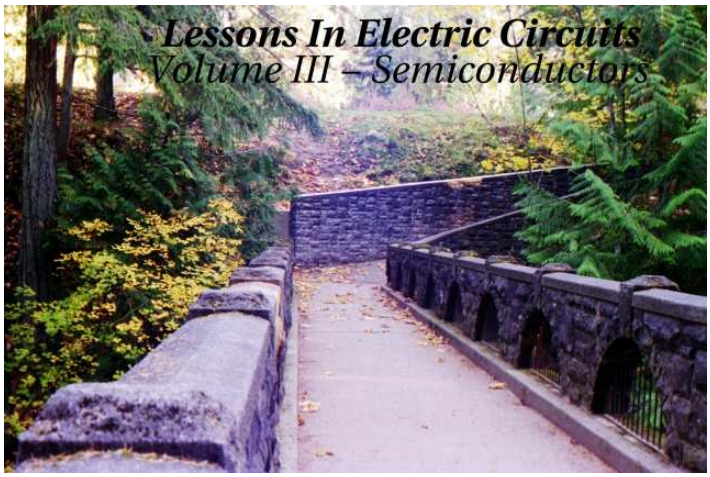


Fifth Edition, last update January 18, 2006

Lessons In Electric Circuits, Volume III – Semiconductors

By Tony R. Kuphaldt

Fourth Edition, last update January 18, 2006

PRINTING HISTORY

- First Edition: Printed in June of 2000. Plain-ASCII illustrations for universal computer readability.
- Second Edition: Printed in September of 2000. Illustrations reworked in standard graphic (eps and jpeg) format. Source files translated to *Texinfo* format for easy online and printed publication.
- Third Edition: Printed in January 2002. Source files translated to *SubML* format. SubML is a simple markup language designed to easily convert to other markups like LaTeX, HTML, or DocBook using nothing but search-and-replace substitutions.
- Fourth Edition: Printed in December 2002. New sections added, and error corrections made, since third edition.

Contents

Chapter 1

AMPLIFIERS AND ACTIVE DEVICES

Contents

1.1 From electric to electronic

This third volume of the book series *Lessons In Electric Circuits* makes a departure from the former two in that the transition between *electric* circuits and *electronic* circuits is formally crossed. Electric circuits are connections of conductive wires and other devices whereby the uniform flow of electrons occurs. Electronic circuits add a new dimension to electric circuits in that some means of *control* is exerted over the flow of electrons by another electrical signal, either a voltage or a current.

In and of itself, the control of electron flow is nothing new to the student of electric circuits. Switches control the flow of electrons, as do potentiometers, especially when connected as variable resistors (rheostats). Neither the switch nor the potentiometer should be new to your experience by this point in your study. The threshold marking the transition from electric to electronic, then, is defined by *how* the flow of electrons is controlled rather than whether or not any form of control exists in a circuit. Switches and rheostats control the flow of electrons according to the positioning of a mechanical device, which is actuated by some physical force external to the circuit. In electronics, however, we are dealing with special devices able to control the flow of electrons according to another flow of electrons, or by the application of a static voltage. In other words, in an electronic circuit, *electricity is able to control electricity.*

Historically, the era of electronics began with the invention of the *Audion tube*, a device controlling the flow of an electron stream through a vacuum by the application of a small voltage between two metal structures within the tube. A more detailed summary of so-called *electron tube* or *vacuum tube* technology is available in the last chapter of this volume for those who are interested.

Electronics technology experienced a revolution in 1948 with the invention of the *transistor*. This tiny device achieved approximately the same effect as the Audion tube, but in a vastly smaller amount of space and with less material. Transistors control the flow of electrons through solid *semiconductor* substances rather than through a vacuum, and so transistor technology is often referred to as *solid-state* electronics.

1.2 Active versus passive devices

An *active* device is any type of circuit component with the ability to electrically control electron flow (electricity controlling electricity). In order for a circuit to be properly called *electronic*, it must contain at least one active device. Components incapable of controlling current by means of another electrical signal are called *passive* devices. Resistors, capacitors, inductors, transformers, and even diodes are all considered passive devices. Active devices include, but are not limited to, vacuum tubes, transistors, silicon-controlled rectifiers (SCRs), and TRIACs. A case might be made for the saturable reactor to be defined as an active device, since it is able to control an AC current with a DC current, but I've never heard it referred to as such. The operation of each of these active devices will be explored in later chapters of this volume.

All active devices control the flow of electrons through them. Some active devices allow a voltage to control this current while other active devices allow another current to do the job. Devices utilizing a static voltage as the controlling signal are, not surprisingly, called *voltage-controlled* devices. Devices working on the principle of one current controlling another current are known as *current-controlled* devices. For the record, vacuum tubes are voltage-controlled devices while transistors are made as either voltage-controlled or current controlled types. The first type of transistor successfully demonstrated was a current-controlled device.

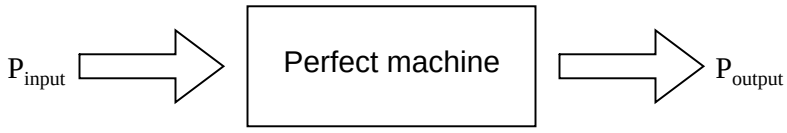
1.3 Amplifiers

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.

Because of this disparity between *controlling* and *controlled* powers, active devices may be employed to govern a large amount of power (controlled) by the application of a small amount of power (controlling). This behavior is known as *amplification*.

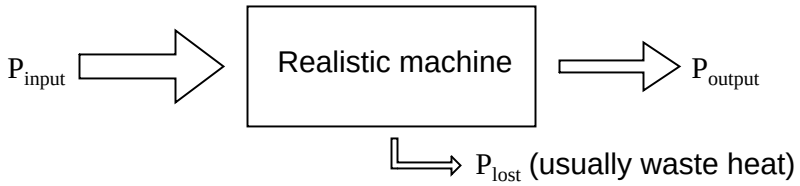
It is a fundamental rule of physics that energy can neither be created nor destroyed. Stated formally, this rule is known as the Law of Conservation of Energy, and no exceptions to it have been discovered to date. If this Law is true – and an overwhelming mass of experimental data suggests that it is – then it is impossible to build a device capable of taking a small amount of energy and magically transforming it into a large amount of energy. All machines, electric and electronic circuits

included, have an upper efficiency limit of 100 percent. At best, power out equals power in:



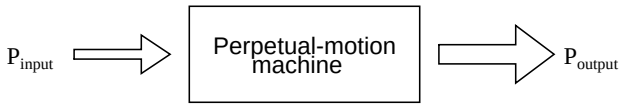
$$\text{Efficiency} = \frac{P_{output}}{P_{input}} = 1 = 100\%$$

Usually, machines fail even to meet this limit, losing some of their input energy in the form of heat which is radiated into surrounding space and therefore not part of the output energy stream.

