everything to do with negative feedback:

A common-collector amplifier



Note that the common-collector amplifier has its load resistor placed in exactly the same spot as we had the $R_{feedback}$ resistor in the last circuit: between emitter and ground. This means that the only voltage impressed across the transistor's base-emitter junction is the *difference* between V_{input} and V_{output}, resulting in a very low voltage gain (usually close to 1 for a common-collector amplifier). Thermal runaway is impossible for this amplifier: if base current happens to increase due to transistor heating, emitter current will likewise increase, dropping more voltage across the load, which in turn *subtracts* from V_{input} to reduce the amount of voltage dropped between base and emitter. In other words, the negative feedback afforded by placement of the load resistor makes the problem of thermal runaway *self-correcting*. In exchange for a greatly reduced voltage gain, we get superb stability and immunity from thermal runaway.

By adding a "feedback" resistor between emitter and ground in a common-emitter amplifier, we make the amplifier behave a little less like an "ideal" common-emitter and a little more like a common-collector. The feedback resistor value is typically quite a bit less than the load, minimizing the amount of negative feedback and keeping the voltage gain fairly high.

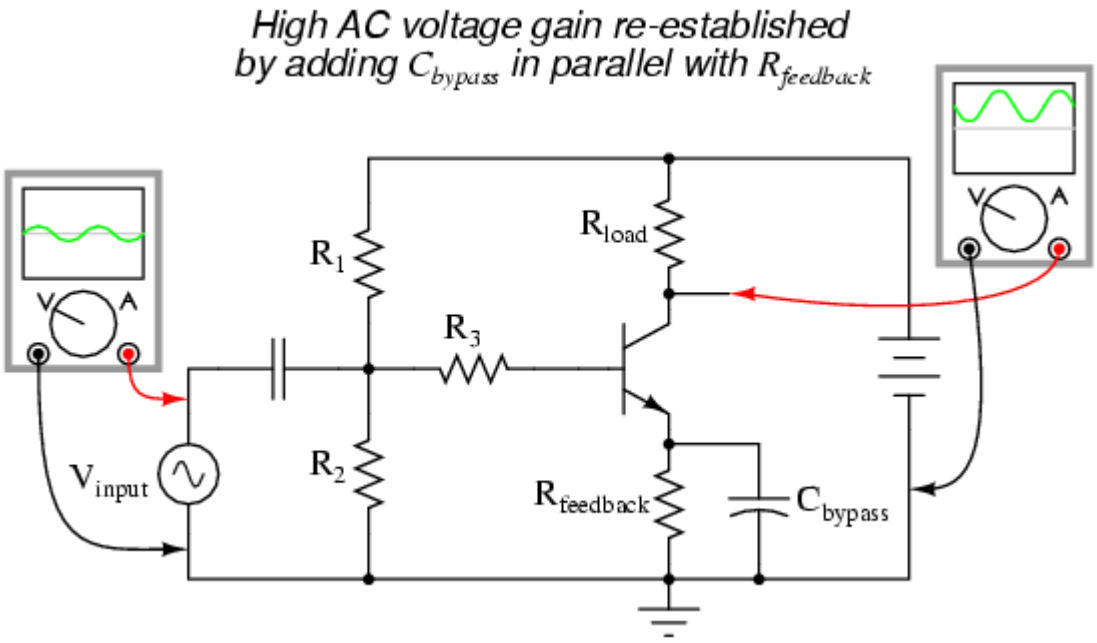
Another benefit of negative feedback, seen clearly in the common-collector circuit, is that it tends to make the voltage gain of the amplifier less dependent on the characteristics of the transistor. Note that in a common-collector amplifier, voltage gain is nearly equal to unity (1), regardless of the transistor's β. This means, among other things, that we could replace the transistor in a common-collector amplifier with one having a different β and not see any significant changes in voltage gain. In a common-emitter circuit, the voltage gain is highly dependent on β. If we were to replace the transistor in a common-emitter circuit with another of differing β, the voltage gain for the amplifier would change significantly. In a common-emitter amplifier equipped with negative feedback, the voltage gain will still be dependent upon transistor β to some degree, but not as much as before, making the circuit more predictable despite variations in transistor β.

The fact that we have to introduce negative feedback into a common-emitter amplifier to avoid thermal runaway is an unsatisfying solution. It would be nice, after all, to avoid thermal runaway without having to suppress the amplifier's inherently high voltage gain. A best-of-

both-worlds solution to this dilemma is available to us if we closely examine the nature of the problem: the voltage gain that we have to minimize in order to avoid thermal runaway is the *DC* voltage gain, not the *AC* voltage gain. After all, it isn't the AC input signal that fuels thermal runaway: it's the DC bias voltage required for a certain class of operation: that quiescent DC signal that we use to "trick" the transistor (fundamentally a DC device) into amplifying an AC signal. We can suppress DC voltage gain in a common-emitter amplifier circuit without suppressing AC voltage gain if we figure out a way to make the negative feedback function with DC only. That is, if we only feed back an inverted DC signal from output to input, but not an inverted AC signal.

The $R_{feedback}$ emitter resistor provides negative feedback by dropping a voltage proportional to load current. In other words, negative feedback is accomplished by inserting an impedance into the emitter current path. If we want to feed back DC but not AC, we need an impedance that is high for DC but low for AC. What kind of circuit presents a high impedance to DC but a low impedance to AC? A high-pass filter, of course!

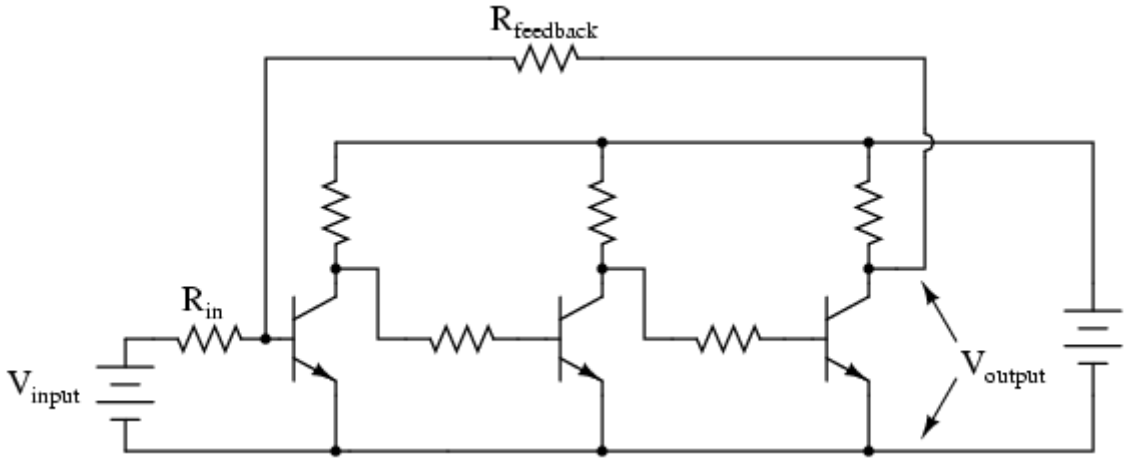
By connecting a capacitor in parallel with the feedback resistor, we create the very situation we need: a path from emitter to ground that is easier for AC than it is for DC:



The new capacitor "bypasses" AC from the transistor's emitter to ground, so that no appreciable AC voltage will be dropped from emitter to ground to "feed back" to the input and suppress voltage gain. Direct current, on the other hand, cannot go through the bypass capacitor, and so must travel through the feedback resistor, dropping a DC voltage between emitter and ground which lowers the DC voltage gain and stabilizes the amplifier's DC response, preventing thermal runaway. Because we want the reactance of this capacitor (X_C) to be as low as possible, C_{bypass} should be sized relatively large. Because the polarity across this capacitor will never change, it is safe to use a polarized (electrolytic) capacitor for the task.

Another approach to the problem of negative feedback reducing voltage gain is to use multi-stage amplifiers rather than single-transistor amplifiers. If the attenuated gain of a single transistor is insufficient for the task at hand, we can use more than one transistor to make up

for the reduction caused by feedback. Here is an example circuit showing negative feedback in a three-stage common-emitter amplifier:



Note how there is but one "path" for feedback, from the final output to the input through a single resistor, $R_{feedback}$. Since each stage is a common-emitter amplifier -- and thus inverting in nature -- and there are an odd number of stages from input to output, the output signal will be inverted with respect to the input signal, and the feedback will be negative (degenerative). Relatively large amounts of feedback may be used without sacrificing voltage gain, because the three amplifier stages provide so much gain to begin with.

At first, this design philosophy may seem inelegant and perhaps even counter-productive. Isn't this a rather crude way to overcome the loss in gain incurred through the use of negative feedback, to simply recover gain by adding stage after stage? What is the point of creating a huge voltage gain using three transistor stages if we're just going to attenuate all that gain anyway with negative feedback? The point, though perhaps not apparent at first, is increased predictability and stability from the circuit as a whole. If the three transistor stages are designed to provide an arbitrarily high voltage gain (in the tens of thousands, or greater) with no feedback, it will be found that the addition of negative feedback causes the overall voltage gain to become less dependent of the individual stage gains, and approximately equal to the simple ratio $R_{feedback}/R_{in}$. The more voltage gain the circuit has (without feedback), the more closely the voltage gain will approximate $R_{feedback}/R_{in}$ once feedback is established. In other words, voltage gain in this circuit is fixed by the values of two resistors, and nothing more.

This advantage has profound impact on mass-production of electronic circuitry: if amplifiers of predictable gain may be constructed using transistors of widely varied β values, it makes the selection and replacement of components very easy and inexpensive. It also means the amplifier's gain varies little with changes in temperature. This principle of stable gain control through a high-gain amplifier "tamed" by negative feedback is elevated almost to an art form in electronic circuits called *operational amplifiers*, or *op-amps*. You may read much more about these circuits in a later chapter of this book!

- **REVIEW:**
- *Feedback* is the coupling of an amplifier's output to its input.
- *Positive*, or *regenerative* feedback has the tendency of making an amplifier circuit unstable, so that it produces oscillations (AC). The frequency of these oscillations is largely determined by the components in the feedback network.

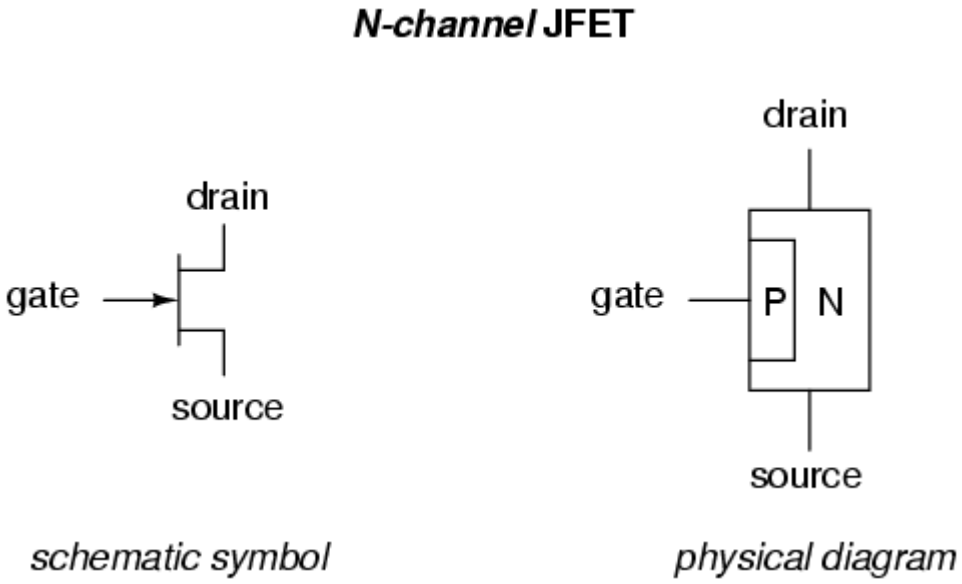
- *Negative*, or *degenerative* feedback has the tendency of making an amplifier circuit more stable, so that its output changes *less* for a given input signal than without feedback. This reduces the gain of the amplifier, but has the advantage of decreasing distortion and increasing bandwidth (the range of frequencies the amplifier can handle).
- Negative feedback may be introduced into a common-emitter circuit by coupling collector to base, or by inserting a resistor between emitter and ground.
- An emitter-to-ground "feedback" resistor is usually found in common-emitter circuits as a preventative measure against *thermal runaway*.
- Negative feedback also has the advantage of making amplifier voltage gain more dependent on resistor values and less dependent on the transistor's characteristics.
- Common-collector amplifiers have a lot of negative feedback, due to the placement of the load resistor between emitter and ground. This feedback accounts for the extremely stable voltage gain of the amplifier, as well as its immunity against thermal runaway.
- Voltage gain for a common-emitter circuit may be re-established without sacrificing immunity to thermal runaway, by connecting a *bypass capacitor* in parallel with the emitter "feedback resistor."
- If the voltage gain of an amplifier is arbitrarily high (tens of thousands, or greater), and negative feedback is used to reduce the gain to reasonable levels, it will be found that the gain will approximately equal $R_{feedback}/R_{in}$. Changes in transistor β or other internal component values will have comparatively little effect on voltage gain with feedback in operation, resulting in an amplifier that is stable and easy to design.

4.FIELD-EFFECT TRANSISTORS

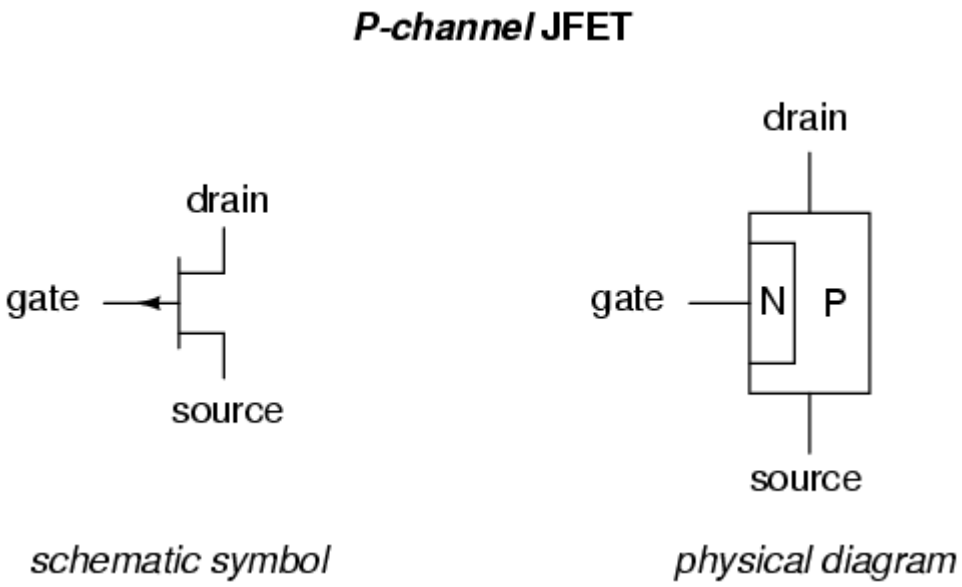
Introduction

A *transistor* is a linear semiconductor device that controls current with the application of a lower-power electrical signal. Transistors may be roughly grouped into two major divisions: *bipolar* and *field-effect*. In the last chapter we studied bipolar transistors, which utilize a small current to control a large current. In this chapter, we'll introduce the general concept of the field-effect transistor -- a device utilizing a small *voltage* to control current -- and then focus on one particular type: the *junction* field-effect transistor. In the next chapter we'll explore another type of field-effect transistor, the *insulated gate* variety.

All field-effect transistors are *unipolar* rather than *bipolar* devices. That is, the main current through them is comprised either of electrons through an N-type semiconductor or holes through a P-type semiconductor. This becomes more evident when a physical diagram of the device is seen:

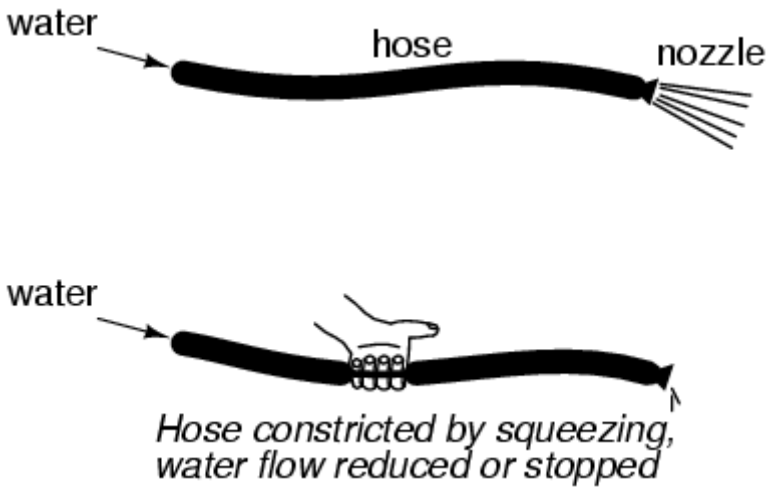


In a junction field-effect transistor, or JFET, the controlled current passes from source to drain, or from drain to source as the case may be. The controlling voltage is applied between the gate and source. Note how the current does not have to cross through a PN junction on its way between source and drain: the path (called a *channel*) is an uninterrupted block of semiconductor material. In the image just shown, this channel is an N-type semiconductor. P-type channel JFETs are also manufactured:



Generally, N-channel JFETs are more commonly used than P-channel. The reasons for this have to do with obscure details of semiconductor theory, which I'd rather not discuss in this chapter. As with bipolar transistors, I believe the best way to introduce field-effect transistor usage is to avoid theory whenever possible and concentrate instead on operational characteristics. The only practical difference between N- and P-channel JFETs you need to concern yourself with now is biasing of the PN junction formed between the gate material and the channel.

With no voltage applied between gate and source, the channel is a wide-open path for electrons to flow. However, if a voltage is applied between gate and source of such polarity that it reverse-biases the PN junction, the flow between source and drain connections becomes limited, or regulated, just as it was for bipolar transistors with a set amount of base current. Maximum gate-source voltage "pinches off" all current through source and drain, thus forcing the JFET into cutoff mode. This behavior is due to the depletion region of the PN junction expanding under the influence of a reverse-bias voltage, eventually occupying the entire width of the channel if the voltage is great enough. This action may be likened to reducing the flow of a liquid through a flexible hose by squeezing it: with enough force, the hose will be constricted enough to completely block the flow.



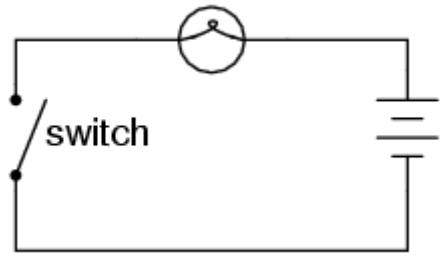
Note how this operational behavior is exactly opposite of the bipolar junction transistor. Bipolar transistors are *normally-off* devices: no current through the base, no current through the collector or the emitter. JFETs, on the other hand, are *normally-on* devices: no voltage applied to the gate allows maximum current through the source and drain. Also take note that the amount of current allowed through a JFET is determined by a *voltage* signal rather than a *current* signal as with bipolar transistors. In fact, with the gate-source PN junction reverse-biased, there should be nearly zero current through the gate connection. For this reason, we classify the JFET as a *voltage-controlled device*, and the bipolar transistor as a *current-controlled device*.

If the gate-source PN junction is forward-biased with a small voltage, the JFET channel will "open" a little more to allow greater currents through. However, the PN junction of a JFET is not built to handle any substantial current itself, and thus it is not recommended to forward-bias the junction under any circumstances.

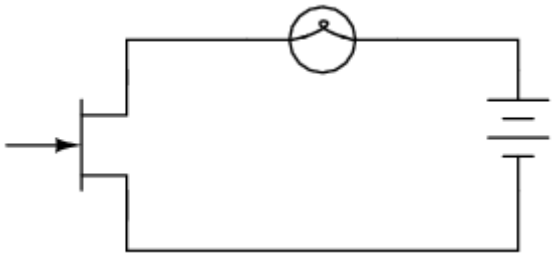
This is a very condensed overview of JFET operation. In the next section, we'll explore the use of the JFET as a switching device.