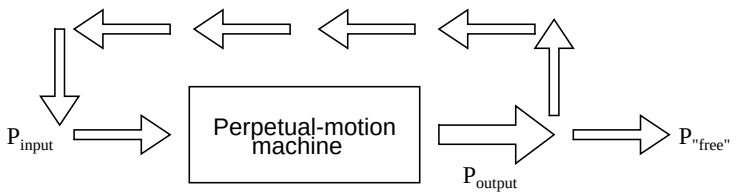


$$\text{Efficiency} = \frac{P_{output}}{P_{input}} < 1 = \text{less than } 100\%$$

Many people have attempted, without success, to design and build machines that output more power than they take in. Not only would such a *perpetual motion* machine prove that the Law of Energy Conservation was not a Law after all, but it would usher in a technological revolution such as the world has never seen, for it could power itself in a circular loop and generate excess power for "free:"



$$\text{Efficiency} = \frac{P_{output}}{P_{input}} > 1 = \text{more than } 100\%$$

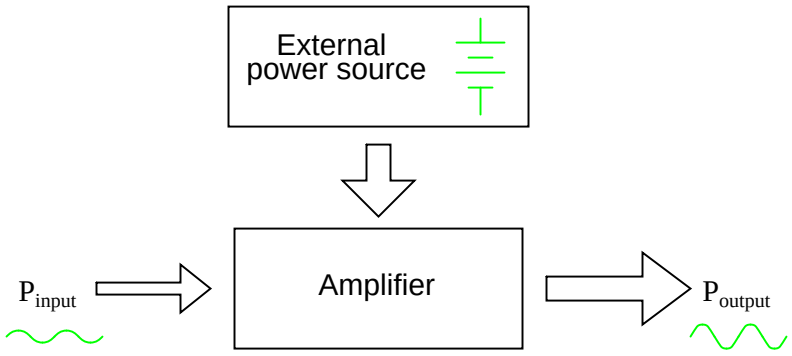


Despite much effort and many unscrupulous claims of "free energy" or *over-unity* machines, not one has ever passed the simple test of powering itself with its own energy output and generating

energy to spare.

There does exist, however, a class of machines known as *amplifiers*, which are able to take in small-power signals and output signals of much greater power. The key to understanding how amplifiers can exist without violating the Law of Energy Conservation lies in the behavior of active devices.

Because active devices have the ability to *control* a large amount of electrical power with a small amount of electrical power, they may be arranged in circuit so as to duplicate the form of the input signal power from a larger amount of power supplied by an external power source. The result is a device that appears to magically magnify the power of a small electrical signal (usually an AC voltage waveform) into an identically-shaped waveform of larger magnitude. The Law of Energy Conservation is not violated because the additional power is supplied by an external source, usually a DC battery or equivalent. The amplifier neither creates nor destroys energy, but merely reshapes it into the waveform desired:



In other words, the current-controlling behavior of active devices is employed to *shape* DC power from the external power source into the same waveform as the input signal, producing an output signal of like shape but different (greater) power magnitude. The transistor or other active device within an amplifier merely forms a larger *copy* of the input signal waveform out of the "raw" DC power provided by a battery or other power source.

Amplifiers, like all machines, are limited in efficiency to a maximum of 100 percent. Usually, electronic amplifiers are far less efficient than that, dissipating considerable amounts of energy in the form of waste heat. Because the efficiency of an amplifier is always 100 percent or less, one can never be made to function as a "perpetual motion" device.

The requirement of an external source of power is common to all types of amplifiers, electrical and non-electrical. A common example of a non-electrical amplification system would be power steering in an automobile, amplifying the power of the driver's arms in turning the steering wheel to move the front wheels of the car. The source of power necessary for the amplification comes from the engine. The active device controlling the driver's "input signal" is a hydraulic valve shuttling fluid power from a pump attached to the engine to a hydraulic piston assisting wheel motion. If the engine stops running, the amplification system fails to amplify the driver's arm power and the car becomes very difficult to turn.

1.4 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:

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

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

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

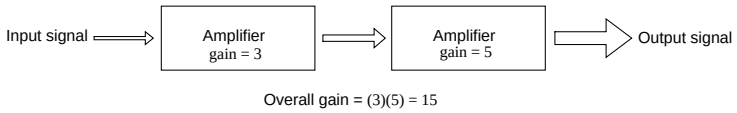
In the last two examples I specifically identified the gains and signal magnitudes in terms of "AC." This was intentional, and illustrates an important concept: electronic amplifiers often respond differently to AC and DC input signals, and may amplify them to different extents. Another way of saying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



1.5 Decibels

In its simplest form, an amplifier's *gain* is a ratio of output over input. Like all ratios, this form of gain is unitless. However, there is an actual unit intended to represent gain, and it is called the *bel*.

As a unit, the bel was actually devised as a convenient way to represent power *loss* in telephone system wiring rather than *gain* in amplifiers. The unit's name is derived from Alexander Graham Bell, the famous Scottish inventor whose work was instrumental in developing telephone systems. Originally, the bel represented the amount of signal power loss due to resistance over a standard length of electrical cable. Now, it is defined in terms of the common (base 10) logarithm of a power ratio (output power divided by input power):

$$A_{P(ratio)} = \frac{P_{output}}{P_{input}}$$

$$A_{P(Bel)} = \log \frac{P_{output}}{P_{input}}$$

Because the bel is a logarithmic unit, it is nonlinear. To give you an idea of how this works, consider the following table of figures, comparing power losses and gains in bels versus simple ratios:

Loss/gain as a ratio	Loss/gain in bels
$\frac{P_{output}}{P_{input}}$	$\log \frac{P_{output}}{P_{input}}$
1000	3 B
100	2 B
10	1 B
1 (no loss or gain)	0 B
0.1	-1 B
0.01	-2 B
0.001	-3 B

It was later decided that the bel was too large of a unit to be used directly, and so it became customary to apply the metric prefix *deci* (meaning 1/10) to it, making it *deci*bels, or dB. Now, the expression "dB" is so common that many people do not realize it is a combination of "deci-" and "-bel," or that there even is such a unit as the "bel." To put this into perspective, here is another table contrasting power gain/loss ratios against decibels:

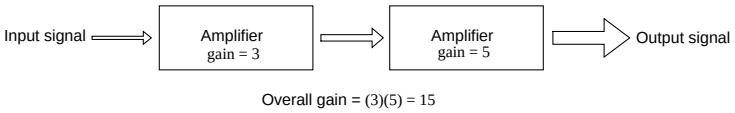
Loss/gain as a ratio	Loss/gain in decibels
$\frac{P_{output}}{P_{input}}$	$10 \log \frac{P_{output}}{P_{input}}$
1000	30 dB
100	20 dB
10	10 dB
1 (no loss or gain)	0 dB
0.1	-10 dB
0.01	-20 dB
0.001	-30 dB

As a logarithmic unit, this mode of power gain expression covers a wide range of ratios with a minimal span in figures. It is reasonable to ask, "why did anyone feel the need to invent a *logarithmic* unit for electrical signal power loss in a telephone system?" The answer is related to the dynamics of human hearing, the perceptive intensity of which is logarithmic in nature.

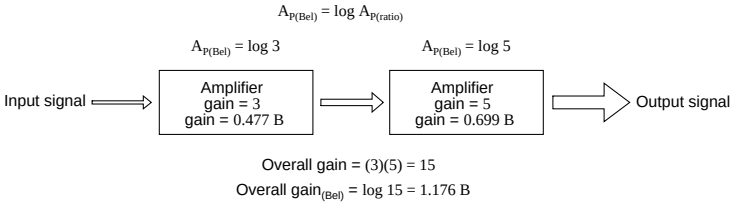
Human hearing is highly nonlinear: in order to double the perceived intensity of a sound, the actual sound power must be multiplied by a factor of ten. Relating telephone signal power loss in terms of the logarithmic "bel" scale makes perfect sense in this context: a power loss of 1 bel translates to a perceived sound loss of 50 percent, or 1/2. A power gain of 1 bel translates to a doubling in the perceived intensity of the sound.

An almost perfect analogy to the bel scale is the Richter scale used to describe earthquake intensity: a 6.0 Richter earthquake is 10 times more powerful than a 5.0 Richter earthquake; a 7.0 Richter earthquake 100 times more powerful than a 5.0 Richter earthquake; a 4.0 Richter earthquake is 1/10 as powerful as a 5.0 Richter earthquake, and so on. The measurement scale for chemical pH is likewise logarithmic, a difference of 1 on the scale is equivalent to a tenfold difference in hydrogen ion concentration of a chemical solution. An advantage of using a logarithmic measurement scale is the tremendous range of expression afforded by a relatively small span of numerical values, and it is this advantage which secures the use of Richter numbers for earthquakes and pH for hydrogen ion activity.

Another reason for the adoption of the bel as a unit for gain is for simple expression of system gains and losses. Consider the last system example where two amplifiers were connected tandem to amplify a signal. The respective gain for each amplifier was expressed as a ratio, and the overall gain for the system was the product (multiplication) of those two ratios:

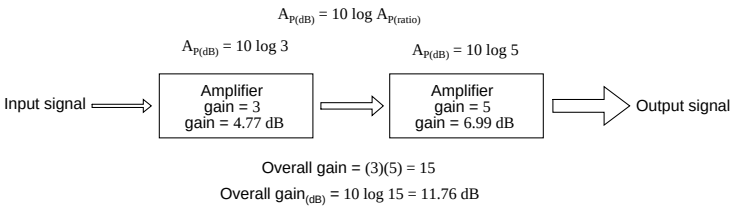


If these figures represented *power* gains, we could directly apply the unit of bels to the task of representing the gain of each amplifier, and of the system altogether:



Close inspection of these gain figures in the unit of "bel" yields a discovery: they're additive. Ratio gain figures are multiplicative for staged amplifiers, but gains expressed in bels *add* rather than *multiply* to equal the overall system gain. The first amplifier with its power gain of 0.477 B adds to the second amplifier's power gain of 0.699 B to make a system with an overall power gain of 1.176 B.

Recalculating for decibels rather than bels, we notice the same phenomenon:



To those already familiar with the arithmetic properties of logarithms, this is no surprise. It is an elementary rule of algebra that the antilogarithm of the sum of two numbers' logarithm values equals the product of the two original numbers. In other words, if we take two numbers and determine the logarithm of each, then add those two logarithm figures together, then determine the "antilogarithm" of that sum (elevate the base number of the logarithm – in this case, 10 – to the power of that sum), the result will be the same as if we had simply multiplied the two original numbers together. This algebraic rule forms the heart of a device called a *slide rule*, an analog computer which could, among other things, determine the products and quotients of numbers by addition (adding together physical lengths marked on sliding wood, metal, or plastic scales). Given a table of logarithm figures, the same mathematical trick could be used to perform otherwise complex multiplications and divisions by only having to do additions and subtractions, respectively. With the advent of high-speed, handheld, digital calculator devices, this elegant calculation technique virtually disappeared from popular use. However, it is still important to understand when working with measurement scales that are logarithmic in nature, such as the bel (decibel) and Richter scales.

When converting a power gain from units of bels or decibels to a unitless ratio, the mathematical inverse function of common logarithms is used: powers of 10, or the *antilog*.

If:

$$A_{P(Bel)} = \log A_{P(ratio)}$$

Then:

$$A_{P(ratio)} = 10^{A_{P(Bel)}}$$

Converting decibels into unitless ratios for power gain is much the same, only a division factor of 10 is included in the exponent term:

If:

$$A_{P(dB)} = 10 \log A_{P(ratio)}$$

Then:

$$A_{P(ratio)} = 10^{\frac{A_{P(dB)}}{10}}$$

Because the bel is fundamentally a unit of *power* gain or loss in a system, voltage or current gains and losses don't convert to bels or dB in quite the same way. When using bels or decibels to express a gain other than power, be it voltage or current, we must perform the calculation in terms of how much power gain there would be for that amount of voltage or current gain. For a constant load impedance, a voltage or current gain of 2 equates to a power gain of 4 (2^2); a voltage or current gain of 3 equates to a power gain of 9 (3^2). If we multiply either voltage or current by a given factor, then the power gain incurred by that multiplication will be the square of that factor. This relates back to the forms of Joule's Law where power was calculated from either voltage or current, and resistance:

$$P = \frac{E^2}{R}$$

$$P = I^2R$$

Power is proportional to the *square* of either voltage or current

Thus, when translating a voltage or current gain *ratio* into a respective gain in terms of the bel unit, we must include this exponent in the equation(s):

$$A_{P(Bel)} = \log A_{P(ratio)}$$

$$A_{V(Bel)} = \log A_{V(ratio)}^{2}$$

$$A_{I(Bel)} = \log A_{I(ratio)}^{2}$$

Exponent required

The same exponent requirement holds true when expressing voltage or current gains in terms of decibels:

$$A_{P(dB)} = 10 \log A_{P(ratio)}$$

$$A_{V(dB)} = 10 \log A_{V(ratio)}^{2}$$

$$A_{I(dB)} = 10 \log A_{I(ratio)}^{2}$$

Exponent required

However, thanks to another interesting property of logarithms, we can simplify these equations to eliminate the exponent by including the "2" as a *multiplying factor* for the logarithm function. In other words, instead of taking the logarithm of the *square* of the voltage or current gain, we just multiply the voltage or current gain's logarithm figure by 2 and the final result in bels or decibels will be the same:

For bels:

$$A_{V(Bel)} = \log A_{V(ratio)}^{2} \qquad A_{I(Bel)} = \log A_{I(ratio)}^{2}$$

. . . is the same as . . .

$$A_{V(Bel)} = 2 \log A_{V(ratio)} \qquad A_{I(Bel)} = 2 \log A_{I(ratio)}$$

For decibels:

$$A_{V(dB)} = 10 \log A_{V(ratio)}^{2} \qquad A_{I(dB)} = 10 \log A_{I(ratio)}^{2}$$

. . . is the same as . . .

$$A_{V(dB)} = 20 \log A_{V(ratio)} \qquad A_{I(dB)} = 20 \log A_{I(ratio)}$$

The process of converting voltage or current gains from bels or decibels into unitless ratios is much the same as it is for power gains:

If:

$$A_{V(Bel)} = 2 \log A_{V(ratio)} \qquad A_{I(Bel)} = 2 \log A_{I(ratio)}$$

Then:

$$A_{V(ratio)} = 10^{\frac{A_{V(Bel)}}{2}} \qquad A_{I(ratio)} = 10^{\frac{A_{I(Bel)}}{2}}$$

Here are the equations used for converting voltage or current gains in decibels into unitless ratios:

If:

$$A_{V(dB)} = 20 \log A_{V(ratio)} \qquad A_{I(dB)} = 20 \log A_{I(ratio)}$$

Then:

$$A_{V(ratio)} = 10^{\frac{A_{V(dB)}}{20}} \qquad A_{I(ratio)} = 10^{\frac{A_{I(dB)}}{20}}$$

While the bel is a unit naturally scaled for power, another logarithmic unit has been invented to directly express voltage or current gains/losses, and it is based on the *natural* logarithm rather than

the *common* logarithm as bels and decibels are. Called the *neper*, its unit symbol is a lower-case "n."

$$A_{V(ratio)} = \frac{V_{output}}{V_{input}} \qquad A_{I(ratio)} = \frac{I_{output}}{I_{input}}$$

$$A_{V(neper)} = \ln A_{V(ratio)} \qquad A_{I(neper)} = \ln A_{I(ratio)}$$

For better or for worse, neither the neper nor its attenuated cousin, the *decineper*, is popularly used as a unit in American engineering applications.