The transistor as a switch

Like its bipolar cousin, the field-effect transistor may be used as an on/off switch controlling electrical power to a load. Let's begin our investigation of the JFET as a switch with our familiar switch/lamp circuit:



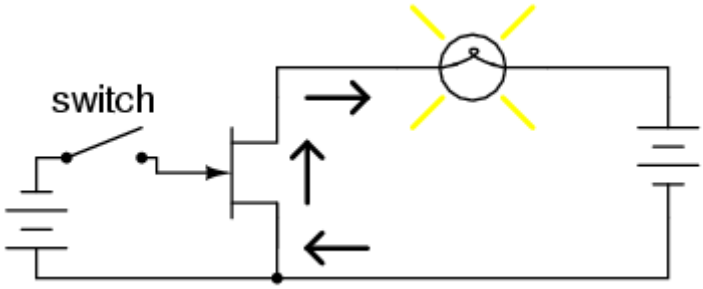
Remembering that the *controlled* current in a JFET flows between source and drain, we substitute the source and drain connections of a JFET for the two ends of the switch in the above circuit:



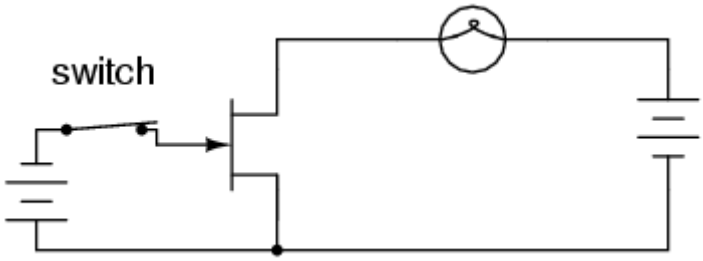
If you haven't noticed by now, the source and drain connections on a JFET look identical on the schematic symbol. Unlike the bipolar junction transistor where the emitter is clearly distinguished from the collector by the arrowhead, a JFET's source and drain lines both run perpendicular into the bar representing the semiconductor channel. This is no accident, as the

source and drain lines of a JFET are often interchangeable in practice! In other words, JFETs are usually able to handle channel current in either direction, from source to drain or from drain to source.

Now all we need in the circuit is a way to control the JFET's conduction. With zero applied voltage between gate and source, the JFET's channel will be "open," allowing full current to the lamp. In order to turn the lamp off, we will need to connect another source of DC voltage between the gate and source connections of the JFET like this:

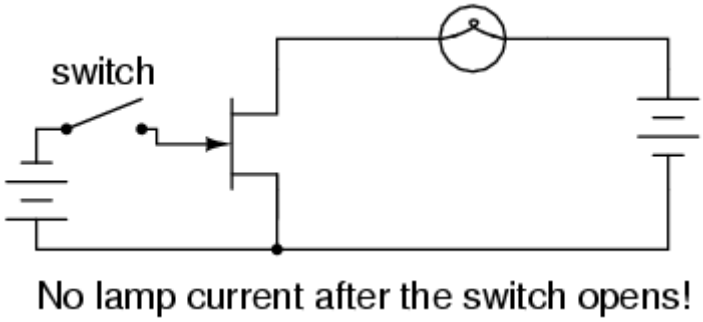


Closing this switch will "pinch off" the JFET's channel, thus forcing it into cutoff and turning the lamp off:



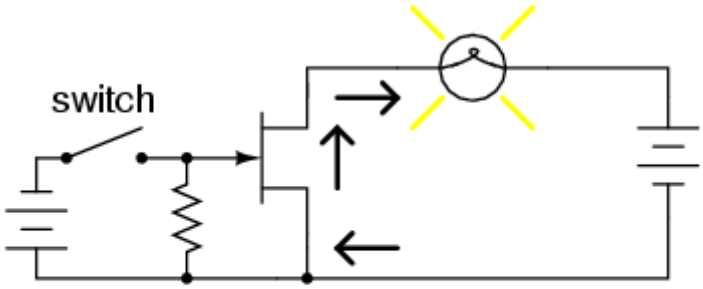
Note that there is no current going through the gate. As a reverse-biased PN junction, it firmly opposes the flow of any electrons through it. As a voltage-controlled device, the JFET requires negligible input current. This is an advantageous trait of the JFET over the bipolar transistor: there is virtually zero power required of the controlling signal.

Opening the control switch again should disconnect the reverse-biasing DC voltage from the gate, thus allowing the transistor to turn back on. Ideally, anyway, this is how it works. In practice this may not work at all:



Why is this? Why doesn't the JFET's channel open up again and allow lamp current through like it did before with no voltage applied between gate and source? The answer lies in the

operation of the reverse-biased gate-source junction. The depletion region within that junction acts as an insulating barrier separating gate from source. As such, it possesses a certain amount of *capacitance* capable of storing an electric charge potential. After this junction has been forcibly reverse-biased by the application of an external voltage, it will tend to hold that reverse-biasing voltage as a stored charge even after the source of that voltage has been disconnected. What is needed to turn the JFET on again is to bleed off that stored charge between the gate and source through a resistor:



Resistor bleeds off stored charge in PN junction to allow transistor to turn on once again.

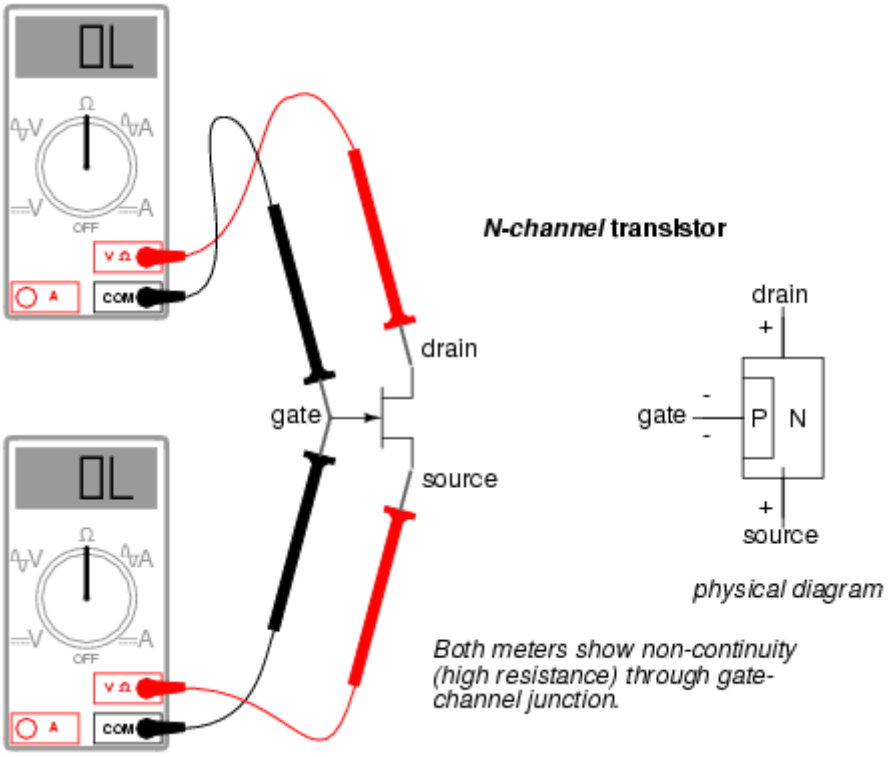
This resistor's value is not very important. The capacitance of the JFET's gate-source junction is very small, and so even a rather high-value bleed resistor creates a fast RC time constant, allowing the transistor to resume conduction with little delay once the switch is opened.

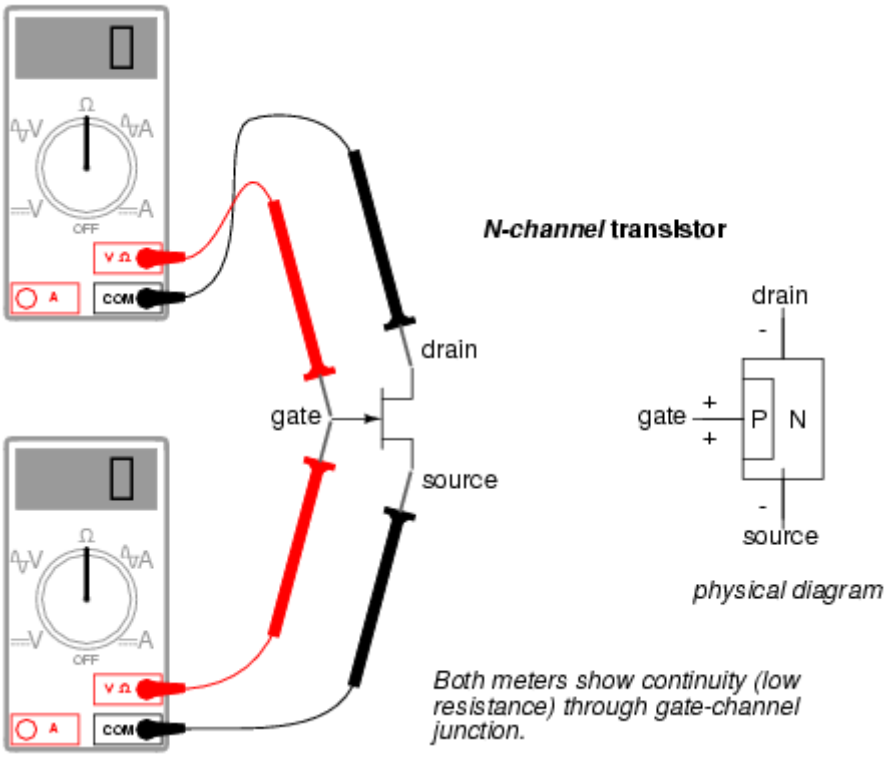
Like the bipolar transistor, it matters little where or what the controlling voltage comes from. We could use a solar cell, thermocouple, or any other sort of voltage-generating device to supply the voltage controlling the JFET's conduction. All that is required of a voltage source for JFET switch operation is *sufficient* voltage to achieve pinch-off of the JFET channel. This level is usually in the realm of a few volts DC, and is termed the *pinch-off* or *cutoff* voltage. The exact pinch-off voltage for any given JFET is a function of its unique design, and is not a universal figure like 0.7 volts is for a silicon BJT's base-emitter junction voltage.

- **REVIEW:**
- Field-effect transistors control the current between source and drain connections by a voltage applied between the gate and source. In a *junction* field-effect transistor (JFET), there is a PN junction between the gate and source which is normally reverse-biased for control of source-drain current.
- JFETs are normally-on (normally-saturated) devices. The application of a reverse-biasing voltage between gate and source causes the depletion region of that junction to expand, thereby "pinching off" the channel between source and drain through which the controlled current travels.
- It may be necessary to attach a "bleed-off" resistor between gate and source to discharge the stored charge built up across the junction's natural capacitance when the controlling voltage is removed. Otherwise, a charge may remain to keep the JFET in cutoff mode even after the voltage source has been disconnected.

Meter check of a transistor

Testing a JFET with a multimeter might seem to be a relatively easy task, seeing as how it has only one PN junction to test: either measured between gate and source, or between gate and drain.





Testing continuity through the drain-source channel is another matter, though. Remember from the last section how a stored charge across the capacitance of the gate-channel PN junction could hold the JFET in a pinched-off state without any external voltage being applied across it? This can occur even when you're holding the JFET in your hand to test it! Consequently, any meter reading of continuity through that channel will be unpredictable, since you don't necessarily know if a charge is being stored by the gate-channel junction. Of course, if you know beforehand which terminals on the device are the gate, source, and drain, you may connect a jumper wire between gate and source to eliminate any stored charge and then proceed to test source-drain continuity with no problem. However, if you *don't* know which terminals are which, the unpredictability of the source-drain connection may confuse your determination of terminal identity.

A good strategy to follow when testing a JFET is to insert the pins of the transistor into anti-static foam (the material used to ship and store static-sensitive electronic components) just prior to testing. The conductivity of the foam will make a resistive connection between all terminals of the transistor when it is inserted. This connection will ensure that all residual voltage built up across the gate-channel PN junction will be neutralized, thus "opening up" the channel for an accurate meter test of source-to-drain continuity.

Since the JFET channel is a single, uninterrupted piece of semiconductor material, there is usually no difference between the source and drain terminals. A resistance check from source to drain should yield the same value as a check from drain to source. This resistance should be relatively low (a few hundred ohms at most) when the gate-source PN junction voltage is zero. By applying a reverse-bias voltage between gate and source, pinch-off of the channel should be apparent by an increased resistance reading on the meter.

5.INSULATED-GATE FIELD-EFFECT TRANSISTORS

Introduction

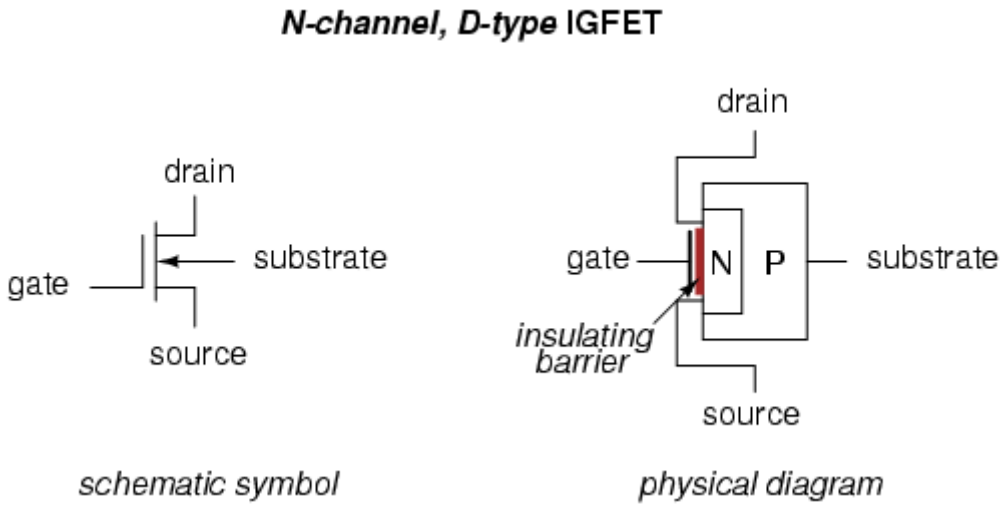
As was stated in the last chapter, there is more than one type of field-effect transistor. The junction field-effect transistor, or JFET, uses voltage applied across a reverse-biased PN junction to control the width of that junction's depletion region, which then controls the conductivity of a semiconductor channel through which the controlled current moves. Another type of field-effect device -- the insulated gate field-effect transistor, or IGFET -- exploits a similar principle of a depletion region controlling conductivity through a semiconductor channel, but it differs primarily from the JFET in that there is no *direct* connection between the gate lead and the semiconductor material itself. Rather, the gate lead is insulated from the transistor body by a thin barrier, hence the term *insulated gate*. This insulating barrier acts like the dielectric layer of a capacitor, and allows gate-to-source voltage to influence the depletion region electrostatically rather than by direct connection.

In addition to a choice of N-channel versus P-channel design, IGFETs come in two major types: *enhancement* and *depletion*. The depletion type is more closely related to the JFET, so we will begin our study of IGFETs with it.

Depletion-type IGFETs

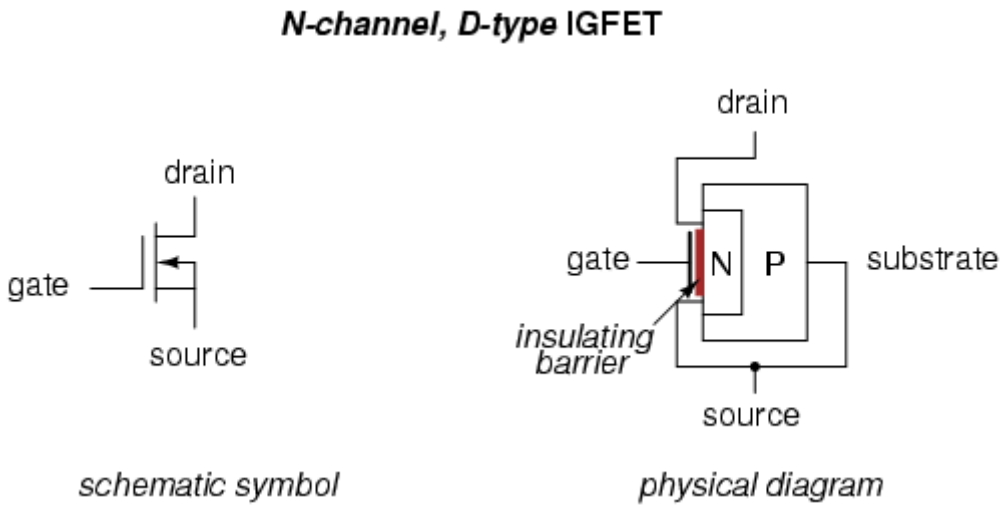
Insulated gate field-effect transistors are unipolar devices just like JFETs: that is, the controlled current does not have to cross a PN junction. There is a PN junction inside the transistor, but its only purpose is to provide that nonconducting depletion region which is used to restrict current through the channel.

Here is a diagram of an N-channel IGFET of the "depletion" type:



Notice how the source and drain leads connect to either end of the N channel, and how the gate lead attaches to a metal plate separated from the channel by a thin insulating barrier. That barrier is sometimes made from silicon dioxide (the primary chemical compound found in sand), which is a very good insulator. Due to this **M**etal (gate) - **O**xide (barrier) - **S**emiconductor (channel) construction, the IGFET is sometimes referred to as a MOSFET. There are other types of IGFET construction, though, and so "IGFET" is the better descriptor for this general class of transistors.

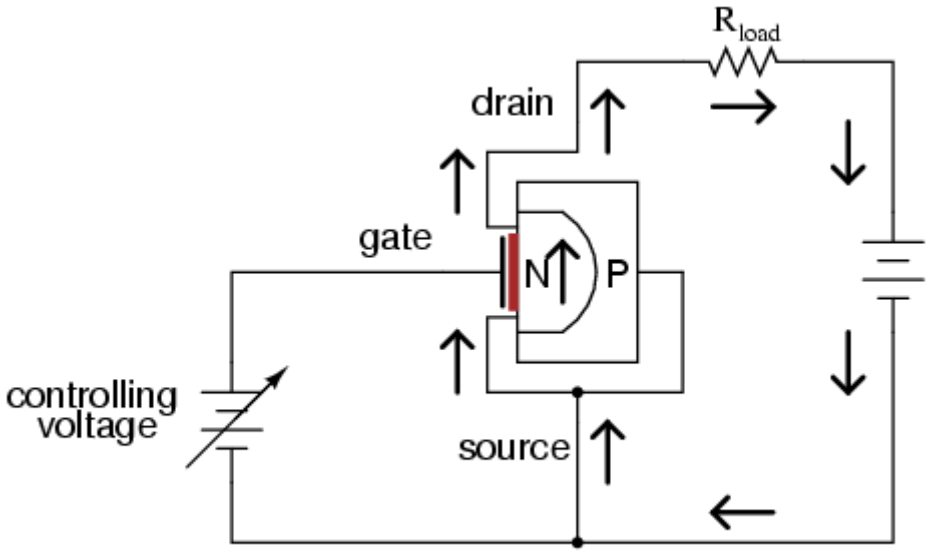
Notice also how there are four connections to the IGFET. In practice, the *substrate* lead is directly connected to the *source* lead to make the two electrically common. Usually, this connection is made internally to the IGFET, eliminating the separate substrate connection, resulting in a three-terminal device with a slightly different schematic symbol:



With source and substrate common to each other, the N and P layers of the IGFET end up being directly connected to each other through the outside wire. This connection prevents any voltage from being impressed across the PN junction. As a result, a depletion region exists between the two materials, but it can never be expanded or collapsed. JFET operation is based on the expansion of the PN junction's depletion region, but here in the IGFET that cannot happen, so IGFET operation must be based on a different effect.

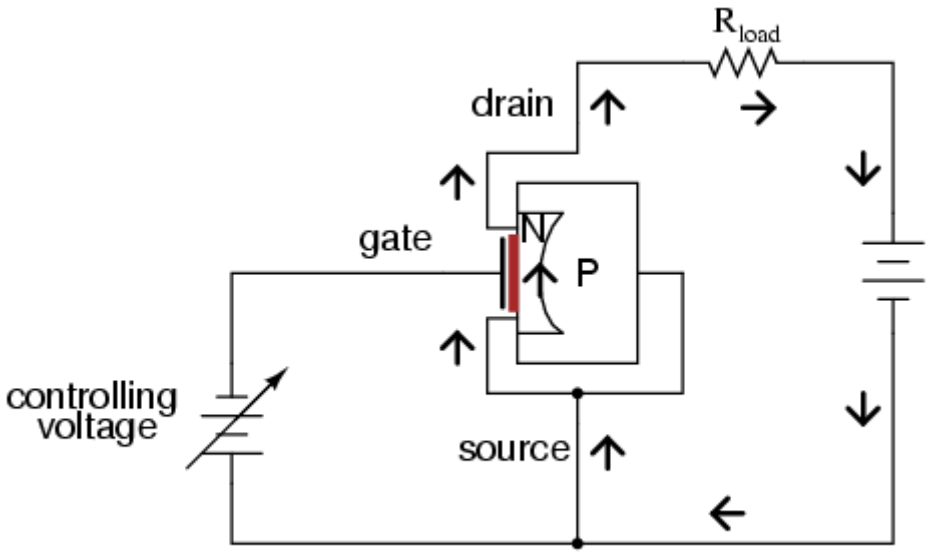
Indeed it is, for when a controlling voltage is applied between gate and source, the conductivity of the channel is changed as a result of the depletion region *moving* closer to or further away from the gate. In other words, the channel's effective width changes just as with the JFET, but this change in channel width is due to depletion region *displacement* rather than depletion region *expansion*.

In an N-channel IGFET, a controlling voltage applied positive (+) to the gate and negative (-) to the source has the effect of repelling the PN junction's depletion region, expanding the N-type channel and increasing conductivity:



Channel expands for greater conductivity

Reversing the controlling voltage's polarity has the opposite effect, attracting the depletion region and narrowing the channel, consequently reducing channel conductivity:



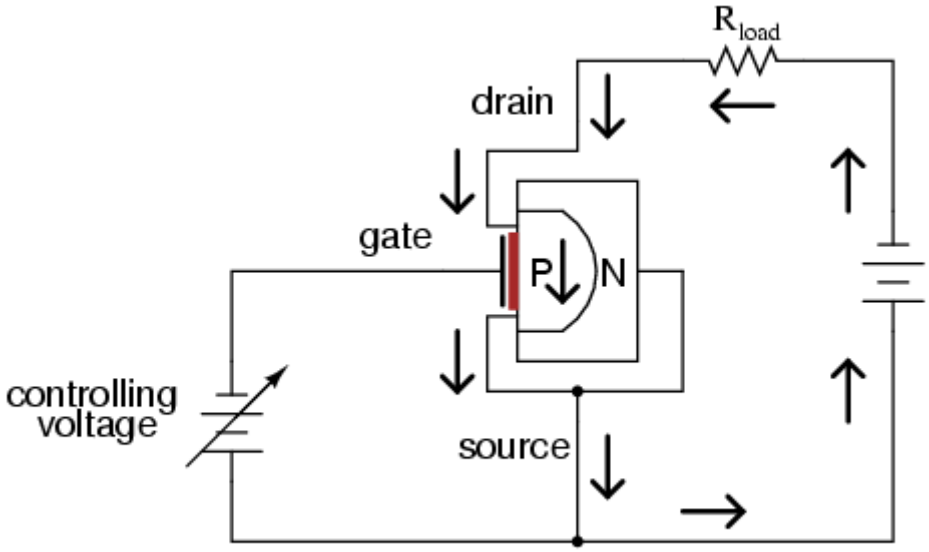
Channel narrows for less conductivity

The insulated gate allows for controlling voltages of any polarity without danger of forward-biasing a junction, as was the concern with JFETs. This type of IGFET, although it's called a "depletion-type," actually has the capability of having its channel *either* depleted (channel narrowed) *or* enhanced (channel expanded). Input voltage polarity determines which way the channel will be influenced.

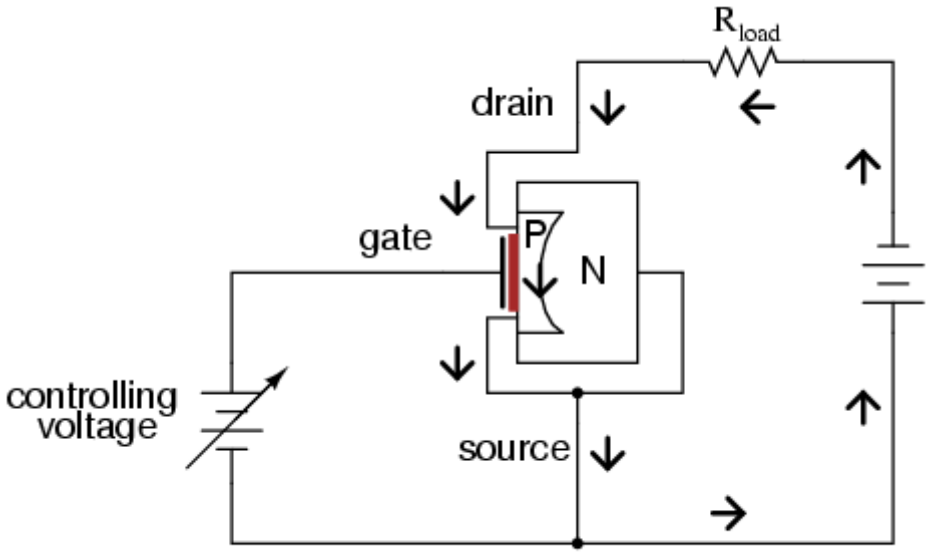
Understanding which polarity has which effect is not as difficult as it may seem. The key is to consider the type of semiconductor doping used in the channel (N-channel or P-channel?), then relate that doping type to the side of the input voltage source connected to the channel by means of the source lead. If the IGFET is an N-channel and the input voltage is connected so that the positive (+) side is on the gate while the negative (-) side is on the source, the channel

will be enhanced as extra electrons build up on the channel side of the dielectric barrier. Think, "negative (-) correlates with *N*-type, thus enhancing the channel with the right type of charge carrier (electrons) and making it more conductive." Conversely, if the input voltage is connected to an N-channel IGFET the other way, so that negative (-) connects to the gate while positive (+) connects to the source, free electrons will be "robbed" from the channel as the gate-channel capacitor charges, thus depleting the channel of majority charge carriers and making it less conductive.

For P-channel IGFETs, the input voltage polarity and channel effects follow the same rule. That is to say, it takes just the opposite polarity as an N-channel IGFET to either deplete or enhance:

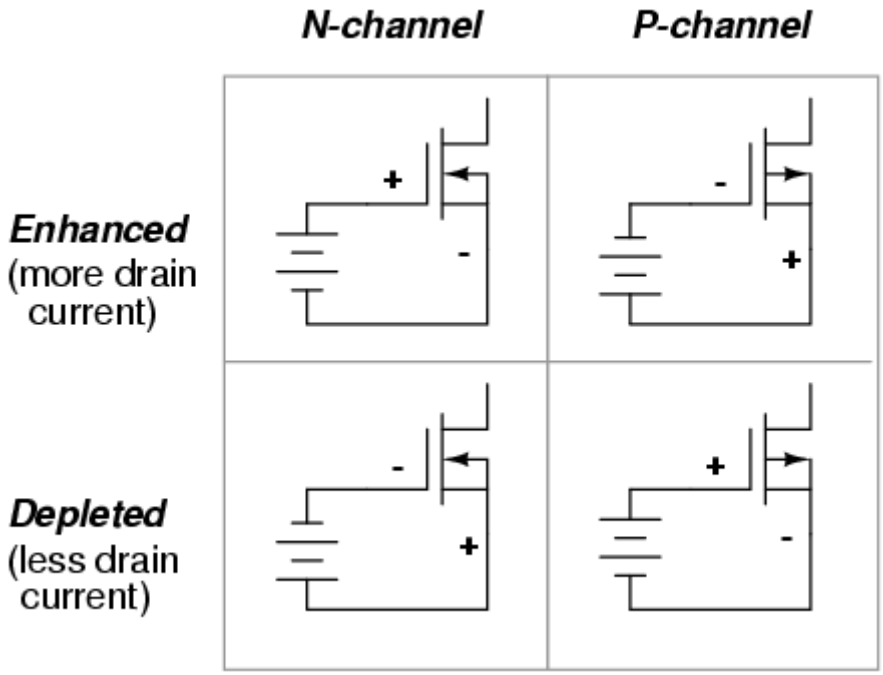


Channel expands for greater conductivity



Channel narrows for less conductivity

Illustrating the proper biasing polarities with standard IGFET symbols:



When there is zero voltage applied between gate and source, the IGFET will conduct current between source and drain, but not as much current as it would if it were enhanced by the proper gate voltage. This places the depletion-type, or simply *D-type*, IGFET in a category of its own in the transistor world. Bipolar junction transistors are *normally-off* devices: with no base current, they block any current from going through the collector. Junction field-effect transistors are *normally-on* devices: with zero applied gate-to-source voltage, they allow maximum drain current (actually, you can coax a JFET into greater drain currents by applying a very small forward-bias voltage between gate and source, but this should never be done in practice for risk of damaging its fragile PN junction). D-type IGFETs, however, are *normally half-on* devices: with no gate-to-source voltage, their conduction level is somewhere between cutoff and full saturation. Also, they will tolerate applied gate-source voltages of any polarity, the PN junction being immune from damage due to the insulating barrier and especially the direct connection between source and substrate preventing any voltage differential across the junction.

Ironically, the conduction behavior of a D-type IGFET is strikingly similar to that of an electron tube of the triode/tetrode/pentode variety. These devices were voltage-controlled current regulators that likewise allowed current through them with zero controlling voltage applied. A controlling voltage of one polarity (grid negative and cathode positive) would diminish conductivity through the tube while a voltage of the other polarity (grid positive and cathode negative) would enhance conductivity. I find it curious that one of the later transistor designs invented exhibits the same basic properties of the very first active (electronic) device.

6.OPERATIONAL AMPLIFIERS

Introduction

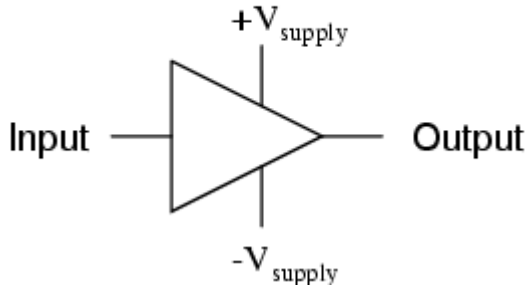
The operational amplifier is arguably the most useful single device in analog electronic circuitry. With only a handful of external components, it can be made to perform a wide variety of analog signal processing tasks. It is also quite affordable, most general-purpose amplifiers selling for under a dollar apiece. Modern designs have been engineered with durability in mind as well: several "op-amps" are manufactured that can sustain direct short-circuits on their outputs without damage.

One key to the usefulness of these little circuits is in the engineering principle of feedback, particularly *negative* feedback, which constitutes the foundation of almost all automatic control processes. The principles presented here in operational amplifier circuits, therefore, extend well beyond the immediate scope of electronics. It is well worth the electronics student's time to learn these principles and learn them well.

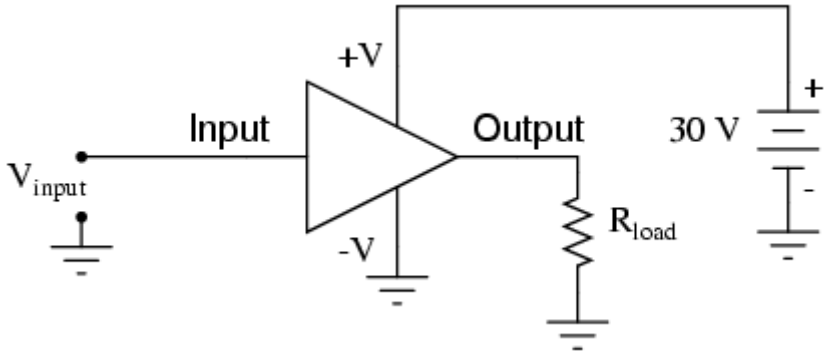
Single-ended and differential amplifiers

For ease of drawing complex circuit diagrams, electronic amplifiers are often symbolized by a simple triangle shape, where the internal components are not individually represented. This symbology is very handy for cases where an amplifier's construction is irrelevant to the greater function of the overall circuit, and it is worthy of familiarization:

General amplifier circuit symbol

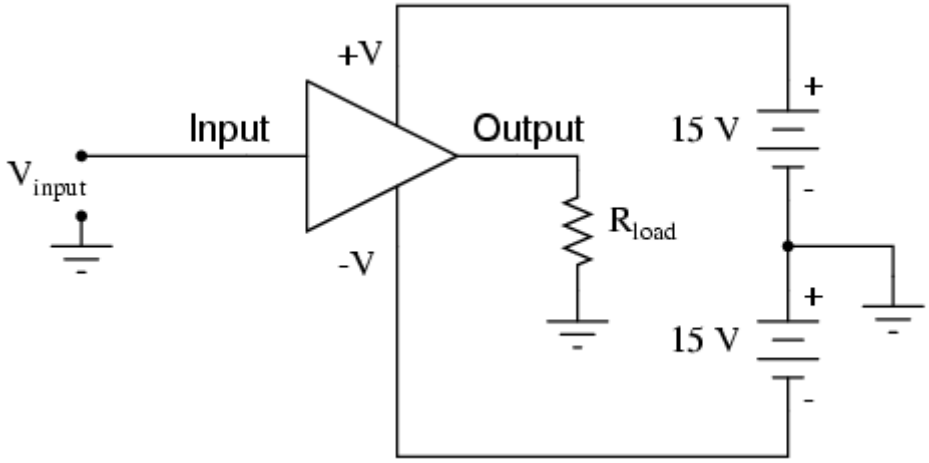


The +V and -V connections denote the positive and negative sides of the DC power supply, respectively. The input and output voltage connections are shown as single conductors, because it is assumed that all signal voltages are referenced to a common connection in the circuit called *ground*. Often (but not always!), one pole of the DC power supply, either positive or negative, is that ground reference point. A practical amplifier circuit (showing the input voltage source, load resistance, and power supply) might look like this:



Without having to analyze the actual transistor design of the amplifier, you can readily discern the whole circuit's function: to take an input signal (V_{in}), amplify it, and drive a load resistance (R_{load}). To complete the above schematic, it would be good to specify the gains of that amplifier (A_V, A_I, A_P) and the Q (bias) point for any needed mathematical analysis.

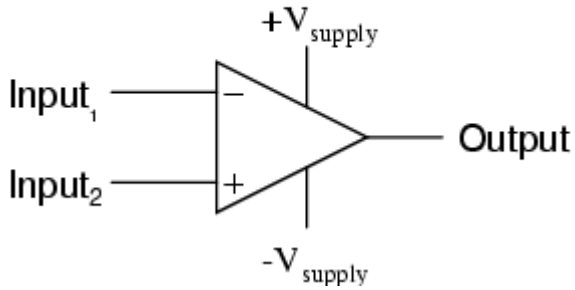
If it is necessary for an amplifier to be able to output true AC voltage (reversing polarity) to the load, a *split* DC power supply may be used, whereby the ground point is electrically "centered" between +V and -V. Sometimes the split power supply configuration is referred to as a *dual* power supply.



The amplifier is still being supplied with 30 volts overall, but with the split voltage DC power supply, the output voltage across the load resistor can now swing from a theoretical maximum of +15 volts to -15 volts, instead of +30 volts to 0 volts. This is an easy way to get true alternating current (AC) output from an amplifier without resorting to capacitive or inductive (transformer) coupling on the output. The peak-to-peak amplitude of this amplifier's output between cutoff and saturation remains unchanged.

By signifying a transistor amplifier within a larger circuit with a triangle symbol, we ease the task of studying and analyzing more complex amplifiers and circuits. One of these more complex amplifier types that we'll be studying is called the *differential amplifier*. Unlike normal amplifiers, which amplify a single input signal (often called *single-ended* amplifiers), differential amplifiers amplify the voltage difference between two input signals. Using the simplified triangle amplifier symbol, a differential amplifier looks like this:

Differential amplifier



The two input leads can be seen on the left-hand side of the triangular amplifier symbol, the output lead on the right-hand side, and the +V and -V power supply leads on top and bottom. As with the other example, all voltages are referenced to the circuit's ground point. Notice that one input lead is marked with a (-) and the other is marked with a (+). Because a differential amplifier amplifies the difference in voltage between the two inputs, each input influences the output voltage in opposite ways. Consider the following table of input/output voltages for a differential amplifier with a voltage gain of 4:

(-) $Input_1$	0	0	0	0	1	2.5	7	3	-3	-2
(+) $Input_2$	0	1	2.5	7	0	0	0	3	3	-7
Output	0	4	10	28	-4	-10	-28	0	24	-20

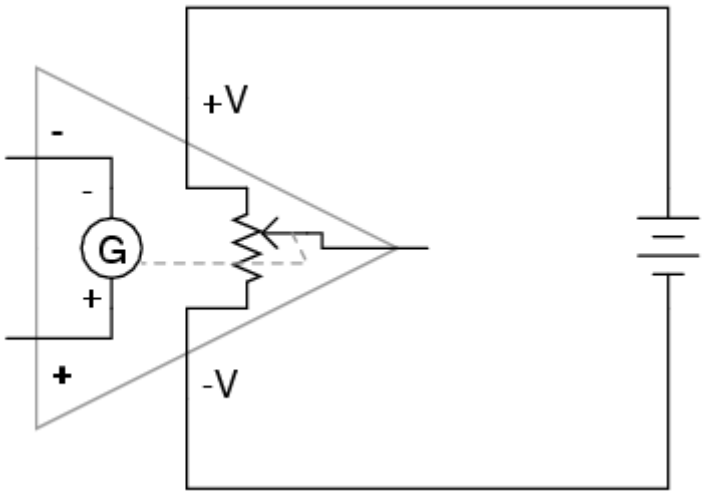
Voltage output equation: $V_{out} = A_V(Input_2 - Input_1)$

or

$V_{out} = A_V(Input_{(+)} - Input_{(-)})$

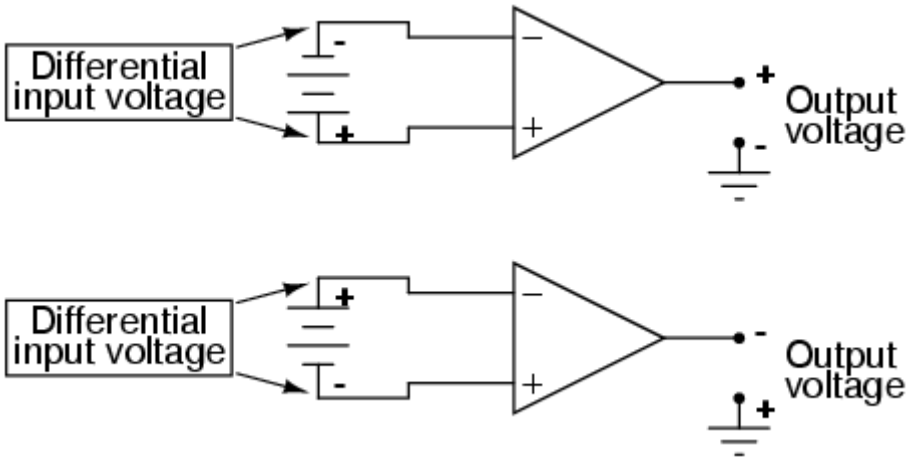
An increasingly positive voltage on the (+) input tends to drive the output voltage more positive, and an increasingly positive voltage on the (-) input tends to drive the output voltage more negative. Likewise, an increasingly negative voltage on the (+) input tends to drive the output negative as well, and an increasingly negative voltage on the (-) input does just the opposite. Because of this relationship between inputs and polarities, the (-) input is commonly referred to as the *inverting* input and the (+) as the *noninverting* input.

It may be helpful to think of a differential amplifier as a variable voltage source controlled by a sensitive voltmeter, as such:

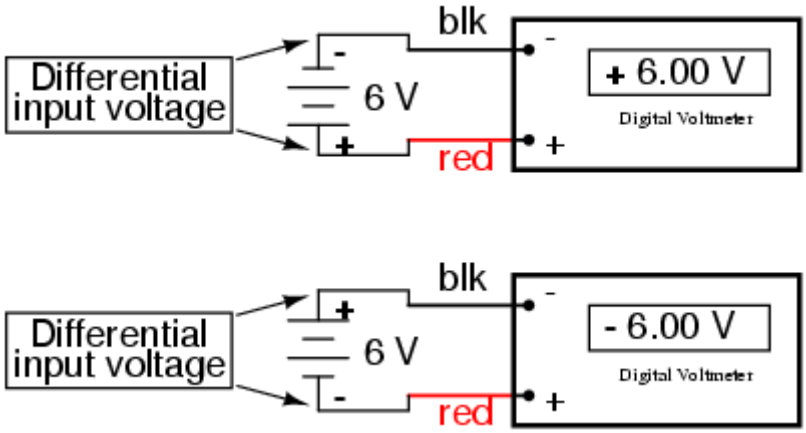


Bear in mind that the above illustration is only a *model* to aid in understanding the behavior of a differential amplifier. It is not a realistic schematic of its actual design. The "G" symbol represents a galvanometer, a sensitive voltmeter movement. The potentiometer connected between +V and -V provides a variable voltage at the output pin (with reference to one side of the DC power supply), that variable voltage set by the reading of the galvanometer. It must be understood that any load powered by the output of a differential amplifier gets its current from the DC power source (battery), *not* the input signal. The input signal (to the galvanometer) merely *controls* the output.