- **REVIEW:**
- Gains and losses may be expressed in terms of a unitless ratio, or in the unit of bels (B) or decibels (dB). A decibel is literally a *deci*-bel: one-tenth of a bel.
- The bel is fundamentally a unit for expressing *power* gain or loss. To convert a power ratio to either bels or decibels, use one of these equations:
- $A_{P(Bel)} = \log A_{P(ratio)}$ $A_{P(db)} = 10 \log A_{P(ratio)}$
- When using the unit of the bel or decibel to express a *voltage* or *current* ratio, it must be cast in terms of the an equivalent *power* ratio. Practically, this means the use of different equations, with a multiplication factor of 2 for the logarithm value corresponding to an exponent of 2 for the voltage or current gain ratio:
- $A_{V(Bel)} = 2 \log A_{V(ratio)}$ $A_{V(dB)} = 20 \log A_{V(ratio)}$

  $A_{I(Bel)} = 2 \log A_{I(ratio)}$ $A_{I(dB)} = 20 \log A_{I(ratio)}$
- To convert a decibel gain into a unitless ratio gain, use one of these equations:
- $A_{V(ratio)} = 10^{\frac{A_{V(dB)}}{20}}$

  $A_{I(ratio)} = 10^{\frac{A_{I(dB)}}{20}}$

  $A_{P(ratio)} = 10^{\frac{A_{P(dB)}}{10}}$
- A gain (amplification) is expressed as a positive bel or decibel figure. A loss (attenuation) is expressed as a negative bel or decibel figure. Unity gain (no gain or loss; ratio = 1) is expressed as zero bels or zero decibels.
- When calculating overall gain for an amplifier system composed of multiple amplifier stages, individual gain ratios are *multiplied* to find the overall gain ratio. Bel or decibel figures for each amplifier stage, on the other hand, are *added* together to determine overall gain.

1.6 Absolute dB scales

It is also possible to use the decibel as a unit of absolute power, in addition to using it as an expression of power gain or loss. A common example of this is the use of decibels as a measurement of sound pressure intensity. In cases like these, the measurement is made in reference to some standardized power level defined as 0 dB. For measurements of sound pressure, 0 dB is loosely defined as the lower threshold of human hearing, objectively quantified as 1 picowatt of sound power per square meter of area.

A sound measuring 40 dB on the decibel sound scale would be 10^4 times greater than the threshold of hearing. A 100 dB sound would be 10^{10} (ten billion) times greater than the threshold of hearing.

Because the human ear is not equally sensitive to all frequencies of sound, variations of the decibel sound-power scale have been developed to represent physiologically equivalent sound intensities at different frequencies. Some sound intensity instruments were equipped with filter networks to give disproportionate indications across the frequency scale, the intent of which to better represent the effects of sound on the human body. Three filtered scales became commonly known as the "A," "B," and "C" weighted scales. Decibel sound intensity indications measured through these respective filtering networks were given in units of dBA, dBB, and dBC. Today, the "A-weighted scale" is most commonly used for expressing the equivalent physiological impact on the human body, and is especially useful for rating dangerously loud noise sources.

Another standard-referenced system of power measurement in the unit of decibels has been established for use in telecommunications systems. This is called the *dBm* scale. The reference point, 0 dBm, is defined as 1 milliwatt of electrical power dissipated by a 600 Ω load. According to this scale, 10 dBm is equal to 10 times the reference power, or 10 milliwatts; 20 dBm is equal to 100 times the reference power, or 100 milliwatts. Some AC voltmeters come equipped with a dBm range or scale (sometimes labeled "DB") intended for use in measuring AC signal power across a 600 Ω load. 0 dBm on this scale is, of course, elevated above zero because it represents something greater than 0 (actually, it represents 0.7746 volts across a 600 Ω load, voltage being equal to the square root of power times resistance; the square root of 0.001 multiplied by 600). When viewed on the face of an analog meter movement, this dBm scale appears compressed on the left side and expanded on the right in a manner not unlike a resistance scale, owing to its logarithmic nature.

An adaptation of the dBm scale for audio signal strength is used in studio recording and broadcast engineering for standardizing volume levels, and is called the *VU* scale. VU meters are frequently seen on electronic recording instruments to indicate whether or not the recorded signal exceeds the maximum signal level limit of the device, where significant distortion will occur. This "volume indicator" scale is calibrated in according to the dBm scale, but does not directly indicate dBm for any signal other than steady sine-wave tones. The proper unit of measurement for a VU meter is *volume units*.

When relatively large signals are dealt with, and an absolute dB scale would be useful for representing signal level, specialized decibel scales are sometimes used with reference points greater than the 1mW used in dBm. Such is the case for the *dBW* scale, with a reference point of 0 dBW established at 1 watt. Another absolute measure of power called the *dBk* scale references 0 dBk at 1 kW, or 1000 watts.

- **REVIEW:**
- The unit of the bel or decibel may also be used to represent an absolute measurement of power

rather than just a relative gain or loss. For sound power measurements, 0 dB is defined as a standardized reference point of power equal to 1 picowatt per square meter. Another dB scale suited for sound intensity measurements is normalized to the same physiological effects as a 1000 Hz tone, and is called the *dBA* scale. In this system, 0 dBA is defined as any frequency sound having the same physiological equivalence as a 1 picowatt-per-square-meter tone at 1000 Hz.

- An electrical dB scale with an absolute reference point has been made for use in telecommunications systems. Called the *dBm* scale, its reference point of 0 dBm is defined as 1 milliwatt of AC signal power dissipated by a 600 Ω load.
- A *VU* meter reads audio signal level according to the dBm for sine-wave signals. Because its response to signals other than steady sine waves is not the same as true dBm, its unit of measurement is *volume units*.
- dB scales with greater absolute reference points than the dBm scale have been invented for high-power signals. The *dBW* scale has its reference point of 0 dBW defined as 1 watt of power. The *dBk* scale sets 1 kW (1000 watts) as the zero-point reference.

1.7 Contributors

Contributors to this chapter are listed in chronological order of their contributions, from most recent to first. See Appendix 2 (Contributor List) for dates and contact information.

Colin Barnard (November 2003): Correction regarding Alexander Graham Bell's country of origin (Scotland, not the United States).

Chapter 2

SOLID-STATE DEVICE THEORY

Contents

***** INCOMPLETE *****

2.1 Introduction

This chapter will cover the physics behind the operation of semiconductor devices and show how these principles are applied in several different types of semiconductor devices. Subsequent chapters will deal primarily with the practical aspects of these devices in circuits and omit theory as much as possible.

2.2 Quantum physics

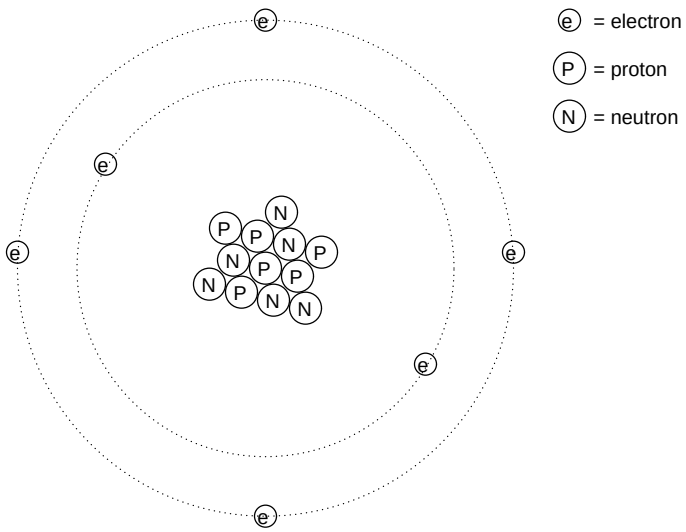
"I think it is safe to say that no one understands quantum mechanics."