This concept may at first be confusing to students new to amplifiers. With all these polarities and polarity markings (- and +) around, it's easy to get confused and not know what the output of a differential amplifier will be. To address this potential confusion, here's a simple rule to remember:



When the polarity of the *differential* voltage matches the markings for inverting and noninverting inputs, the output will be positive. When the polarity of the differential voltage clashes with the input markings, the output will be negative. This bears some similarity to the mathematical sign displayed by digital voltmeters based on input voltage polarity. The red test lead of the voltmeter (often called the "positive" lead because of the color red's popular association with the positive side of a power supply in electronic wiring) is more positive than the black, the meter will display a positive voltage figure, and vice versa:



Just as a voltmeter will only display the voltage *between* its two test leads, an ideal differential amplifier only amplifies the potential difference between its two input connections, not the voltage between any one of those connections and ground. The output polarity of a differential amplifier, just like the signed indication of a digital voltmeter, depends on the relative polarities of the differential voltage between the two input connections.

If the input voltages to this amplifier represented mathematical quantities (as is the case within analog computer circuitry), or physical process measurements (as is the case within analog electronic instrumentation circuitry), you can see how a device such as a differential amplifier could be very useful. We could use it to compare two quantities to see which is greater (by the polarity of the output voltage), or perhaps we could compare the difference between two quantities (such as the level of liquid in two tanks) and flag an alarm (based on the absolute value of the amplifier output) if the difference became too great. In basic automatic control circuitry, the quantity being controlled (called the *process variable*) is compared with a target value (called the *setpoint*), and decisions are made as to how to act based on the discrepancy between these two values. The first step in electronically controlling such a scheme is to amplify the difference between the process variable and the setpoint with a differential amplifier. In simple controller designs, the output of this differential amplifier can be directly utilized to drive the final control element (such as a valve) and keep the process reasonably close to setpoint.

- **REVIEW:**
- A "shorthand" symbol for an electronic amplifier is a triangle, the wide end signifying the input side and the narrow end signifying the output. Power supply lines are often omitted in the drawing for simplicity.
- To facilitate true AC output from an amplifier, we can use what is called a *split* or *dual* power supply, with two DC voltage sources connected in series with the middle point grounded, giving a positive voltage to ground (+V) and a negative voltage to ground (-V). Split power supplies like this are frequently used in differential amplifier circuits.
- Most amplifiers have one input and one output. *Differential amplifiers* have two inputs and one output, the output signal being proportional to the difference in signals between the two inputs.
- The voltage output of a differential amplifier is determined by the following equation: $V_{out} = A_V(V_{noninv} - V_{inv})$

The "operational" amplifier

Long before the advent of digital electronic technology, computers were built to electronically perform calculations by employing voltages and currents to represent numerical quantities. This was especially useful for the simulation of physical processes. A variable voltage, for instance, might represent velocity or force in a physical system. Through the use of resistive voltage dividers and voltage amplifiers, the mathematical operations of division and multiplication could be easily performed on these signals.

The reactive properties of capacitors and inductors lend themselves well to the simulation of variables related by calculus functions. Remember how the current through a capacitor was a function of the voltage's rate of change, and how that rate of change was designated in calculus as the *derivative*? Well, if voltage across a capacitor were made to represent the velocity of an object, the current through the capacitor would represent the force required to accelerate or decelerate that object, the capacitor's capacitance representing the object's mass:

$$i_C = C \frac{dv}{dt}$$

Where,

i_C = Instantaneous current through capacitor

C = Capacitance in farads

$\frac{dv}{dt}$ = Rate of change of voltage over time

$$F = m \frac{dv}{dt}$$

Where,

F = Force applied to object

m = Mass of object

$\frac{dv}{dt}$ = Rate of change of velocity over time

This analog electronic computation of the calculus derivative function is technically known as *differentiation*, and it is a natural function of a capacitor's current in relation to the voltage applied across it. Note that this circuit requires no "programming" to perform this relatively advanced mathematical function as a digital computer would.

Electronic circuits are very easy and inexpensive to create compared to complex physical systems, so this kind of analog electronic simulation was widely used in the research and development of mechanical systems. For realistic simulation, though, amplifier circuits of high accuracy and easy configurability were needed in these early computers.

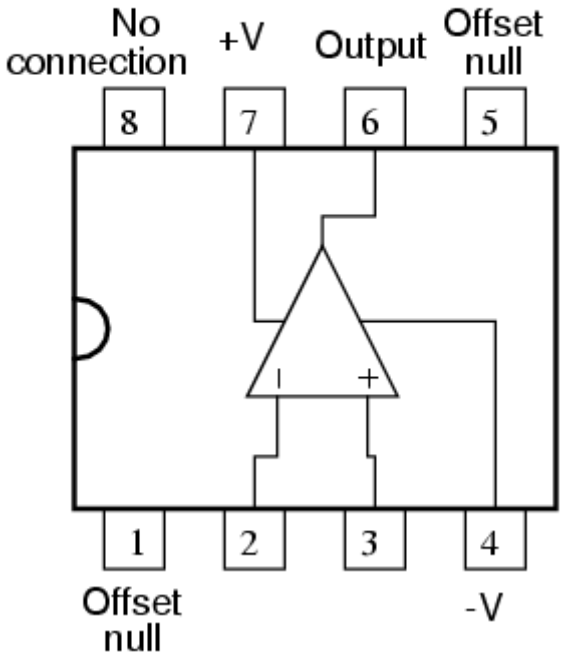
It was found in the course of analog computer design that differential amplifiers with extremely high voltage gains met these requirements of accuracy and configurability better than single-ended amplifiers with custom-designed gains. Using simple components connected to the inputs and output of the high-gain differential amplifier, virtually any gain and any function could be obtained from the circuit, overall, without adjusting or modifying the internal circuitry of the amplifier itself. These high-gain differential amplifiers came to be known as *operational amplifiers*, or *op-amps*, because of their application in analog computers' mathematical *operations*.

Modern op-amps, like the popular model 741, are high-performance, inexpensive integrated circuits. Their input impedances are quite high, the inputs drawing currents in the range of half a microamp (maximum) for the 741, and far less for op-amps utilizing field-effect input transistors. Output impedance is typically quite low, about 75 Ω for the model 741, and many

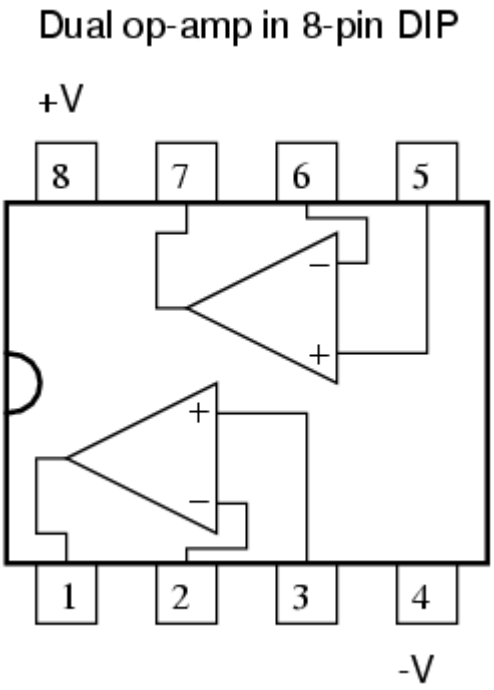
models have built-in output short circuit protection, meaning that their outputs can be directly shorted to ground without causing harm to the internal circuitry. With direct coupling between op-amps' internal transistor stages, they can amplify DC signals just as well as AC (up to certain maximum voltage-risetime limits). It would cost far more in money and time to design a comparable discrete-transistor amplifier circuit to match that kind of performance, unless high power capability was required. For these reasons, op-amps have all but obsoleted discrete-transistor signal amplifiers in many applications.

The following diagram shows the pin connections for single op-amps (741 included) when housed in an 8-pin DIP (**D**ual **I**nline **P**ackage) integrated circuit:

Typical 8-pin "DIP" op-amp integrated circuit



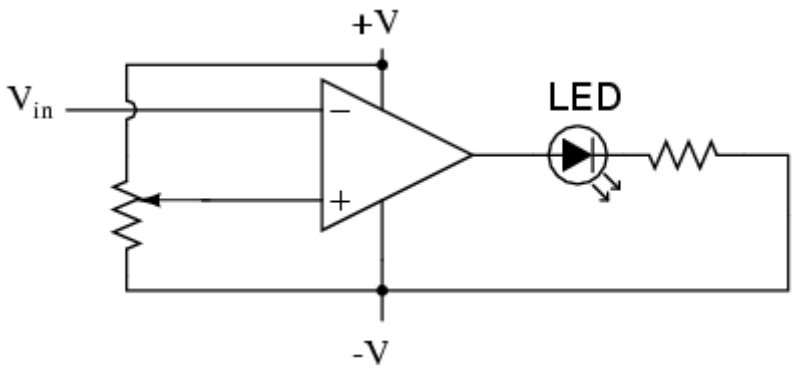
Some models of op-amp come two to a package, including the popular models TL082 and 1458. These are called "dual" units, and are typically housed in an 8-pin DIP package as well, with the following pin connections:



Operational amplifiers are also available four to a package, usually in 14-pin DIP arrangements. Unfortunately, pin assignments aren't as standard for these "quad" op-amps as they are for the "dual" or single units. Consult the manufacturer datasheet(s) for details.

Practical operational amplifier voltage gains are in the range of 200,000 or more, which makes them almost useless as an analog differential amplifier by themselves. For an op-amp with a voltage gain (A_V) of 200,000 and a maximum output voltage swing of +15V/-15V, all it would take is a differential input voltage of 75 µV (microvolts) to drive it to saturation or cutoff! Before we take a look at how external components are used to bring the gain down to a reasonable level, let's investigate applications for the "bare" op-amp by itself.

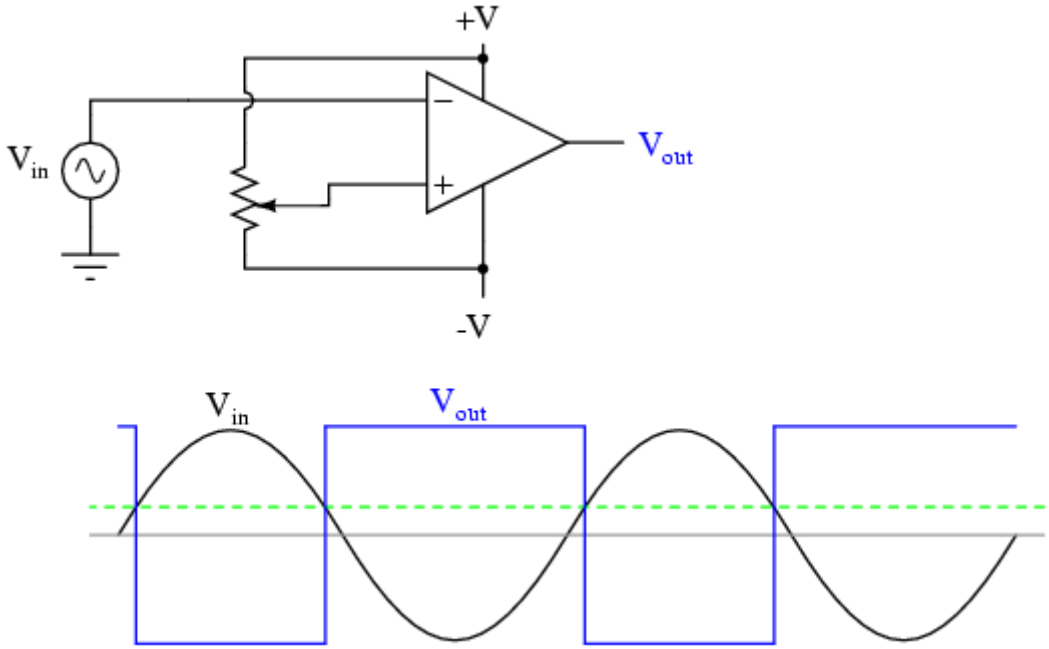
One application is called the *comparator*. For all practical purposes, we can say that the output of an op-amp will be saturated fully positive if the (+) input is more positive than the (-) input, and saturated fully negative if the (+) input is less positive than the (-) input. In other words, an op-amp's extremely high voltage gain makes it useful as a device to compare two voltages and change output voltage states when one input exceeds the other in magnitude.



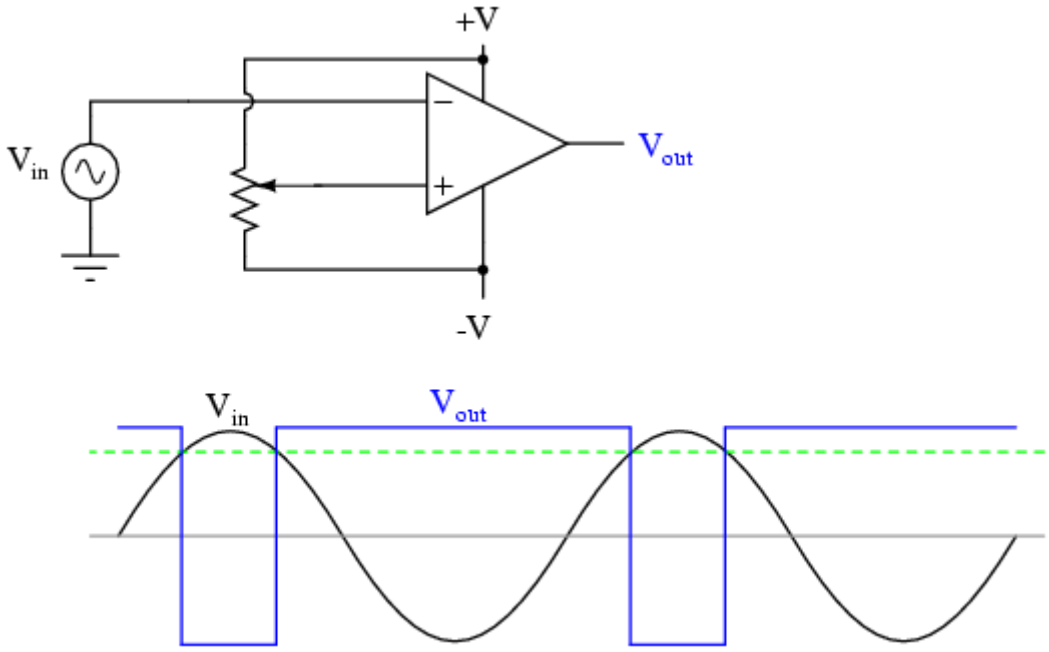
In the above circuit, we have an op-amp connected as a comparator, comparing the input voltage with a reference voltage set by the potentiometer (R_1). If V_{in} drops below the voltage

set by R_1, the op-amp's output will saturate to +V, thereby lighting up the LED. Otherwise, if V_{in} is above the reference voltage, the LED will remain off. If V_{in} is a voltage signal produced by a measuring instrument, this comparator circuit could function as a "low" alarm, with the trip-point set by R_1. Instead of an LED, the op-amp output could drive a relay, a transistor, an SCR, or any other device capable of switching power to a load such as a solenoid valve, to take action in the event of a low alarm.

Another application for the comparator circuit shown is a square-wave converter. Suppose that the input voltage applied to the inverting (-) input was an AC sine wave rather than a stable DC voltage. In that case, the output voltage would transition between opposing states of saturation whenever the input voltage was equal to the reference voltage produced by the potentiometer. The result would be a square wave:

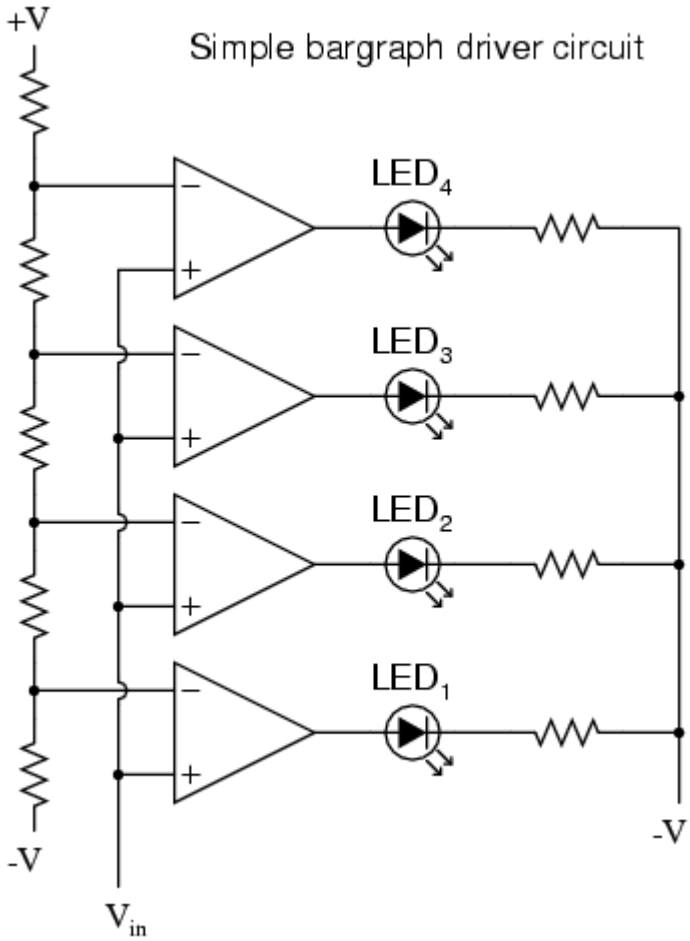


Adjustments to the potentiometer setting would change the reference voltage applied to the noninverting (+) input, which would change the points at which the sine wave would cross, changing the on/off times, or *duty cycle* of the square wave:



It should be evident that the AC input voltage would not have to be a sine wave in particular for this circuit to perform the same function. The input voltage could be a triangle wave, sawtooth wave, or any other sort of wave that ramped smoothly from positive to negative to positive again. This sort of comparator circuit is very useful for creating square waves of varying duty cycle. This technique is sometimes referred to as *pulse-width modulation*, or PWM (varying, or *modulating* a waveform according to a controlling signal, in this case the signal produced by the potentiometer).

Another comparator application is that of the bargraph driver. If we had several op-amps connected as comparators, each with its own reference voltage connected to the inverting input, but each one monitoring the same voltage signal on their noninverting inputs, we could build a bargraph-style meter such as what is commonly seen on the face of stereo tuners and graphic equalizers. As the signal voltage (representing radio signal strength or audio sound level) increased, each comparator would "turn on" in sequence and send power to its respective LED. With each comparator switching "on" at a different level of audio sound, the number of LED's illuminated would indicate how strong the signal was.



In the circuit shown above, LED_1 would be the first to light up as the input voltage increased in a positive direction. As the input voltage continued to increase, the other LED's would illuminate in succession, until all were lit.

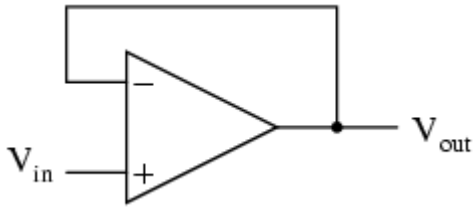
This very same technology is used in some analog-to-digital signal converters, namely the *flash converter*, to translate an analog signal quantity into a series of on/off voltages representing a digital number.

- **REVIEW:**
- A triangle shape is the generic symbol for an amplifier circuit, the wide end signifying the input and the narrow end signifying the output.
- Unless otherwise specified, *all* voltages in amplifier circuits are referenced to a common *ground* point, usually connected to one terminal of the power supply. This way, we can speak of a certain amount of voltage being "on" a single wire, while realizing that voltage is *always* measured between two points.
- A *differential amplifier* is one amplifying the voltage *difference* between two signal inputs. In such a circuit, one input tends to drive the output voltage to the same polarity of the input signal, while the other input does just the opposite. Consequently, the first input is called the *noninverting* (+) input and the second is called the *inverting* (-) input.
- An *operational amplifier* (or *op-amp* for short) is a differential amplifier with an extremely high voltage gain (A_V = 200,000 or more). Its name hails from its original use in analog computer circuitry (performing mathematical *operations*).

- Op-amps typically have very high input impedances and fairly low output impedances.
- Sometimes op-amps are used as signal *comparators*, operating in full cutoff or saturation mode depending on which input (inverting or noninverting) has the greatest voltage. Comparators are useful in detecting "greater-than" signal conditions (comparing one to the other).
- One comparator application is called the *pulse-width modulator*, and is made by comparing a sine-wave AC signal against a DC reference voltage. As the DC reference voltage is adjusted, the square-wave output of the comparator changes its duty cycle (positive versus negative times). Thus, the DC reference voltage controls, or *modulates* the pulse width of the output voltage.

Negative feedback

If we connect the output of an op-amp to its inverting input and apply a voltage signal to the noninverting input, we find that the output voltage of the op-amp closely follows that input voltage (I've neglected to draw in the power supply, +V/-V wires, and ground symbol for simplicity):

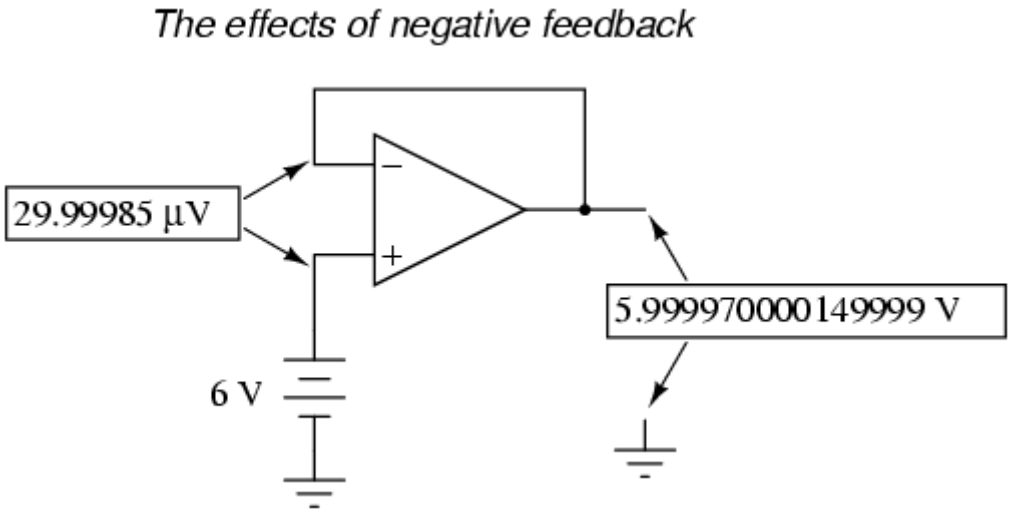


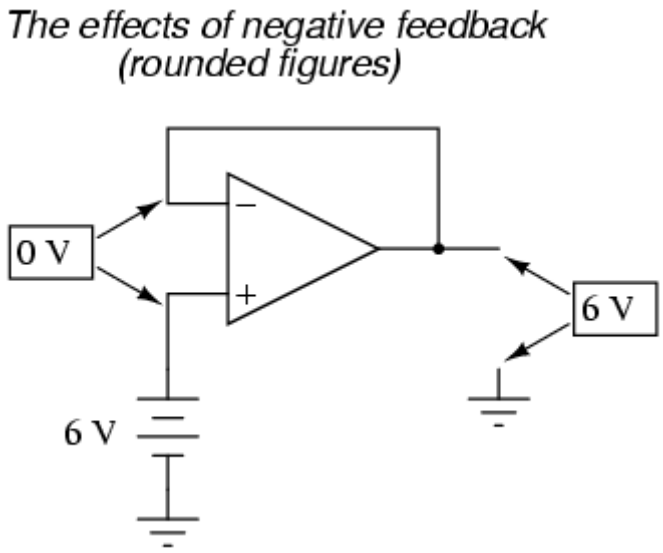
As V_{in} increases, V_{out} will increase in accordance with the differential gain. However, as V_{out} increases, that output voltage is fed back to the inverting input, thereby acting to decrease the voltage differential between inputs, which acts to bring the output down. What will happen for any given voltage input is that the op-amp will output a voltage very nearly equal to V_{in}, but just low enough so that there's enough voltage difference left between V_{in} and the (-) input to be amplified to generate the output voltage.

The circuit will quickly reach a point of stability (known as *equilibrium* in physics), where the output voltage is just the right amount to maintain the right amount of differential, which in turn produces the right amount of output voltage. Taking the op-amp's output voltage and coupling it to the inverting input is a technique known as *negative feedback*, and it is the key to having a self-stabilizing system (this is true not only of op-amps, but of any dynamic system in general). This stability gives the op-amp the capacity to work in its linear (active) mode, as opposed to merely being saturated fully "on" or "off" as it was when used as a comparator, with no feedback at all.

Because the op-amp's gain is so high, the voltage on the inverting input can be maintained almost equal to V_{in}. Let's say that our op-amp has a differential voltage gain of 200,000. If V_{in} equals 6 volts, the output voltage will be 5.999970000149999 volts. This creates just enough differential voltage (6 volts - 5.999970000149999 volts = 29.99985 µV) to cause 5.999970000149999 volts to be manifested at the output terminal, and the system holds there in balance. As you can see, 29.99985 µV is not a lot of differential, so for practical

calculations, we can assume that the differential voltage between the two input wires is held by negative feedback exactly at 0 volts.

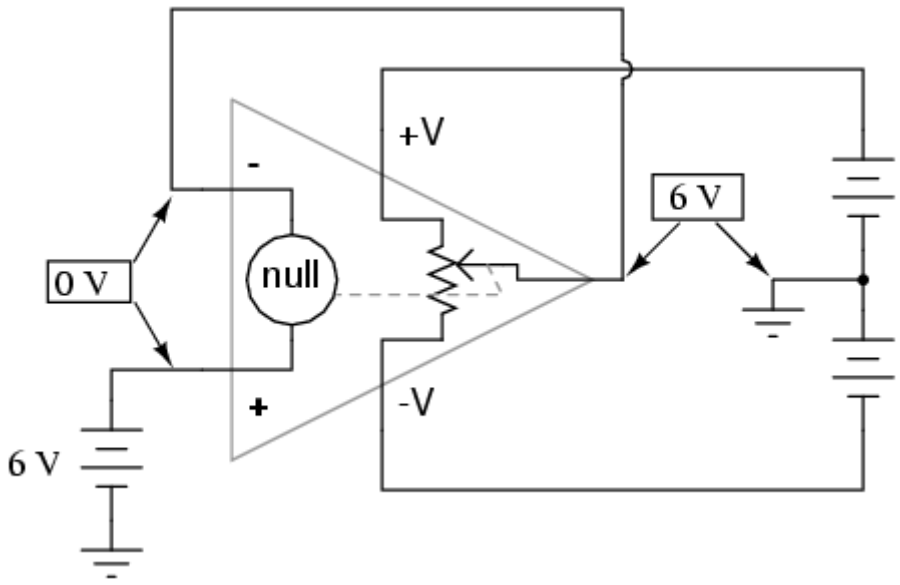




One great advantage to using an op-amp with negative feedback is that the actual voltage gain of the op-amp doesn't matter, so long as it's very large. If the op-amp's differential gain were 250,000 instead of 200,000, all it would mean is that the output voltage would hold just a little closer to V_{in} (less differential voltage needed between inputs to generate the required output). In the circuit just illustrated, the output voltage would still be (for all practical purposes) equal to the non-inverting input voltage. Op-amp gains, therefore, do not have to be precisely set by the factory in order for the circuit designer to build an amplifier circuit with precise gain. Negative feedback makes the system self-correcting. The above circuit as a whole will simply follow the input voltage with a stable gain of 1.

Going back to our differential amplifier model, we can think of the operational amplifier as being a variable voltage source controlled by an extremely sensitive *null detector*, the kind of meter movement or other sensitive measurement device used in bridge circuits to detect a condition of balance (zero volts). The "potentiometer" inside the op-amp creating the variable

voltage will move to whatever position it must to "balance" the inverting and noninverting input voltages so that the "null detector" has zero voltage across it:



As the "potentiometer" will move to provide an output voltage necessary to satisfy the "null detector" at an "indication" of zero volts, the output voltage becomes equal to the input voltage: in this case, 6 volts. If the input voltage changes at all, the "potentiometer" inside the op-amp will change position to hold the "null detector" in balance (indicating zero volts), resulting in an output voltage approximately equal to the input voltage at all times.

This will hold true within the range of voltages that the op-amp can output. With a power supply of +15V/-15V, and an ideal amplifier that can swing its output voltage just as far, it will faithfully "follow" the input voltage between the limits of +15 volts and -15 volts. For this reason, the above circuit is known as a *voltage follower*. Like its one-transistor counterpart, the common-collector ("emitter-follower") amplifier, it has a voltage gain of 1, a high input impedance, a low output impedance, and a high current gain. Voltage followers are also known as *voltage buffers*, and are used to boost the current-sourcing ability of voltage signals too weak (too high of source impedance) to directly drive a load. The op-amp model shown in the last illustration depicts how the output voltage is essentially isolated from the input voltage, so that current on the output pin is not supplied by the input voltage source at all, but rather from the power supply powering the op-amp.

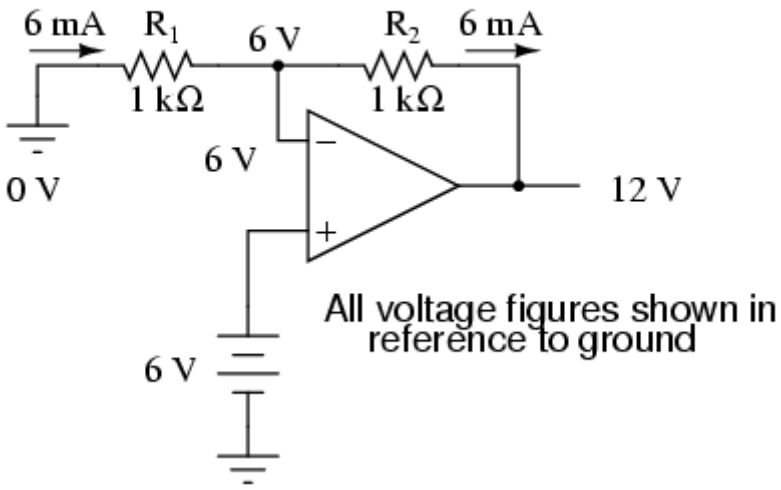
It should be mentioned that many op-amps cannot swing their output voltages exactly to +V/-V power supply rail voltages. The model 741 is one of those that cannot: when saturated, its output voltage peaks within about one volt of the +V power supply voltage and within about 2 volts of the -V power supply voltage. Therefore, with a split power supply of +15/-15 volts, a 741 op-amp's output may go as high as +14 volts or as low as -13 volts (approximately), but no further. This is due to its bipolar transistor design. These two voltage limits are known as the *positive saturation voltage* and *negative saturation voltage*, respectively. Other op-amps, such as the model 3130 with field-effect transistors in the final output stage, have the ability to swing their output voltages within millivolts of either power supply *rail* voltage. Consequently, their positive and negative saturation voltages are practically equal to the supply voltages.

- **REVIEW:**
- Connecting the output of an op-amp to its inverting (-) input is called *negative feedback*. This term can be broadly applied to any dynamic system where the output signal is "fed back" to the input somehow so as to reach a point of equilibrium (balance).
- When the output of an op-amp is *directly* connected to its inverting (-) input, a *voltage follower* will be created. Whatever signal voltage is impressed upon the noninverting (+) input will be seen on the output.
- An op-amp with negative feedback will try to drive its output voltage to whatever level necessary so that the differential voltage between the two inputs is practically zero. The higher the op-amp differential gain, the closer that differential voltage will be to zero.
- Some op-amps cannot produce an output voltage equal to their supply voltage when saturated. The model 741 is one of these. The upper and lower limits of an op-amp's output voltage swing are known as *positive saturation voltage* and *negative saturation voltage*, respectively.

Divided feedback

If we add a voltage divider to the negative feedback wiring so that only a *fraction* of the output voltage is fed back to the inverting input instead of the full amount, the output voltage will be a *multiple* of the input voltage (please bear in mind that the power supply connections to the op-amp have been omitted once again for simplicity's sake):

The effects of divided negative feedback

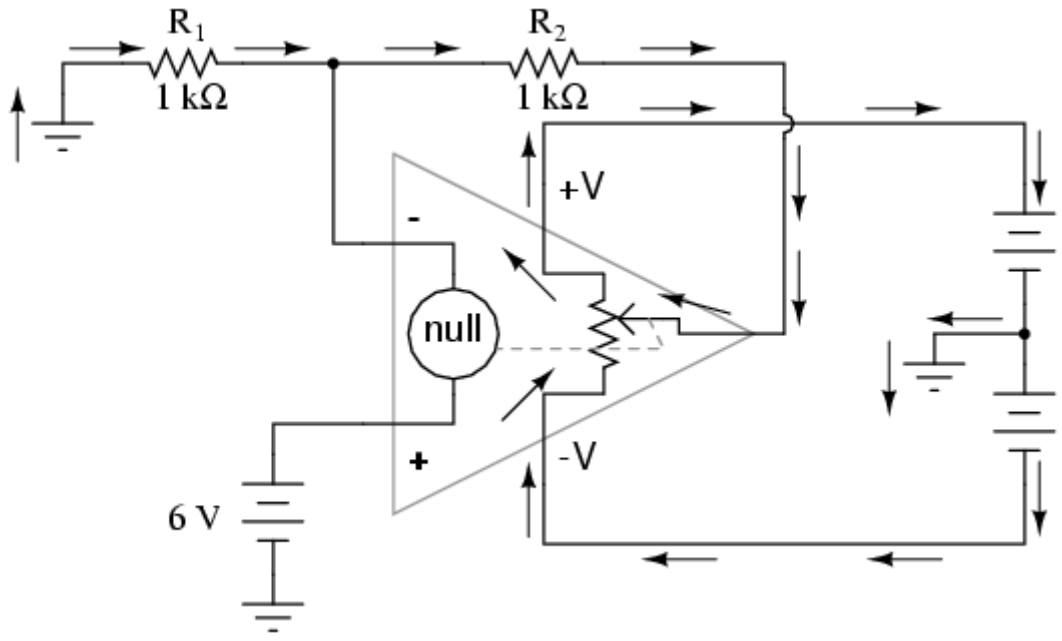


If R_1 and R_2 are both equal and V_{in} is 6 volts, the op-amp will output whatever voltage is needed to drop 6 volts across R_1 (to make the inverting input voltage equal to 6 volts, as well, keeping the voltage difference between the two inputs equal to zero). With the 2:1 voltage divider of R_1 and R_2, this will take 12 volts at the output of the op-amp to accomplish.

Another way of analyzing this circuit is to start by calculating the magnitude and direction of current through R_1, knowing the voltage on either side (and therefore, by subtraction, the

voltage across R_1), and R_1's resistance. Since the left-hand side of R_1 is connected to ground (0 volts) and the right-hand side is at a potential of 6 volts (due to the negative feedback holding that point equal to V_{in}), we can see that we have 6 volts across R_1. This gives us 6 mA of current through R_1 from left to right. Because we know that both inputs of the op-amp have extremely high impedance, we can safely assume they won't add or subtract any current through the divider. In other words, we can treat R_1 and R_2 as being in series with each other: all of the electrons flowing through R_1 must flow through R_2. Knowing the current through R_2 and the resistance of R_2, we can calculate the voltage across R_2 (6 volts), and its polarity. Counting up voltages from ground (0 volts) to the right-hand side of R_2, we arrive at 12 volts on the output.

Upon examining the last illustration, one might wonder, "where does that 1 mA of current go?" The last illustration doesn't show the entire current path, but in reality it comes from the negative side of the DC power supply, through ground, through R_1, through R_2, through the output pin of the op-amp, and then back to the positive side of the DC power supply through the output transistor(s) of the op-amp. Using the null detector/potentiometer model of the op-amp, the current path looks like this:



The 6 volt signal source does not have to supply any current for the circuit: it merely commands the op-amp to balance voltage between the inverting (-) and noninverting (+) input pins, and in so doing produce an output voltage that is twice the input due to the dividing effect of the two 1 kΩ resistors.

We can change the voltage gain of this circuit, overall, just by adjusting the values of R_1 and R_2 (changing the ratio of output voltage that is fed back to the inverting input). Gain can be calculated by the following formula:

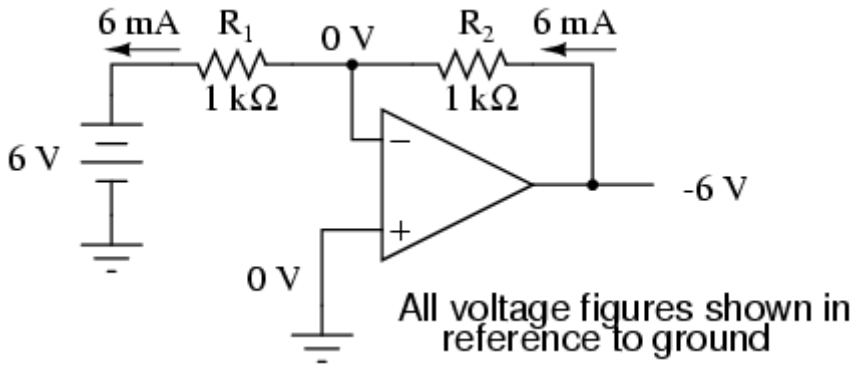
$$A_V = \frac{R_2}{R_1} + 1$$

Note that the voltage gain for this design of amplifier circuit can never be less than 1. If we were to lower R_2 to a value of zero ohms, our circuit would be essentially identical to the voltage follower, with the output directly connected to the inverting input. Since the voltage follower has a gain of 1, this sets the lower gain limit of the noninverting amplifier. However, the gain can be increased far beyond 1, by increasing R_2 in proportion to R_1.

Also note that the polarity of the output matches that of the input, just as with a voltage follower. A positive input voltage results in a positive output voltage, and vice versa (with respect to ground). For this reason, this circuit is referred to as a *noninverting amplifier*.

Just as with the voltage follower, we see that the differential gain of the op-amp is irrelevant, so long as it's very high. The voltages and currents in this circuit would hardly change at all if the op-amp's voltage gain were 250,000 instead of 200,000. This stands as a stark contrast to single-transistor amplifier circuit designs, where the Beta of the individual transistor greatly influenced the overall gains of the amplifier. With negative feedback, we have a self-correcting system that amplifies voltage according to the ratios set by the feedback resistors, not the gains internal to the op-amp.

Let's see what happens if we retain negative feedback through a voltage divider, but apply the input voltage at a different location:



By grounding the noninverting input, the negative feedback from the output seeks to hold the inverting input's voltage at 0 volts, as well. For this reason, the inverting input is referred to in this circuit as a *virtual ground*, being held at ground potential (0 volts) by the feedback, yet not directly connected to (electrically common with) ground. The input voltage this time is applied to the left-hand end of the voltage divider ($R_1 = R_2 = 1$ kΩ again), so the output voltage must swing to -6 volts in order to balance the middle at ground potential (0 volts). Using the same techniques as with the noninverting amplifier, we can analyze this circuit's operation by determining current magnitudes and directions, starting with R_1, and continuing on to determining the output voltage.

We can change the overall voltage gain of this circuit, overall, just by adjusting the values of R_1 and R_2 (changing the ratio of output voltage that is fed back to the inverting input). Gain can be calculated by the following formula:

$$A_V = \frac{R_2}{R_1}$$

Note that this circuit's voltage gain *can* be less than 1, depending solely on the ratio of R_2 to R_1. Also note that the output voltage is always the opposite polarity of the input voltage. A positive input voltage results in a negative output voltage, and vice versa (with respect to ground). For this reason, this circuit is referred to as an *inverting amplifier*. Sometimes, the gain formula contains a negative sign (before the R_2/R_1 fraction) to reflect this reversal of polarities.