Physicist Richard P. Feynman

To say that the invention of semiconductor devices was a revolution would not be an exaggeration. Not only was this an impressive technological accomplishment, but it paved the way for developments that would indelibly alter modern society. Semiconductor devices made possible miniaturized electronics, including computers, certain types of medical diagnostic and treatment equipment, and popular telecommunication devices, to name a few applications of this technology.

But behind this revolution in technology stands an even greater revolution in general science: the field of *quantum physics*. Without this leap in understanding the natural world, the development of semiconductor devices (and more advanced electronic devices still under development) would never have been possible. Quantum physics is an incredibly complicated realm of science, and this chapter is by no means a complete discussion of it, but rather a brief overview. When scientists of Feynman's caliber say that "no one understands [it]," you can be sure it is a complex subject. Without a basic understanding of quantum physics, or at least an understanding of the scientific discoveries that led to its formulation, though, it is impossible to understand how and why semiconductor electronic devices function. Most introductory electronics textbooks I've read attempt to explain semiconductors in terms of "classical" physics, resulting in more confusion than comprehension.

Many of us have seen diagrams of atoms that look something like this:



Tiny particles of matter called *protons* and *neutrons* make up the center of the atom, while *electrons* orbit around not unlike planets around a star. The nucleus carries a positive electrical charge, owing to the presence of protons (the neutrons have no electrical charge whatsoever), while the atom's balancing negative charge resides in the orbiting electrons. The negative electrons tend to be attracted to the positive protons just as planets are gravitationally attracted toward whatever object(s) they orbit, yet the orbits are stable due to the electrons' motion. We owe this popular model of the atom to the work of Ernest Rutherford, who around the year 1911 experimentally determined that atoms' positive charges were concentrated in a tiny, dense core rather than being spread evenly about the diameter as was proposed by an earlier researcher, J.J. Thompson.

While Rutherford's atomic model accounted for experimental data better than Thompson's, it still wasn't perfect. Further attempts at defining atomic structure were undertaken, and these efforts

helped pave the way for the bizarre discoveries of quantum physics. Today our understanding of the atom is quite a bit more complex. However, despite the revolution of quantum physics and the impact it had on our understanding of atomic structure, Rutherford's solar-system picture of the atom embedded itself in the popular conscience to such a degree that it persists in some areas of study even when inappropriate.

Consider this short description of electrons in an atom, taken from a popular electronics textbook:

> *Orbiting negative electrons are therefore attracted toward the positive nucleus, which leads us to the question of why the electrons do not fly into the atom's nucleus. The answer is that the orbiting electrons remain in their stable orbit due to two equal but opposite forces. The centrifugal outward force exerted on the electrons due to the orbit counteracts the attractive inward force (centripetal) trying to pull the electrons toward the nucleus due to the unlike charges.*

In keeping with the Rutherford model, this author casts the electrons as solid chunks of matter engaged in circular orbits, their inward attraction to the oppositely charged nucleus balanced by their motion. The reference to "centrifugal force" is technically incorrect (even for orbiting planets), but is easily forgiven due to its popular acceptance: in reality, there is no such thing as a force pushing *any* orbiting body *away* from its center of orbit. It only seems that way because a body's inertia tends to keep it traveling in a straight line, and since an orbit is a constant deviation (acceleration) from straight-line travel, there is constant inertial opposition to whatever force is attracting the body toward the orbit center (centripetal), be it gravity, electrostatic attraction, or even the tension of a mechanical link.

The real problem with this explanation, however, is the idea of electrons traveling in circular orbits in the first place. It is a verifiable fact that accelerating electric charges emit electromagnetic radiation, and this fact was known even in Rutherford's time. Since orbiting motion is a form of acceleration (the orbiting object in constant acceleration away from normal, straight-line motion), electrons in an orbiting state should be throwing off radiation like mud from a spinning tire. Electrons accelerated around circular paths in particle accelerators called *synchrotrons* are known to do this, and the result is called *synchrotron radiation*. If electrons were losing energy in this way, their orbits would eventually decay, resulting in collisions with the positively charged nucleus. However, this doesn't ordinarily happen within atoms. Indeed, electron "orbits" are remarkably stable over a wide range of conditions.

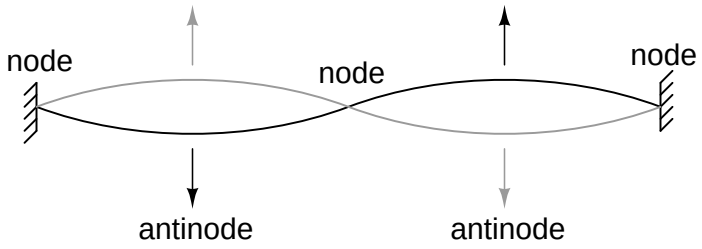
Furthermore, experiments with "excited" atoms demonstrated that electromagnetic energy emitted by an atom occurs only at certain, definite frequencies. Atoms that are "excited" by outside influences such as light are known to absorb that energy and return it as electromagnetic waves of very specific frequencies, like a tuning fork that rings at a fixed pitch no matter how it is struck. When the light emitted by an excited atom is divided into its constituent frequencies (colors) by a prism, distinct lines of color appear in the spectrum, the pattern of spectral lines being unique to that element. So regular is this phenomenon that it is commonly used to identify atomic elements, and even measure the proportions of each element in a compound or chemical mixture. According to Rutherford's solar-system atomic model (regarding electrons as chunks of matter free to orbit at any radius) and the laws of classical physics, excited atoms should be able to return energy over a virtually limitless range of frequencies rather than a select few. In other words, if Rutherford's model were correct, there would be no "tuning fork" effect, and the light spectrum emitted by any atom would appear as a continuous band of colors rather than as a few distinct lines.

A pioneering researcher by the name of Niels Bohr attempted to improve upon Rutherford's model after studying in Rutherford's laboratory for several months in 1912. Trying to harmonize the findings of other physicists (most notably, Max Planck and Albert Einstein), Bohr suggested that each electron possessed a certain, specific amount of energy, and that their orbits were likewise *quantized* such that they could only occupy certain places around the nucleus, somewhat like marbles fixed in circular tracks around the nucleus rather than the free-ranging satellites they were formerly imagined to be. In deference to the laws of electromagnetics and accelerating charges, Bohr referred to these "orbits" as *stationary states* so as to escape the implication that they were in motion.

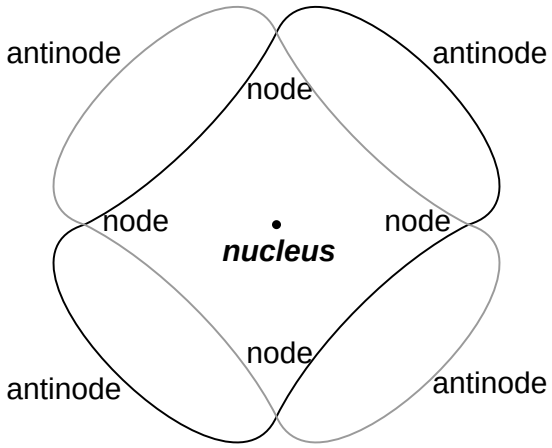
While Bohr's ambitious attempt at re-framing the structure of the atom in terms that agreed closer to experimental results was a milestone in physics, it was by no means complete. His mathematical analyses produced better predictions of experimental events than analyses belonging to previous models, but there were still some unanswered questions as to *why* electrons would behave in such strange ways. The assertion that electrons existed in stationary, quantized states around the nucleus certainly accounted for experimental data better than Rutherford's model, but he had no idea what would force electrons to manifest those particular states. The answer to that question had to come from another physicist, Louis de Broglie, about a decade later.