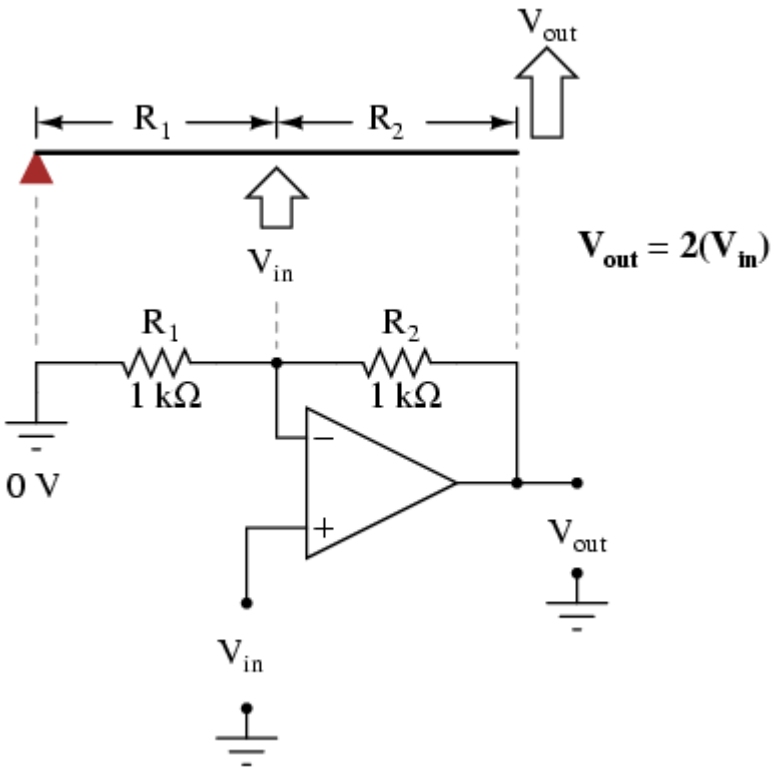
These two amplifier circuits we've just investigated serve the purpose of multiplying or dividing the magnitude of the input voltage signal. This is exactly how the mathematical operations of multiplication and division are typically handled in analog computer circuitry.

- **REVIEW:**
- By connecting the inverting (-) input of an op-amp directly to the output, we get negative feedback, which gives us a *voltage follower* circuit. By connecting that negative feedback through a resistive voltage divider (feeding back a *fraction* of the output voltage to the inverting input), the output voltage becomes a *multiple* of the input voltage.
- A negative-feedback op-amp circuit with the input signal going to the noninverting (+) input is called a *noninverting amplifier*. The output voltage will be the same polarity as the input. Voltage gain is given by the following equation: $A_V = (R_2/R_1) + 1$
- A negative-feedback op-amp circuit with the input signal going to the "bottom" of the resistive voltage divider, with the noninverting (+) input grounded, is called an *inverting amplifier*. Its output voltage will be the opposite polarity of the input. Voltage gain is given by the following equation: $A_V = R_2/R_1$

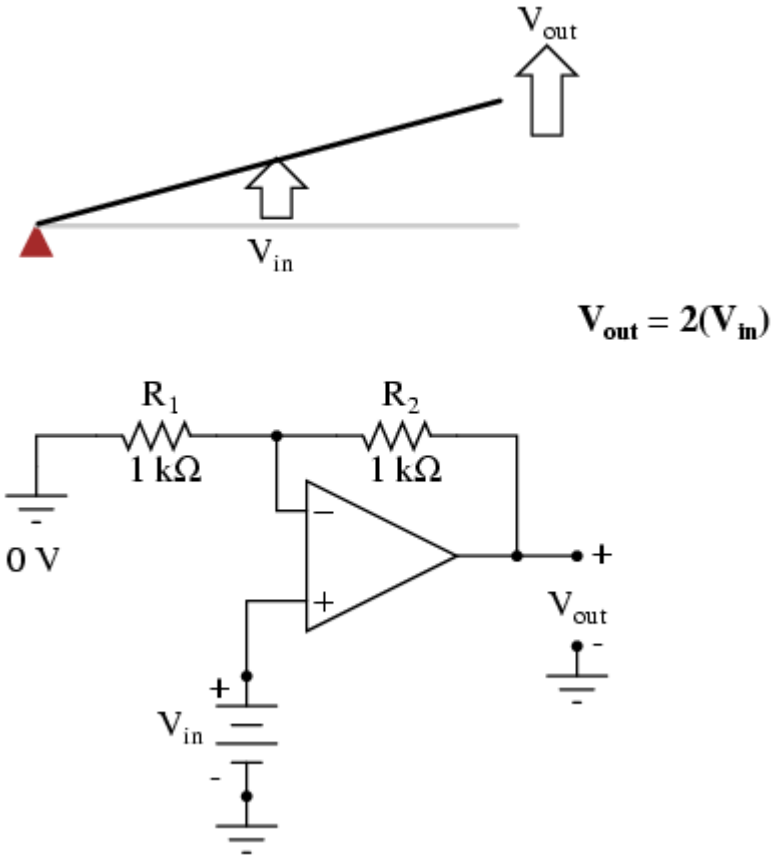
An analogy for divided feedback

A helpful analogy for understanding divided feedback amplifier circuits is that of a mechanical lever, with relative motion of the lever's ends representing change in input and output voltages, and the fulcrum (pivot point) representing the location of the ground point, real or virtual.

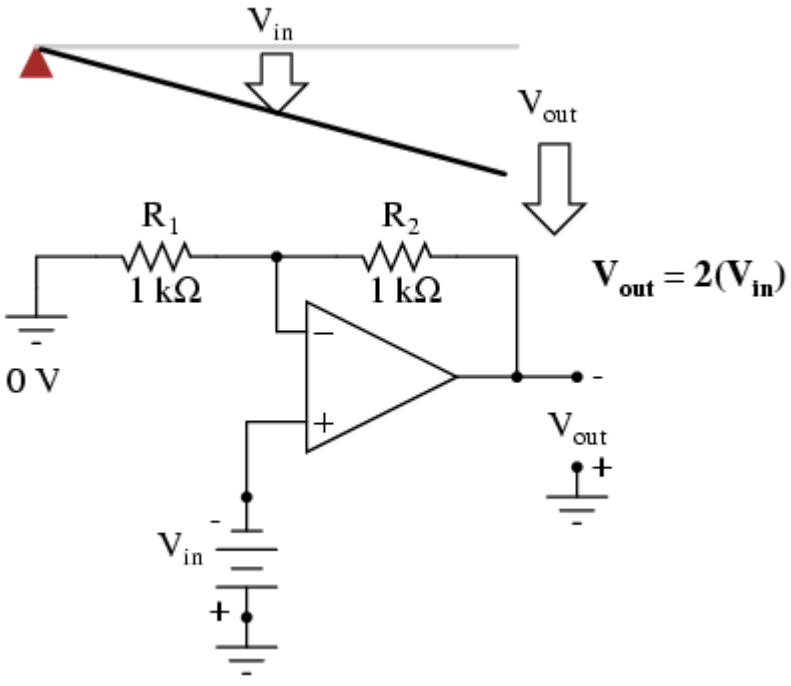
Take for example the following noninverting op-amp circuit. We know from the prior section that the voltage gain of a noninverting amplifier configuration can never be less than unity (1). If we draw a lever diagram next to the amplifier schematic, with the distance between fulcrum and lever ends representative of resistor values, the motion of the lever will signify changes in voltage at the input and output terminals of the amplifier:



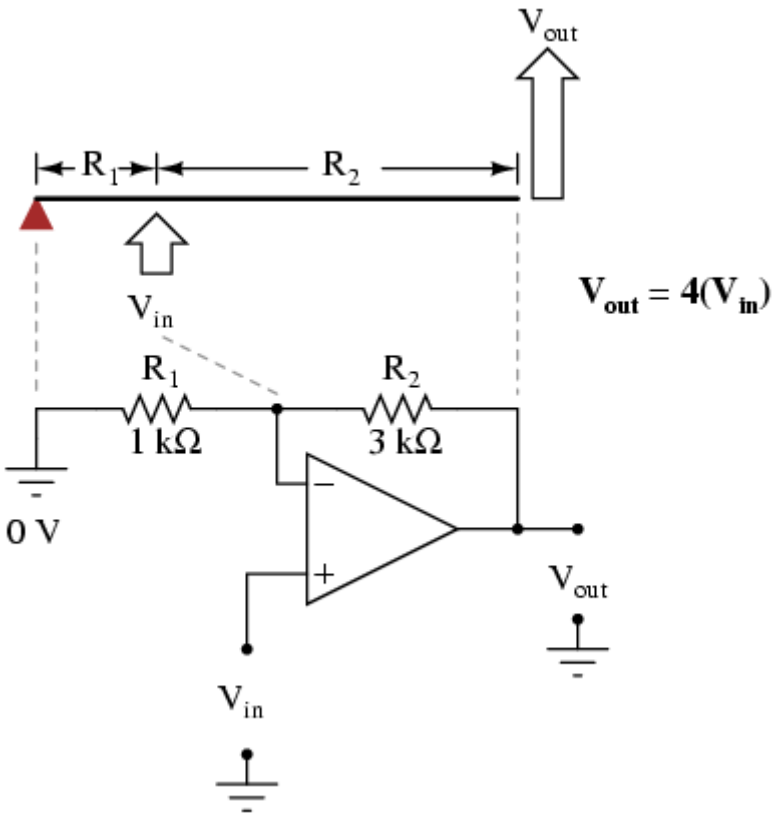
Physicists call this type of lever, with the input force (effort) applied between the fulcrum and output (load), a *third-class* lever. It is characterized by an output displacement (motion) at least as large than the input displacement -- a "gain" of at least 1 -- and in the same direction. Applying a positive input voltage to this op-amp circuit is analogous to displacing the "input" point on the lever upward:



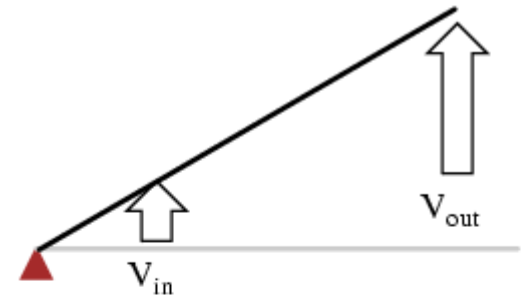
Due to the displacement-amplifying characteristics of the lever, the "output" point will move twice as far as the "input" point, and in the same direction. In the electronic circuit, the output voltage will equal twice the input, with the same polarity. Applying a negative input voltage is analogous to moving the lever downward from its level "zero" position, resulting in an amplified output displacement that is also negative:



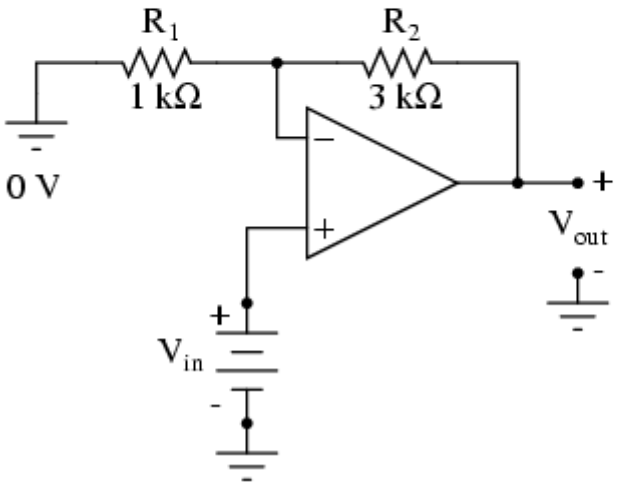
If we alter the resistor ratio R_2/R_1, we change the gain of the op-amp circuit. In lever terms, this means moving the input point in relation to the fulcrum and lever end, which similarly changes the displacement "gain" of the machine:



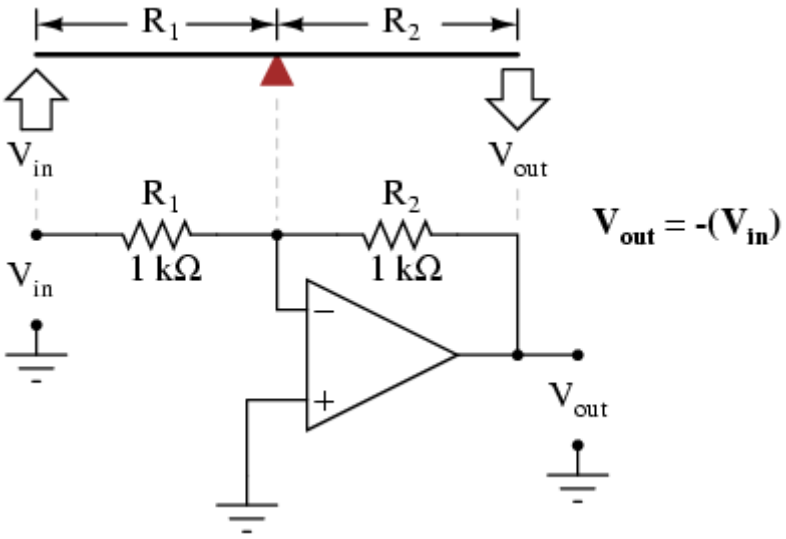
Now, any input signal will become amplified by a factor of four instead of by a factor of two:



$$\mathbf{V_{out} = 4(V_{in})}$$

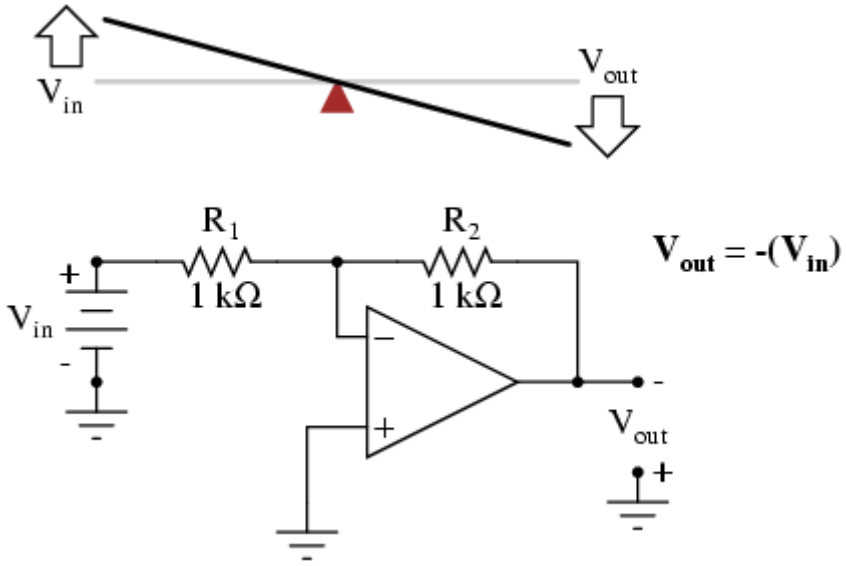


Inverting op-amp circuits may be modeled using the lever analogy as well. With the inverting configuration, the ground point of the feedback voltage divider is the op-amp's inverting input with the input to the left and the output to the right. This is mechanically equivalent to a *first-class* lever, where the input force (effort) is on the opposite side of the fulcrum from the output (load):

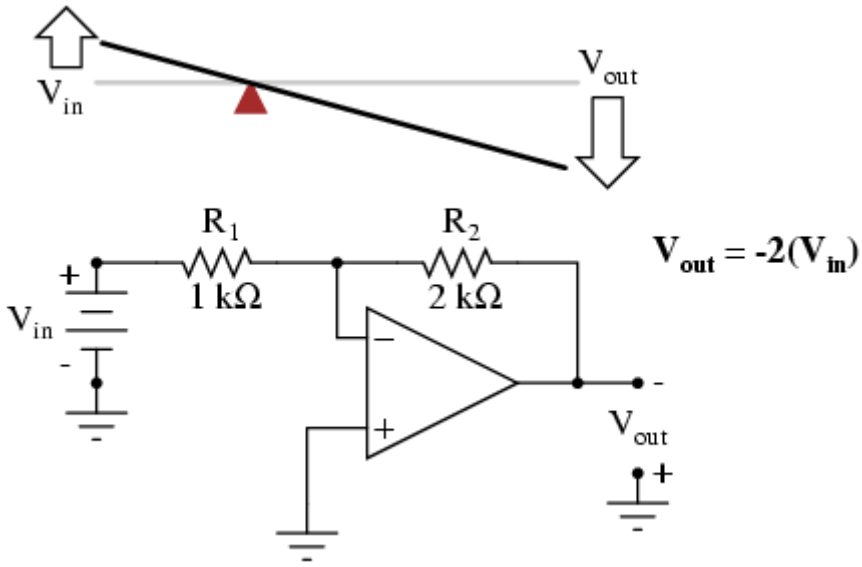


$$\mathbf{V_{out} = -(V_{in})}$$

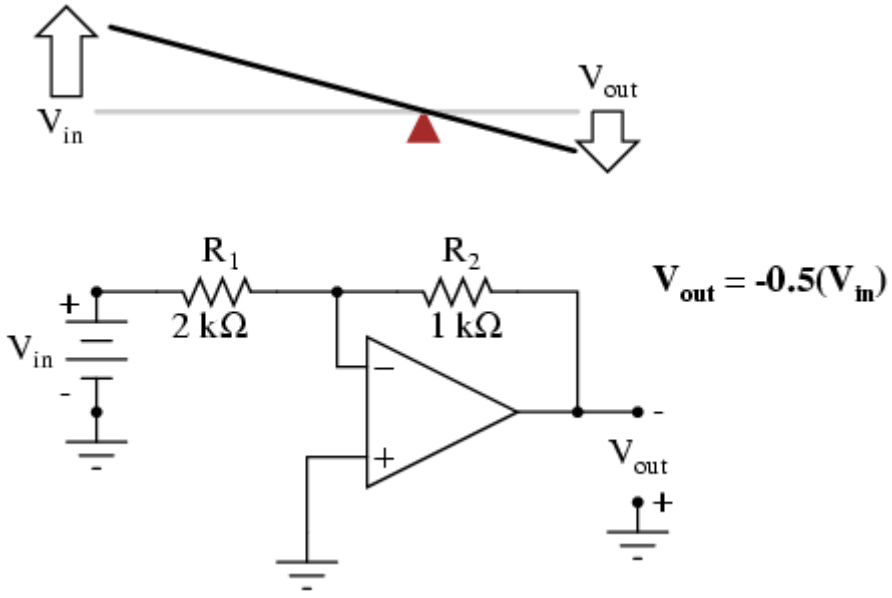
With equal-value resistors (equal-lengths of lever on each side of the fulcrum), the output voltage (displacement) will be equal in magnitude to the input voltage (displacement), but of the opposite polarity (direction). A positive input results in a negative output:



Changing the resistor ratio R_2/R_1 changes the gain of the amplifier circuit, just as changing the fulcrum position on the lever changes its mechanical displacement "gain." Consider the following example, where R_2 is made twice as large as R_1:



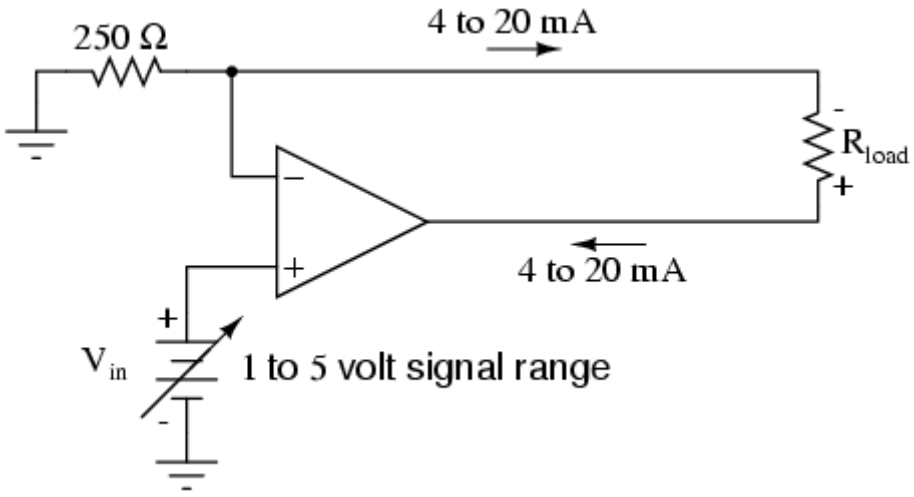
With the inverting amplifier configuration, though, gains of less than 1 are possible, just as with first-class levers. Reversing R_2 and R_1 values is analogous to moving the fulcrum to its complementary position on the lever: one-third of the way from the output end. There, the output displacement will be one-half the input displacement:



Voltage-to-current signal conversion

In instrumentation circuitry, DC signals are often used as analog representations of physical measurements such as temperature, pressure, flow, weight, and motion. Most commonly, *DC current* signals are used in preference to *DC voltage* signals, because current signals are exactly equal in magnitude throughout the series circuit loop carrying current from the source (measuring device) to the load (indicator, recorder, or controller), whereas voltage signals in a parallel circuit may vary from one end to the other due to resistive wire losses. Furthermore, current-sensing instruments typically have low impedances (while voltage-sensing instruments have high impedances), which gives current-sensing instruments greater electrical noise immunity.