De Broglie proposed that electrons, like photons (particles of light) manifested both particle-like and wave-like properties. Building on this proposal, he suggested that an analysis of orbiting electrons from a wave perspective rather than a particle perspective might make more sense of their quantized nature. Indeed, this was the case, and another breakthrough in understanding was reached.

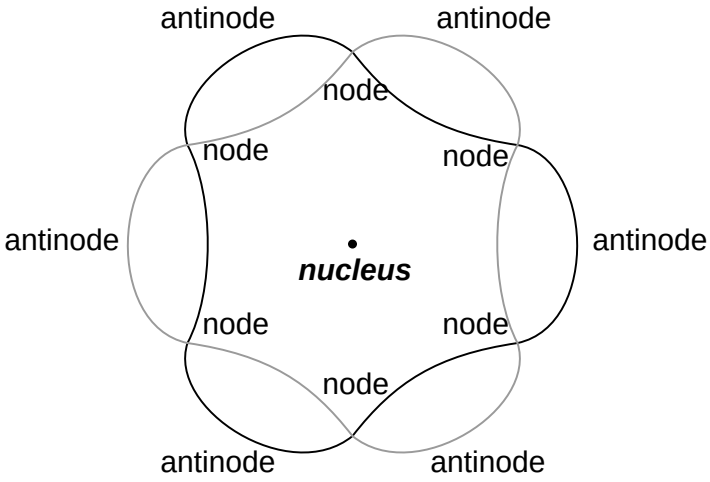
The atom according to de Broglie consisted of electrons existing in the form of *standing waves*, a phenomenon well known to physicists in a variety of forms. Like the plucked string of a musical instrument vibrating at a resonant frequency, with "nodes" and "antinodes" at stable positions along its length, de Broglie envisioned electrons around atoms standing as waves bent around a circle:



String vibrating at resonant frequency between two fixed points forms a ***standing wave****.*



"Orbiting" electron as a standing wave around the nucleus. Two cycles per "orbit" shown.



"Orbiting" electron as a standing wave around the nucleus. Three cycles per "orbit" shown.

Electrons could only exist in certain, definite "orbits" around the nucleus because those were the only distances where the wave ends would match. In any other radius, the wave would destructively interfere with itself and thus cease to exist.

De Broglie's hypothesis gave both mathematical support and a convenient physical analogy to account for the quantized states of electrons within an atom, but his atomic model was still incomplete. Within a few years, though, physicists Werner Heisenberg and Erwin Schrodinger, working independently of each other, built upon de Broglie's concept of a matter-wave duality to create more mathematically rigorous models of subatomic particles.

This theoretical advance from de Broglie's primitive standing wave model to Heisenberg's matrix and Schrodinger's differential equation models was given the name *quantum mechanics*, and it

introduced a rather shocking characteristic to the world of subatomic particles: the trait of probability, or uncertainty. According to the new quantum theory, it was impossible to determine the exact position *and* exact momentum of a particle at the same time. Popular explanations of this "uncertainty principle" usually cast it in terms of error caused by the process of measurement (i.e. by attempting to precisely measure the position of an electron, you interfere with its momentum and thus cannot know what it was before the position measurement was taken, and visa versa), but the truth is actually much more mysterious than simple measurement interference. The startling implication of quantum mechanics is that particles do not actually possess precise positions *and* momenta, but rather balance the two quantities in a such way that their combined uncertainties never diminish below a certain minimum value.

It is interesting to note that this form of "uncertainty" relationship exists in areas other than quantum mechanics. As discussed in the "Mixed-Frequency AC Signals" chapter in volume II of this book series, there is a mutually exclusive relationship between the certainty of a waveform's time-domain data and its frequency-domain data. In simple terms, the more precisely we know its constituent frequency(ies), the less precisely we know its amplitude in time, and vice versa. To quote myself:

> *A waveform of infinite duration (infinite number of cycles) can be analyzed with absolute precision, but the less cycles available to the computer for analysis, the less precise the analysis. . . The fewer times that a wave cycles, the less certain its frequency is. Taking this concept to its logical extreme, a short pulse – a waveform that doesn't even complete a cycle – actually has no frequency, but rather acts as an infinite range of frequencies. This principle is common to all wave-based phenomena, not just AC voltages and currents.*

In order to precisely determine the amplitude of a varying signal, we must sample it over a very narrow span of time. However, doing this limits our view of the wave's frequency. Conversely, to determine a wave's frequency with great precision, we must sample it over many, many cycles, which means we lose view of its amplitude at any given moment. Thus, we cannot simultaneously know the instantaneous amplitude and the overall frequency of any wave with unlimited precision. Stranger yet, this uncertainty is much more than observer imprecision; it resides in the very nature of the wave itself. It is not as though it would be possible, given the proper technology, to obtain precise measurements of *both* instantaneous amplitude and frequency at once. Quite literally, a wave cannot possess both a precise, instantaneous amplitude, and a precise frequency at the same time.

Likewise, the minimum uncertainty of a particle's position and momentum expressed by Heisenberg and Schrodinger has nothing to do with limitation in measurement; rather it is an intrinsic property of the particle's matter-wave dual nature. Electrons, therefore, do not really exist in their "orbits" as precisely defined bits of matter, or even as precisely defined waveshapes, but rather as "clouds" – the technical term is *wavefunction* – of probability distribution, as if each electron were "spread" or "smeared" over a range of positions and momenta.

This radical view of electrons as imprecise clouds at first seems to contradict the original principle of quantized electron states: that electrons exist in discrete, defined "orbits" around atomic nuclei. It was, after all, this discovery that led to the formation of quantum theory to explain it. How odd it seems that a theory developed to explain the discrete behavior of electrons ends up declaring that electrons exist as "clouds" rather than as discrete pieces of matter. However, the quantized behavior of electrons does not depend on electrons having definite position and momentum values,

but rather on other properties called *quantum numbers*. In essence, quantum mechanics dispenses with commonly held notions of absolute position and absolute momentum, and replaces them with absolute notions of a sort having no analogue in common experience.

Even though electrons are known to exist in ethereal, "cloud-like" forms of distributed probability rather than as discrete chunks of matter, those "clouds" possess other characteristics that *are* discrete. Any electron in an atom can be described in terms of four numerical measures (the previously mentioned *quantum numbers*), called the **Principal**, **Angular Momentum**, **Magnetic**, and **Spin** numbers. The following is a synopsis of each of these numbers' meanings:

Principal Quantum Number: Symbolized by the letter **n**, this number describes the *shell* that an electron resides in. An electron "shell" is a region of space around an atom's nucleus that electrons are allowed to exist in, corresponding to the stable "standing wave" patterns of de Broglie and Bohr. Electrons may "leap" from shell to shell, but cannot exist *between* the shell regions.

The principle quantum number can be any positive integer (a whole number, greater than or equal to 1). In other words, there is no such thing as a principle quantum number for an electron of 1/2 or -3. These integer values were not arrived at arbitrarily, but rather through experimental evidence of light spectra: the differing frequencies (colors) of light emitted by excited hydrogen atoms follow a sequence mathematically dependent on specific, integer values.

Each shell has the capacity to hold multiple electrons. An analogy for electron shells is the concentric rows of seats of an amphitheater. Just as a person seated in an amphitheater must choose a row to sit in (for there is no place to sit in the space *between* rows), electrons must "choose" a particular shell to "sit" in. Like amphitheater rows, the outermost shells are able to hold more electrons than the inner shells. Also, electrons tend to seek the lowest available shell, like people in an amphitheater trying to find the closest seat to the center stage. The higher the shell number, the greater the energy of the electrons in it.

The maximum number of electrons that any shell can hold is described by the equation $2n^2$, where "n" is the principle quantum number. Thus, the first shell (n=1) can hold 2 electrons; the second shell (n=2) 8 electrons, and the third shell (n=3) 18 electrons.

Electron shells in an atom are sometimes designated by letter rather than by number. The first shell (n=1) is labeled K, the second shell (n=2) L, the third shell (n=3) M, the fourth shell (n=4) N, the fifth shell (n=5) O, the sixth shell (n=6) P, and the seventh shell (n=7) Q.

Angular Momentum Quantum Number: Within each shell, there are *subshells*. One might be inclined to think of subshells as simple subdivisions of shells, like lanes dividing a road, but the truth is much stranger than this. Subshells are regions of space where electron "clouds" are allowed to exist, and different subshells actually have different *shapes*. The first subshell is shaped like a sphere, which makes sense to most people, visualizing a cloud of electrons surrounding the atomic nucleus in three dimensions. The second subshell, however, resembles a dumbbell, comprised of two "lobes" joined together at a single point near the atom's center. The third subshell typically resembles a set of four "lobes" clustered around the atom's nucleus. These subshell shapes are reminiscent of graphical depictions of radio antenna signal strength, with bulbous lobe-shaped regions extending from the antenna in various directions.

Valid angular momentum quantum numbers are positive integers like principal quantum numbers, but also include zero. These quantum numbers for electrons are symbolized by the letter **l**. The number of subshells in a shell is equal to the shell's principal quantum number. Thus, the first shell

(n=1) has one subshell, numbered 0; the second shell (n=2) has two subshells, numbered 0 and 1; the third shell (n=3) has three subshells, numbered 0, 1, and 2.

An older convention for subshell description used letters rather than numbers. In this notational system, the first subshell (l=0) was designated *s*, the second subshell (l=1) designated *p*, the third subshell (l=2) designated *d*, and the fourth subshell (l=3) designated *f*. The letters come from the words *sharp*, *principal* (not to be confused with the principal quantum number, *n*), *diffuse*, and *fundamental*. You will still see this notational convention in many periodic tables, used to designate the electron configuration of the atoms' outermost, or *valence*, shells.

Magnetic Quantum Number: The magnetic quantum number for an electron classifies which orientation its subshell shape is pointed. For each subshell in each shell, there are multiple directions in which the "lobes" can point, and these different orientations are called *orbitals*. For the first subshell (s; l=0), which resembles a sphere, there is no "direction" it can "point," so there is only one orbital. For the second (p; l=1) subshell in each shell, which resembles a dumbbell, there are three different directions they can be oriented (think of three dumbbells intersecting in the middle, each oriented along a different axis in a three-axis coordinate system).

Valid numerical values for this quantum number consist of integers ranging from -l to l, and are symbolized as $\mathbf{m}_l$ in atomic physics and $\mathbf{l}_z$ in nuclear physics. To calculate the number of orbitals in any given subshell, double the subshell number and add 1 $(2l + 1)$. For example, the first subshell (l=0) in any shell contains a single orbital, numbered 0; the second subshell (l=1) in any shell contains three orbitals, numbered -1, 0, and 1; the third subshell (l=2) contains five orbitals, numbered -2, -1, 0, 1, and 2; and so on.

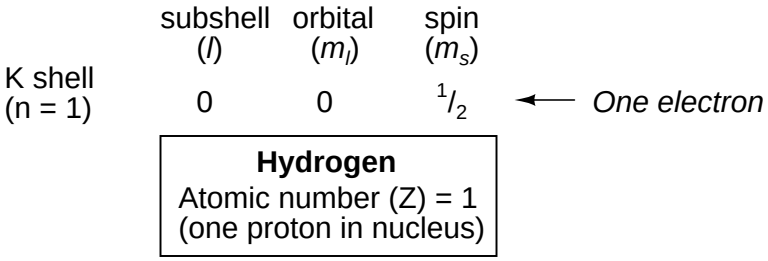
Like principal quantum numbers, the magnetic quantum number arose directly from experimental evidence: the division of spectral lines as a result of exposing an ionized gas to a magnetic field, hence the name "magnetic" quantum number.

Spin Quantum Number: Like the magnetic quantum number, this property of atomic electrons was discovered through experimentation. Close observation of spectral lines revealed that each line was actually a pair of very closely-spaced lines, and this so-called *fine structure* was hypothesized to be the result of each electron "spinning" on an axis like a planet. Electrons with different "spins" would give off slightly different frequencies of light when excited, and so the quantum number of "spin" came to be named as such. The concept of a spinning electron is now obsolete, being better suited to the (incorrect) view of electrons as discrete chunks of matter rather than as the "clouds" they really are, but the name remains.

Spin quantum numbers are symbolized as $\mathbf{m}_s$ in atomic physics and $\mathbf{s}_z$ in nuclear physics. For each orbital in each subshell in each shell, there can be two electrons, one with a spin of $+1/2$ and the other with a spin of $-1/2$.

The physicist Wolfgang Pauli developed a principle explaining the ordering of electrons in an atom according to these quantum numbers. His principle, called the *Pauli exclusion principle*, states that no two electrons in the same atom may occupy the exact same quantum states. That is, each electron in an atom has a unique set of quantum numbers. This limits the number of electrons that may occupy any given orbital, subshell, and shell.

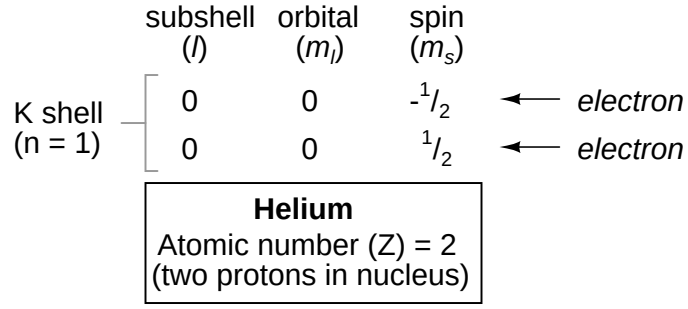
Shown here is the electron arrangement for a hydrogen atom:



Spectroscopic notation: $1s^1$

With one proton in the nucleus, it takes one electron to electrostatically balance the atom (the proton's positive electric charge exactly balanced by the electron's negative electric charge). This one electron resides in the lowest shell (n=1), the first subshell (l=0), in the only orbital (spatial orientation) of that subshell (m_l=0), with a spin value of 1/2. A very common method of describing this organization is by listing the electrons according to their shells and subshells in a convention called *spectroscopic notation*. In this notation, the shell number is shown as an integer, the subshell as a letter (s,p,d,f), and the total number of electrons in the subshell (all orbitals, all spins) as a superscript. Thus, hydrogen, with its lone electron residing in the base level, would be described as $1s^1$.