In order to use current as an analog representation of a physical quantity, we have to have some way of generating a precise amount of current within the signal circuit. But how do we generate a precise current signal when we might not know the resistance of the loop? The answer is to use an amplifier designed to hold current to a prescribed value, applying as much or as little voltage as necessary to the load circuit to maintain that value. Such an amplifier performs the function of a *current source*. An op-amp with negative feedback is a perfect candidate for such a task:



The input voltage to this circuit is assumed to be coming from some type of physical transducer/amplifier arrangement, calibrated to produce 1 volt at 0 percent of physical measurement, and 5 volts at 100 percent of physical measurement. The standard analog current signal range is 4 mA to 20 mA, signifying 0% to 100% of measurement range, respectively. At 5 volts input, the 250 Ω (precision) resistor will have 5 volts applied across it, resulting in 20 mA of current in the large loop circuit (with R_{load}). It does not matter what resistance value R_{load} is, or how much wire resistance is present in that large loop, so long as the op-amp has a high enough power supply voltage to output the voltage necessary to get 20 mA flowing through R_{load}. The 250 Ω resistor establishes the relationship between input voltage and output current, in this case creating the equivalence of 1-5 V in / 4-20 mA out. If we were converting the 1-5 volt input signal to a 10-50 mA output signal (an older, obsolete instrumentation standard for industry), we'd use a 100 Ω precision resistor instead.

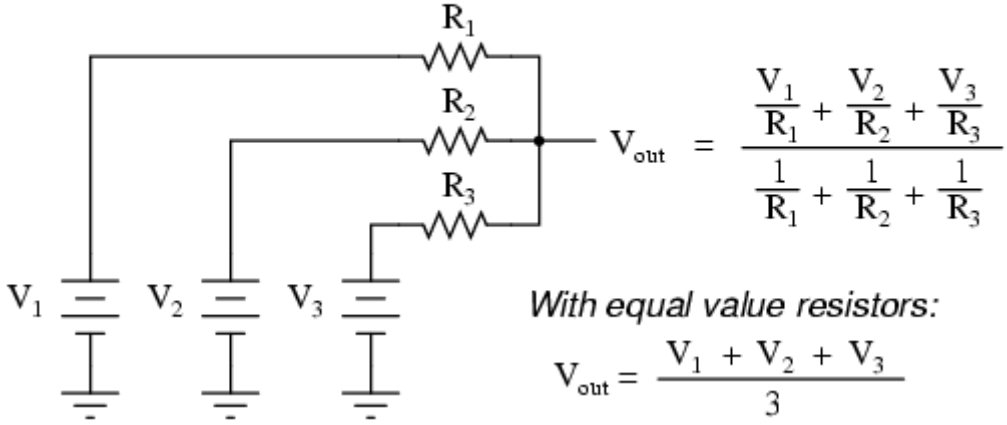
Another name for this circuit is *transconductance amplifier*. In electronics, transconductance is the mathematical ratio of current change divided by voltage change ($\Delta I / \Delta V$), and it is measured in the unit of Siemens, the same unit used to express conductance (the mathematical reciprocal of resistance: current/voltage). In this circuit, the transconductance ratio is fixed by the value of the 250 Ω resistor, giving a linear current-out/voltage-in relationship.

- **REVIEW:**
- In industry, DC current signals are often used in preference to DC voltage signals as analog representations of physical quantities. Current in a series circuit is absolutely equal at all points in that circuit regardless of wiring resistance, whereas voltage in a parallel-connected circuit may vary from end to end because of wire resistance, making current-signaling more accurate from the "transmitting" to the "receiving" instrument.
- Voltage signals are relatively easy to produce directly from transducer devices, whereas accurate current signals are not. Op-amps can be used to "convert" a voltage signal into a current signal quite easily. In this mode, the op-amp will output whatever voltage is necessary to maintain current through the signaling circuit at the proper value.

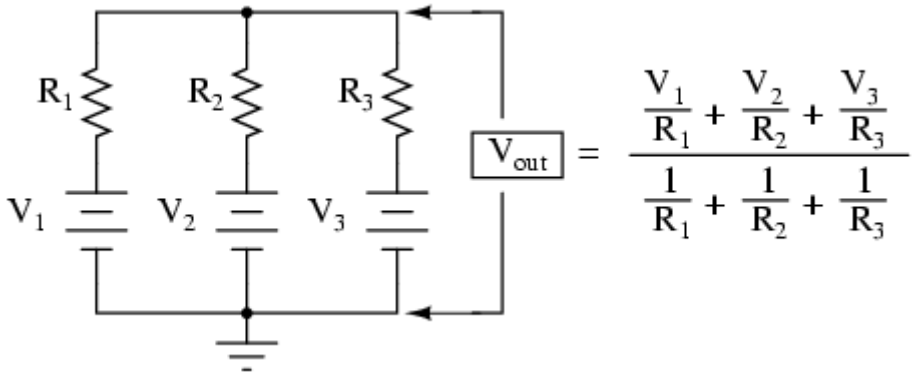
Averager and summer circuits

If we take three equal resistors and connect one end of each to a common point, then apply three input voltages (one to each of the resistors' free ends), the voltage seen at the common point will be the mathematical *average* of the three.

"Passive averager" circuit



This circuit is really nothing more than a practical application of Millman's Theorem:



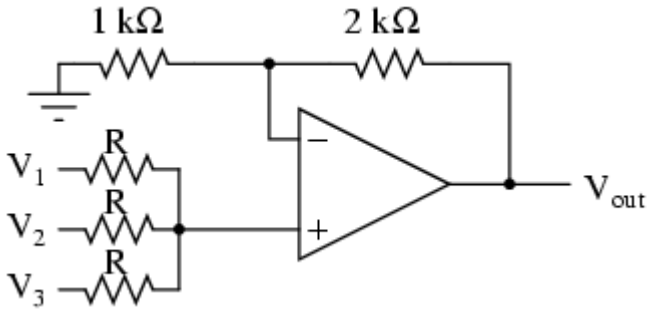
This circuit is commonly known as a *passive averager*, because it generates an average voltage with non-amplifying components. *Passive* simply means that it is an unamplified circuit. The large equation to the right of the averager circuit comes from Millman's Theorem, which describes the voltage produced by multiple voltage sources connected together through individual resistances. Since the three resistors in the averager circuit are equal to each other, we can simplify Millman's formula by writing R_1, R_2, and R_3 simply as R (one, equal resistance instead of three individual resistances):

$$V_{out} = \frac{\frac{V_1}{R} + \frac{V_2}{R} + \frac{V_3}{R}}{\frac{1}{R} + \frac{1}{R} + \frac{1}{R}}$$

$$V_{out} = \frac{\frac{V_1 + V_2 + V_3}{R}}{\frac{3}{R}}$$

$$V_{out} = \frac{V_1 + V_2 + V_3}{3}$$

If we take a passive averager and use it to connect three input voltages into an op-amp amplifier circuit with a gain of 3, we can turn this *averaging* function into an *addition* function. The result is called a *noninverting summer* circuit:

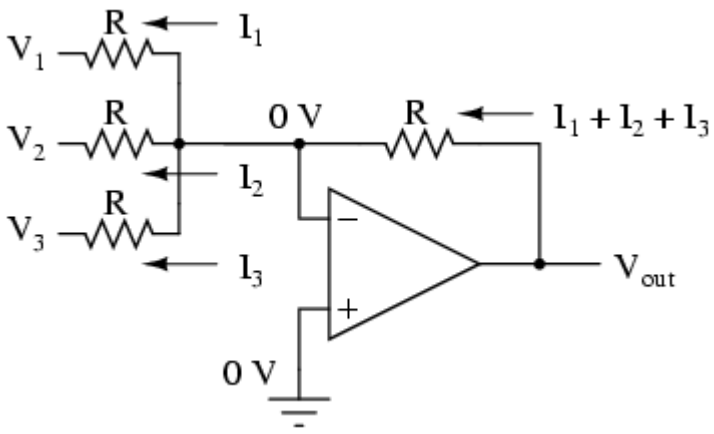


With a voltage divider composed of a 2 kΩ / 1 kΩ combination, the noninverting amplifier circuit will have a voltage gain of 3. By taking the voltage from the passive averager, which is the sum of V_1, V_2, and V_3 divided by 3, and multiplying that average by 3, we arrive at an output voltage equal to the *sum* of V_1, V_2, and V_3:

$$V_{out} = 3 \frac{V_1 + V_2 + V_3}{3}$$

$$V_{out} = V_1 + V_2 + V_3$$

Much the same can be done with an inverting op-amp amplifier, using a passive averager as part of the voltage divider feedback circuit. The result is called an *inverting summer* circuit:



Now, with the right-hand sides of the three averaging resistors connected to the virtual ground point of the op-amp's inverting input, Millman's Theorem no longer directly applies as it did before. The voltage at the virtual ground is now held at 0 volts by the op-amp's negative feedback, whereas before it was free to float to the average value of V_1, V_2, and V_3. However, with all resistor values equal to each other, the currents through each of the three resistors will be proportional to their respective input voltages. Since those three currents will *add* at the virtual ground node, the algebraic sum of those currents through the feedback resistor will produce a voltage at V_{out} equal to $V_1 + V_2 + V_3$, except with reversed polarity. The reversal in polarity is what makes this circuit an *inverting* summer:

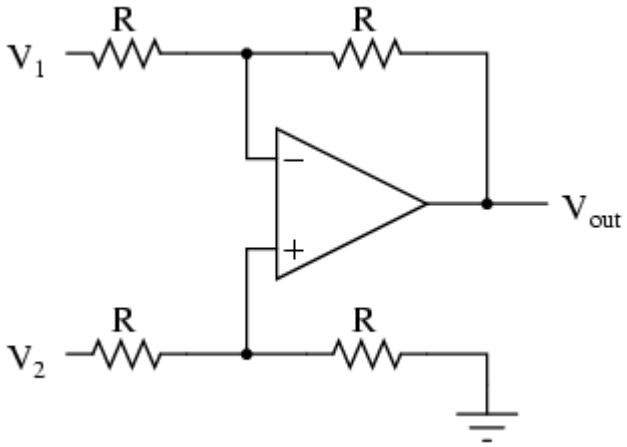
$$V_{out} = -(V_1 + V_2 + V_3)$$

Summer (adder) circuits are quite useful in analog computer design, just as multiplier and divider circuits would be. Again, it is the extremely high differential gain of the op-amp which allows us to build these useful circuits with a bare minimum of components.

- **REVIEW:**
- A *summer* circuit is one that *sums*, or adds, multiple analog voltage signals together. There are two basic varieties of op-amp summer circuits: noninverting and inverting.

Building a differential amplifier

An op-amp with no feedback is already a differential amplifier, amplifying the voltage difference between the two inputs. However, its gain cannot be controlled, and it is generally too high to be of any practical use. So far, our application of negative feedback to op-amps has resulting in the practical loss of one of the inputs, the resulting amplifier only good for amplifying a single voltage signal input. With a little ingenuity, however, we can construct an op-amp circuit maintaining both voltage inputs, yet with a controlled gain set by external resistors.

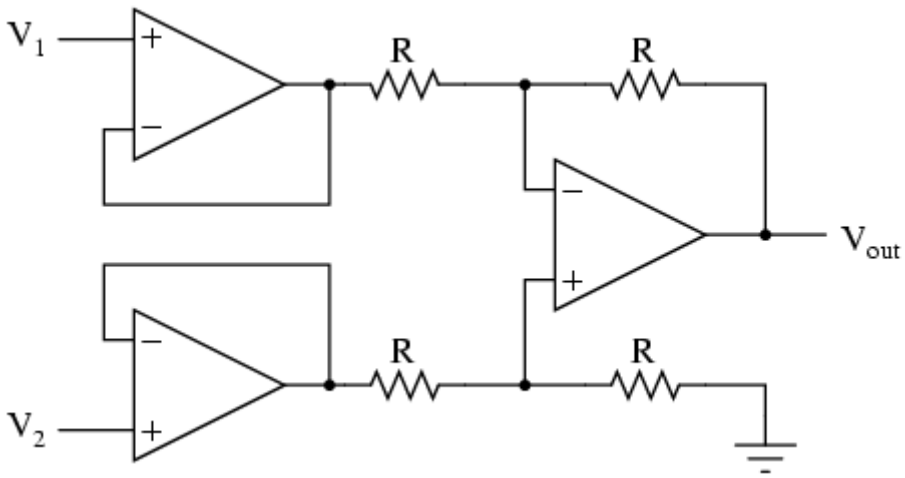


If all the resistor values are equal, this amplifier will have a differential voltage gain of 1. The analysis of this circuit is essentially the same as that of an inverting amplifier, except that the noninverting input (+) of the op-amp is at a voltage equal to a fraction of V_2, rather than being connected directly to ground. As would stand to reason, V_2 functions as the noninverting input and V_1 functions as the inverting input of the final amplifier circuit. Therefore:

$$V_{out} = V_2 - V_1$$

If we wanted to provide a differential gain of anything other than 1, we would have to adjust the resistances in *both* upper and lower voltage dividers, necessitating multiple resistor changes and balancing between the two dividers for symmetrical operation. This is not always practical, for obvious reasons.

Another limitation of this amplifier design is the fact that its input impedances are rather low compared to that of some other op-amp configurations, most notably the noninverting (single-ended input) amplifier. Each input voltage source has to drive current through a resistance, which constitutes far less impedance than the bare input of an op-amp alone. The solution to this problem, fortunately, is quite simple. All we need to do is "buffer" each input voltage signal through a voltage follower like this:

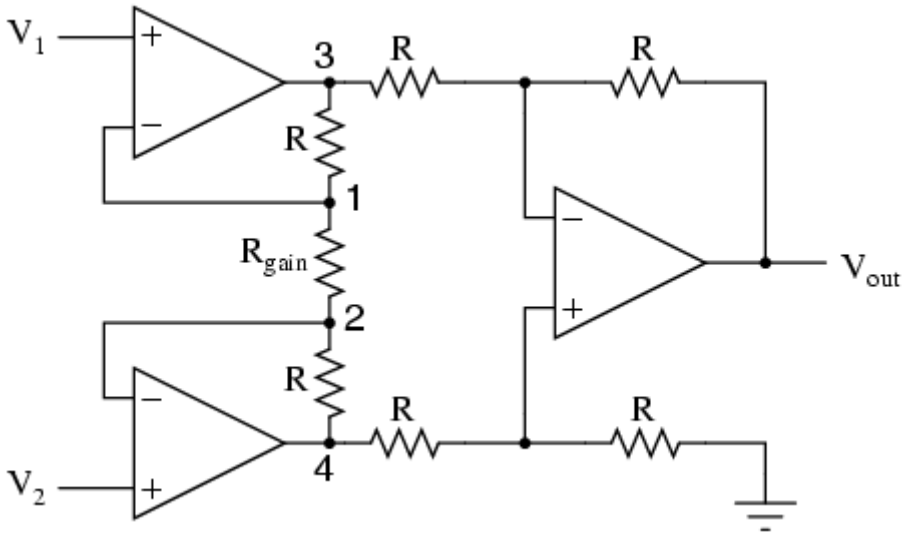


Now the V_1 and V_2 input lines are connected straight to the inputs of two voltage-follower op-amps, giving very high impedance. The two op-amps on the left now handle the driving of

current through the resistors instead of letting the input voltage sources (whatever they may be) do it. The increased complexity to our circuit is minimal for a substantial benefit.

The instrumentation amplifier

As suggested before, it is beneficial to be able to adjust the gain of the amplifier circuit without having to change more than one resistor value, as is necessary with the previous design of differential amplifier. The so-called *instrumentation* builds on the last version of differential amplifier to give us that capability:



This intimidating circuit is constructed from a buffered differential amplifier stage with three new resistors linking the two buffer circuits together. Consider all resistors to be of equal value except for R_{gain}. The negative feedback of the upper-left op-amp causes the voltage at point 1 (top of R_{gain}) to be equal to V_1. Likewise, the voltage at point 2 (bottom of R_{gain}) is held to a value equal to V_2. This establishes a voltage drop across R_{gain} equal to the voltage difference between V_1 and V_2. That voltage drop causes a current through R_{gain}, and since the feedback loops of the two input op-amps draw no current, that same amount of current through R_{gain} must be going through the two "R" resistors above and below it. This produces a voltage drop between points 3 and 4 equal to:

$$V_{3\text{-}4} = (V_2 - V_1)\left(1 + \frac{2R}{R_{gain}}\right)$$

The regular differential amplifier on the right-hand side of the circuit then takes this voltage drop between points 3 and 4, and amplifies it by a gain of 1 (assuming again that all "R" resistors are of equal value). Though this looks like a cumbersome way to build a differential amplifier, it has the distinct advantages of possessing extremely high input impedances on the V_1 and V_2 inputs (because they connect straight into the noninverting inputs of their respective op-amps), and adjustable gain that can be set by a single resistor. Manipulating the above formula a bit, we have a general expression for overall voltage gain in the instrumentation amplifier:

$$A_V = (1 + \frac{2R}{R_{gain}})$$

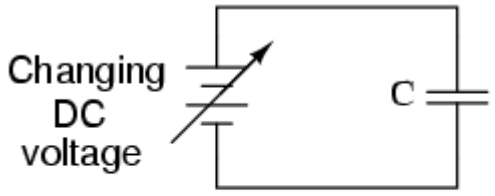
Though it may not be obvious by looking at the schematic, we can change the differential gain of the instrumentation amplifier simply by changing the value of one resistor: R_{gain}. Yes, we could still change the overall gain by changing the values of some of the other resistors, but this would necessitate *balanced* resistor value changes for the circuit to remain symmetrical. Please note that the lowest gain possible with the above circuit is obtained with R_{gain} completely open (infinite resistance), and that gain value is 1.

- **REVIEW:**
- An *instrumentation amplifier* is a differential op-amp circuit providing high input impedances with ease of gain adjustment through the variation of a single resistor.

Differentiator and integrator circuits

By introducing electrical reactance into the feedback loops of op-amp amplifier circuits, we can cause the output to respond to changes in the input voltage over *time*. Drawing their names from their respective calculus functions, the *integrator* produces a voltage output proportional to the product (multiplication) of the input voltage and time; and the *differentiator* (not to be confused with *differential*) produces a voltage output proportional to the input voltage's rate of change.

Capacitance can be defined as the measure of a capacitor's opposition to changes in voltage. The greater the capacitance, the more the opposition. Capacitors oppose voltage change by creating current in the circuit: that is, they either charge or discharge in response to a change in applied voltage. So, the more capacitance a capacitor has, the greater its charge or discharge current will be for any given rate of voltage change across it. The equation for this is quite simple:



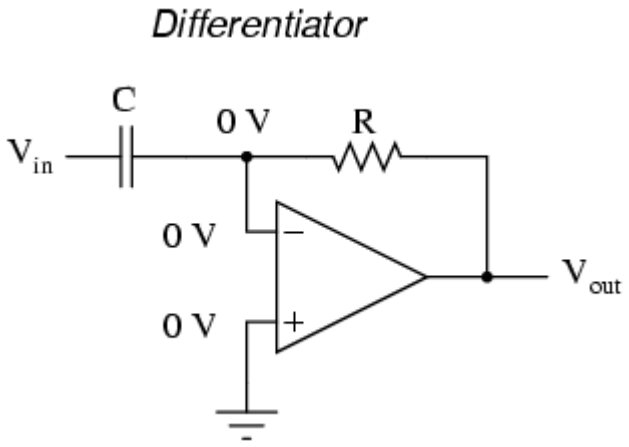
$$i = C \frac{dv}{dt}$$

The *dv/dt* fraction is a calculus expression representing the rate of voltage change over time. If the DC supply in the above circuit were steadily increased from a voltage of 15 volts to a voltage of 16 volts over a time span of 1 hour, the current through the capacitor would most likely be *very* small, because of the very low rate of voltage change (dv/dt = 1 volt / 3600 seconds). However, if we steadily increased the DC supply from 15 volts to 16 volts over a shorter time span of 1 second, the rate of voltage change would be much higher, and thus the charging current would be much higher (3600 times higher, to be exact). Same amount of

change in voltage, but vastly different *rates* of change, resulting in vastly different amounts of current in the circuit.

To put some definite numbers to this formula, if the voltage across a 47 µF capacitor was changing at a linear rate of 3 volts per second, the current "through" the capacitor would be (47 µF)(3 V/s) = 141 µA.

We can build an op-amp circuit which measures change in voltage by measuring current through a capacitor, and outputs a voltage proportional to that current:



The right-hand side of the capacitor is held to a voltage of 0 volts, due to the "virtual ground" effect. Therefore, current "through" the capacitor is solely due to *change* in the input voltage. A steady input voltage won't cause a current through C, but a *changing* input voltage will.

Capacitor current moves through the feedback resistor, producing a drop across it, which is the same as the output voltage. A linear, positive rate of input voltage change will result in a steady negative voltage at the output of the op-amp. Conversely, a linear, negative rate of input voltage change will result in a steady positive voltage at the output of the op-amp. This polarity inversion from input to output is due to the fact that the input signal is being sent (essentially) to the inverting input of the op-amp, so it acts like the inverting amplifier mentioned previously. The faster the rate of voltage change at the input (either positive or negative), the greater the voltage at the output.

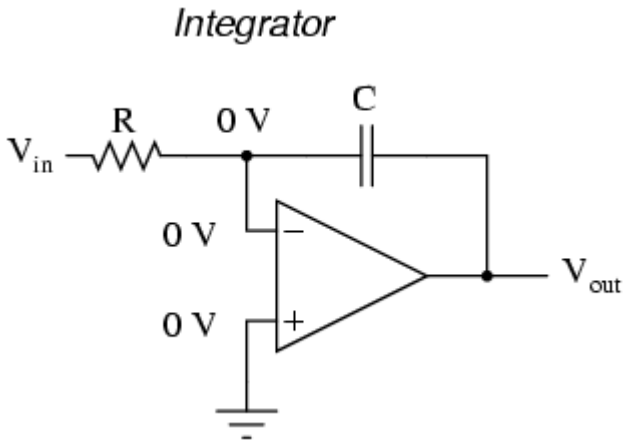
The formula for determining voltage output for the differentiator is as follows:

$$V_{out} = -RC \frac{dv_{in}}{dt}$$

Applications for this, besides representing the derivative calculus function inside of an analog computer, include rate-of-change indicators for process instrumentation. One such rate-of-change signal application might be for monitoring (or controlling) the rate of temperature change in a furnace, where too high or too low of a temperature rise rate could be detrimental. The DC voltage produced by the differentiator circuit could be used to drive a comparator, which would signal an alarm or activate a control if the rate of change exceeded a pre-set level.

In process control, the derivative function is used to make control decisions for maintaining a process at setpoint, by monitoring the rate of process change over time and taking action to prevent excessive rates of change, which can lead to an unstable condition. Analog electronic controllers use variations of this circuitry to perform the derivative function.

On the other hand, there are applications where we need precisely the opposite function, called *integration* in calculus. Here, the op-amp circuit would generate an output voltage proportional to the magnitude and duration that an input voltage signal has deviated from 0 volts. Stated differently, a constant input signal would generate a certain *rate of change* in the output voltage: differentiation in reverse. To do this, all we have to do is swap the capacitor and resistor in the previous circuit:



As before, the negative feedback of the op-amp ensures that the inverting input will be held at 0 volts (the virtual ground). If the input voltage is exactly 0 volts, there will be no current through the resistor, therefore no charging of the capacitor, and therefore the output voltage will not change. We cannot guarantee what voltage will be at the output with respect to ground in this condition, but we can say that the output voltage *will be constant*.

However, if we apply a constant, positive voltage to the input, the op-amp output will fall negative at a linear rate, in an attempt to produce the changing voltage across the capacitor necessary to maintain the current established by the voltage difference across the resistor. Conversely, a constant, negative voltage at the input results in a linear, rising (positive) voltage at the output. The output voltage rate-of-change will be proportional to the value of the input voltage.

The formula for determining voltage output for the integrator is as follows:

$$\frac{dv_{out}}{dt} = - \frac{V_{in}}{RC}$$

or

$$V_{out} = \int_0^t - \frac{V_{in}}{RC}\, dt + c$$

Where,

c = Output voltage at start time (t=0)

One application for this device would be to keep a "running total" of radiation exposure, or dosage, if the input voltage was a proportional signal supplied by an electronic radiation detector. Nuclear radiation can be just as damaging at low intensities for long periods of time as it is at high intensities for short periods of time. An integrator circuit would take both the intensity (input voltage magnitude) and time into account, generating an output voltage representing total radiation dosage.

Another application would be to integrate a signal representing water flow, producing a signal representing total quantity of water that has passed by the flowmeter. This application of an integrator is sometimes called a *totalizer* in the industrial instrumentation trade.

- **REVIEW:**
- A *differentiator* circuit produces a constant output voltage for a steadily changing input voltage.
- An *integrator* circuit produces a steadily changing output voltage for a constant input voltage.
- Both types of devices are easily constructed, using reactive components (usually capacitors rather than inductors) in the feedback part of the circuit.

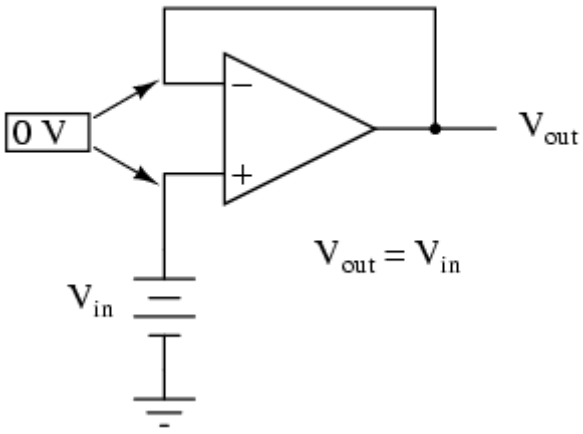
Positive feedback

As we've seen, negative feedback is an incredibly useful principle when applied to operational amplifiers. It is what allows us to create all these practical circuits, being able to precisely set gains, rates, and other significant parameters with just a few changes of resistor values. Negative feedback makes all these circuits stable and self-correcting.

The basic principle of negative feedback is that the output tends to drive in a direction that creates a condition of equilibrium (balance). In an op-amp circuit with no feedback, there is no corrective mechanism, and the output voltage will saturate with the tiniest amount of differential voltage applied between the inputs. The result is a comparator:

With negative feedback (the output voltage "fed back" somehow to the inverting input), the circuit tends to prevent itself from driving the output to full saturation. Rather, the output voltage drives only as high or as low as needed to balance the two inputs' voltages:

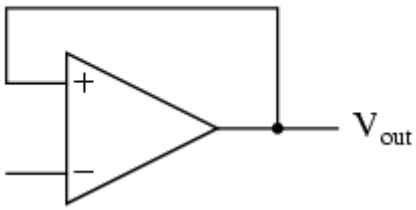
Negative feedback



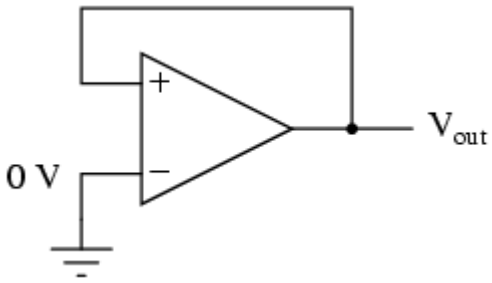
Whether the output is directly fed back to the inverting (-) input or coupled through a set of components, the effect is the same: the extremely high differential voltage gain of the op-amp will be "tamed" and the circuit will respond according to the dictates of the feedback "loop" connecting output to inverting input.

Another type of feedback, namely *positive feedback*, also finds application in op-amp circuits. Unlike negative feedback, where the output voltage is "fed back" to the inverting (-) input, with positive feedback the output voltage is somehow routed back to the noninverting (+) input. In its simplest form, we could connect a straight piece of wire from output to noninverting input and see what happens:

Positive feedback



The inverting input remains disconnected from the feedback loop, and is free to receive an external voltage. Let's see what happens if we ground the inverting input:



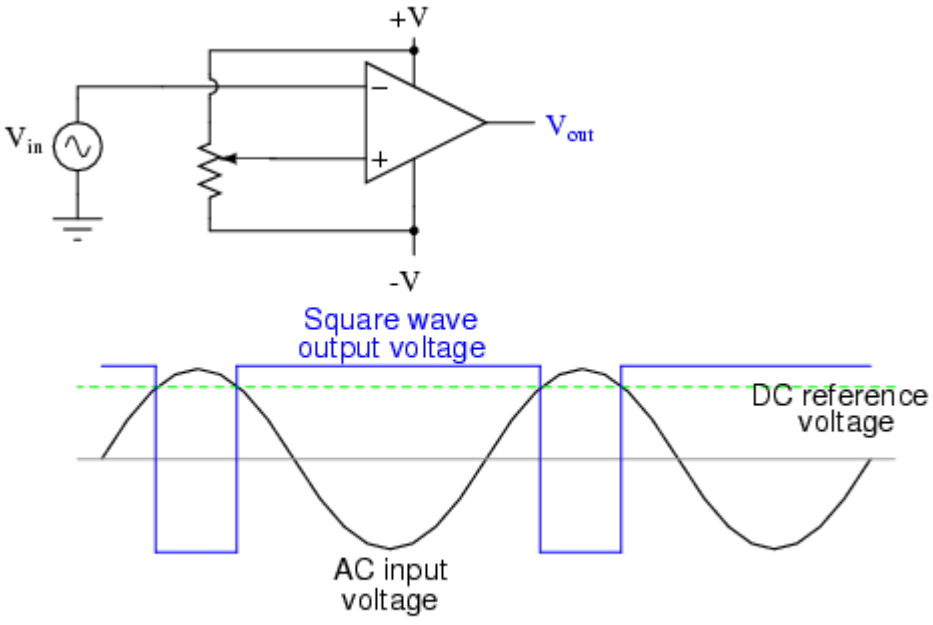
With the inverting input grounded (maintained at zero volts), the output voltage will be dictated by the magnitude and polarity of the voltage at the noninverting input. If that voltage

happens to be positive, the op-amp will drive its output positive as well, feeding that positive voltage back to the noninverting input, which will result in full positive output saturation. On the other hand, if the voltage on the noninverting input happens to start out negative, the op-amp's output will drive in the negative direction, feeding back to the noninverting input and resulting in full negative saturation.

What we have here is a circuit whose output is *bistable*: stable in one of two states (saturated positive or saturated negative). Once it has reached one of those saturated states, it will tend to remain in that state, unchanging. What is necessary to get it to switch states is a voltage placed upon the inverting (-) input of the same polarity, but of a slightly greater magnitude. For example, if our circuit is saturated at an output voltage of +12 volts, it will take an input voltage at the inverting input of at least +12 volts to get the output to change. When it changes, it will saturate fully negative.

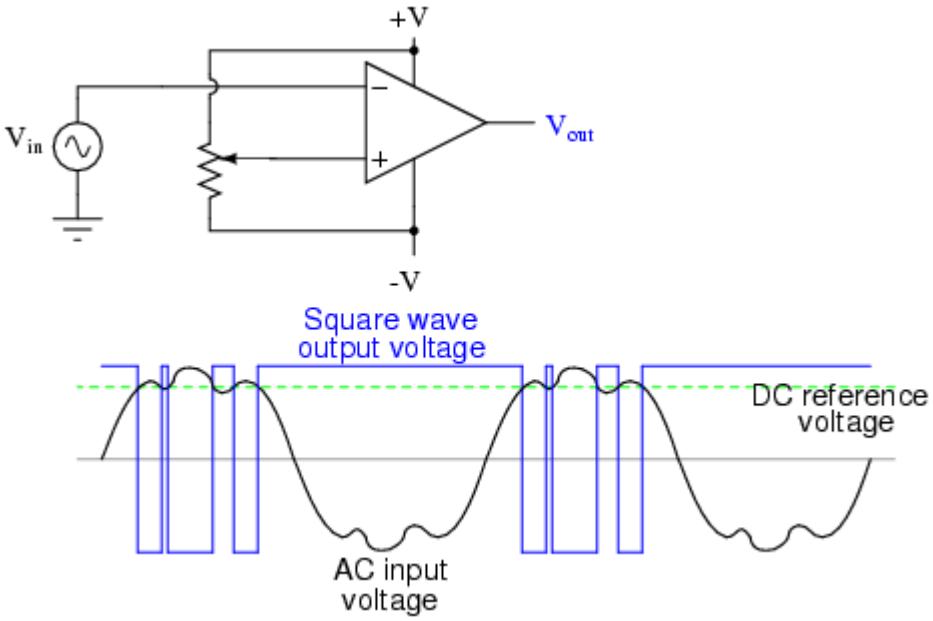
So, an op-amp with positive feedback tends to stay in whatever output state it's already in. It "latches" between one of two states, saturated positive or saturated negative. Technically, this is known as *hysteresis*.

Hysteresis can be a useful property for a comparator circuit to have. As we've seen before, comparators can be used to produce a square wave from any sort of ramping waveform (sine wave, triangle wave, sawtooth wave, etc.) input. If the incoming AC waveform is noise-free (that is, a "pure" waveform), a simple comparator will work just fine.



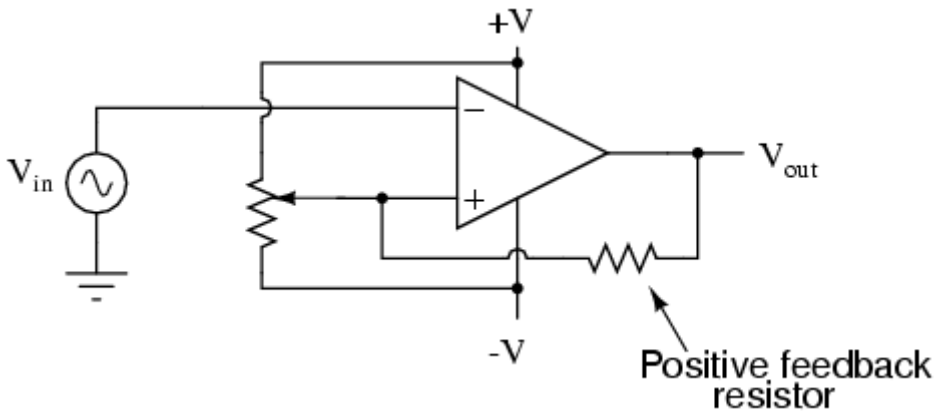
A "clean" AC input waveform produces predictable transition points on the output voltage square wave

However, if there exist any anomalies in the waveform such as harmonics or "spikes" which cause the voltage to rise and fall significantly within the timespan of a single cycle, a comparator's output might switch states unexpectedly:

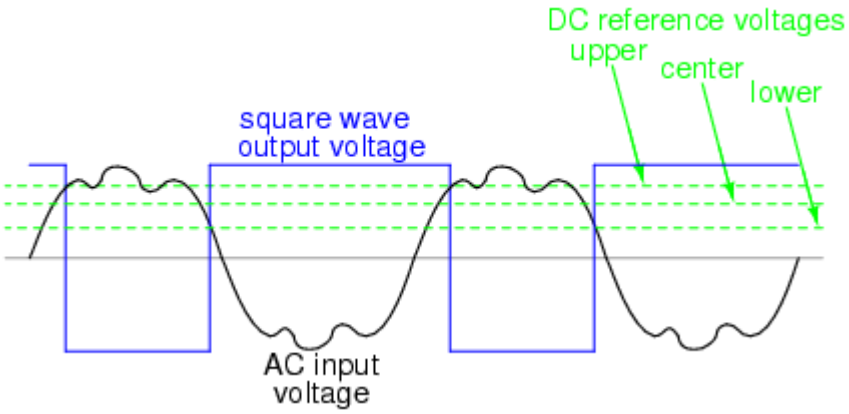


Any time there is a transition through the reference voltage level, no matter how tiny that transition may be, the output of the comparator will switch states, producing a square wave with "glitches."

If we add a little positive feedback to the comparator circuit, we will introduce hysteresis into the output. This hysteresis will cause the output to remain in its current state unless the AC input voltage undergoes a *major* change in magnitude.



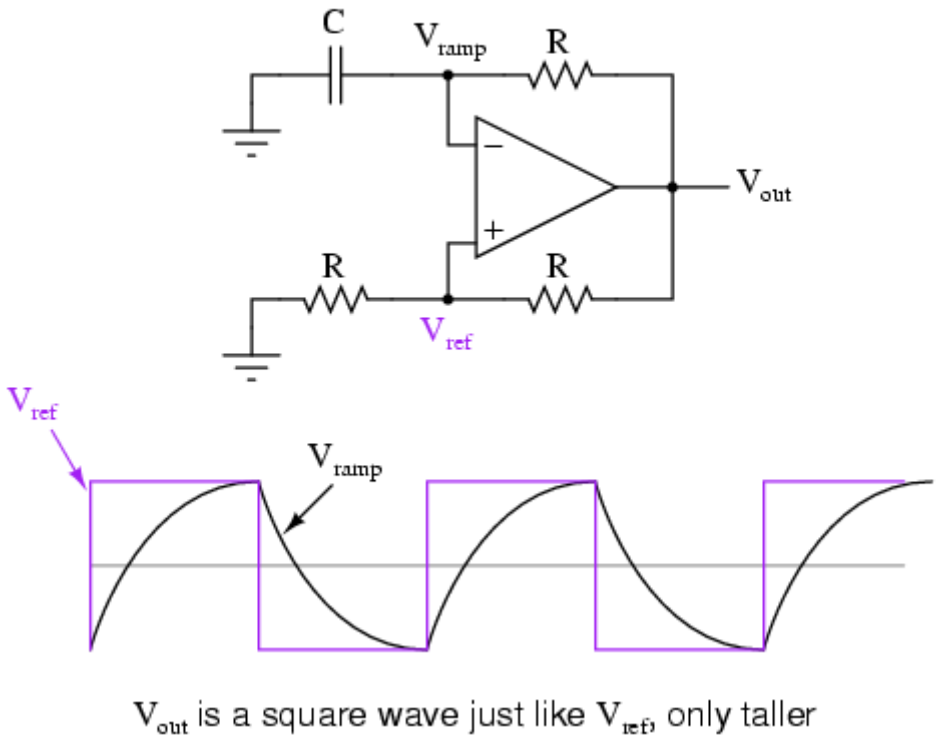
What this feedback resistor creates is a dual-reference for the comparator circuit. The voltage applied to the noninverting (+) input as a reference which to compare with the incoming AC voltage changes depending on the value of the op-amp's output voltage. When the op-amp output is saturated positive, the reference voltage at the noninverting input will be more positive than before. Conversely, when the op-amp output is saturated negative, the reference voltage at the noninverting input will be more negative than before. The result is easier to understand on a graph:



When the op-amp output is saturated positive, the upper reference voltage is in effect, and the output won't drop to a negative saturation level unless the AC input rises *above* that upper reference level. Conversely, when the op-amp output is saturated negative, the lower reference voltage is in effect, and the output won't rise to a positive saturation level unless the AC input drops *below* that lower reference level. The result is a clean square-wave output again, despite significant amounts of distortion in the AC input signal. In order for a "glitch" to cause the comparator to switch from one state to another, it would have to be at least as big (tall) as the difference between the upper and lower reference voltage levels, and at the right point in time to cross both those levels.

Another application of positive feedback in op-amp circuits is in the construction of oscillator circuits. An *oscillator* is a device that produces an alternating (AC), or at least pulsing, output voltage. Technically, it is known as an *astable* device: having no stable output state (no equilibrium whatsoever). Oscillators are very useful devices, and they are easily made with just an op-amp and a few external components.

Oscillator circuit using positive feedback



V_{out} is a square wave just like V_{ref}, only taller

When the output is saturated positive, the V_{ref} will be positive, and the capacitor will charge up in a positive direction. When V_{ramp} exceeds V_{ref} by the tiniest margin, the output will saturate negative, and the capacitor will charge in the opposite direction (polarity). Oscillation occurs because the positive feedback is instantaneous and the negative feedback is delayed (by means of an RC time constant). The frequency of this oscillator may be adjusted by varying the size of any component.

- **REVIEW:**
- Negative feedback creates a condition of *equilibrium* (balance). Positive feedback creates a condition of *hysteresis* (the tendency to "latch" in one of two extreme states).
- An *oscillator* is a device producing an alternating or pulsing output voltage.