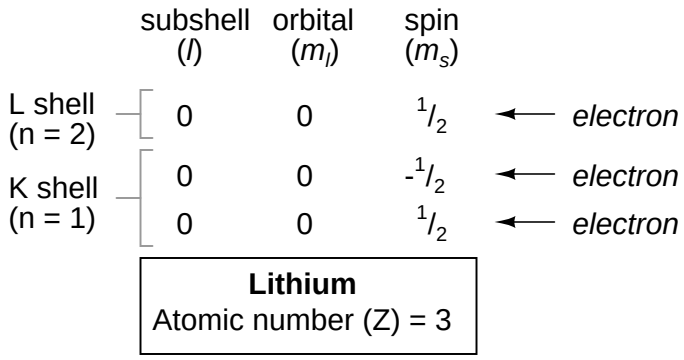
Proceeding to the next atom type (in order of atomic number), we have the element helium:



Spectroscopic notation: $1s^2$

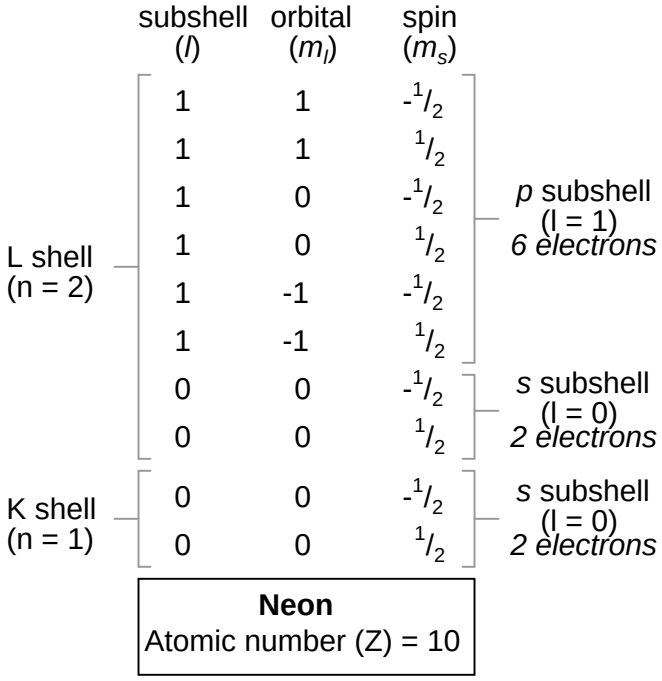
A helium atom has two protons in the nucleus, and this necessitates two electrons to balance the double-positive electric charge. Since two electrons – one with spin=1/2 and the other with spin=-1/2 – will fit into one orbital, the electron configuration of helium requires no additional subshells or shells to hold the second electron.

However, an atom requiring three or more electrons *will* require additional subshells to hold all electrons, since only two electrons will fit into the lowest shell (n=1). Consider the next atom in the sequence of increasing atomic numbers, lithium:



Spectroscopic notation: $1s^2 2s^1$

An atom of lithium only uses a fraction of the L shell's (n=2) capacity. This shell actually has a total capacity of eight electrons (maximum shell capacity = $2n^2$ electrons). If we examine the organization of the atom with a completely filled L shell, we will see how all combinations of subshells, orbitals, and spins are occupied by electrons:



Spectroscopic notation: $1s^2 2s^2 2p^6$

Often, when the spectroscopic notation is given for an atom, any shells that are completely filled are omitted, and only the unfilled, or the highest-level filled shell, is denoted. For example, the element neon (shown in the previous illustration), which has two completely filled shells, may be

spectroscopically described simply as $2p^6$ rather than $1s^2 2s^2 2p^6$. Lithium, with its K shell completely filled and a solitary electron in the L shell, may be described simply as $2s^1$ rather than $1s^2 2s^1$.

The omission of completely filled, lower-level shells is not just a notational convenience. It also illustrates a basic principle of chemistry: that the chemical behavior of an element is primarily determined by its unfilled shells. Both hydrogen and lithium have a single electron in their outermost shells ($1s^1$ and $2s^1$, respectively), and this gives the two elements some similar properties. Both are highly reactive, and reactive in much the same way (bonding to similar elements in similar modes). It matters little that lithium has a completely filled K shell underneath its almost-vacant L shell: the unfilled L shell is the shell that determines its chemical behavior.

Elements having completely filled outer shells are classified as *noble*, and are distinguished by their almost complete non-reactivity with other elements. These elements used to be classified as *inert*, when it was thought that they were completely unreactive, but it is now known that they may form compounds with other elements under certain conditions.

Given the fact that elements with identical electron configurations in their outermost shell(s) exhibit similar chemical properties, it makes sense to organize the different elements in a table accordingly. Such a table is known as a *periodic table of the elements*, and modern tables follow this general form:

Periodic Table of the Elements

H 1 Hydrogen 1.00794 $1s^1$																	He 2 Helium 4.00260 $1s^2$
Li 3 Lithium 6.941 $2s^1$	Be 4 Beryllium 9.012182 $2s^2$											B 5 Boron 10.81 $2p^1$	C 6 Carbon 12.011 $2p^2$	N 7 Nitrogen 14.0067 $2p^3$	O 8 Oxygen 15.9994 $2p^4$	F 9 Fluorine 18.9984 $2p^5$	Ne 10 Neon 20.179 $2p^6$
Na 11 Sodium 22.989768 $3s^1$	Mg 12 Magnesium 24.3050 $3s^2$											Al 13 Aluminum 26.9815 $3p^1$	Si 14 Silicon 28.0855 $3p^2$	P 15 Phosphorus 30.9738 $3p^3$	S 16 Sulfur 32.06 $3p^4$	Cl 17 Chlorine 35.453 $3p^5$	Ar 18 Argon 39.948 $3p^6$
K 19 Potassium 39.0983 $4s^1$	Ca 20 Calcium 40.078 $4s^2$	Sc 21 Scandium 44.955910 $3d^1 4s^2$	Ti 22 Titanium 47.88 $3d^2 4s^2$	V 23 Vanadium 50.9415 $3d^3 4s^2$	Cr 24 Chromium 51.9961 $3d^5 4s^1$	Mn 25 Manganese 54.93805 $3d^5 4s^2$	Fe 26 Iron 55.847 $3d^6 4s^2$	Co 27 Cobalt 58.93320 $3d^7 4s^2$	Ni 28 Nickel 58.69 $3d^8 4s^2$	Cu 29 Copper 63.546 $3d^{10} 4s^1$	Zn 30 Zinc 65.39 $3d^{10} 4s^2$	Ga 31 Gallium 69.723 $4p^1$	Ge 32 Germanium 72.61 $4p^2$	As 33 Arsenic 74.92159 $4p^3$	Se 34 Selenium 78.96 $4p^4$	Br 35 Bromine 79.904 $4p^5$	Kr 36 Krypton 83.80 $4p^6$
Rb 37 Rubidium 85.4678 $5s^1$	Sr 38 Strontium 87.62 $5s^2$	Y 39 Yttrium 88.90585 $4d^1 5s^2$	Zr 40 Zirconium 91.224 $4d^2 5s^2$	Nb 41 Niobium 92.90638 $4d^4 5s^1$	Mo 42 Molybdenum 95.94 $4d^5 5s^1$	Tc 43 Technetium (98) $4d^5 5s^2$	Ru 44 Ruthenium 101.07 $4d^7 5s^1$	Rh 45 Rhodium 102.90550 $4d^8 5s^1$	Pd 46 Palladium 106.42 $4d^{10} 5s^0$	Ag 47 Silver 107.8682 $4d^{10} 5s^1$	Cd 48 Cadmium 112.411 $4d^{10} 5s^2$	In 49 Indium 114.82 $5p^1$	Sn 50 Tin 118.710 $5p^2$	Sb 51 Antimony 121.75 $5p^3$	Te 52 Tellurium 127.60 $5p^4$	I 53 Iodine 126.905 $5p^5$	Xe 54 Xenon 131.30 $5p^6$
Cs 55 Cesium 132.90543 $6s^1$	Ba 56 Barium 137.327 $6s^2$	*57 - 71 Lanthanide series*	Hf 72 Hafnium 178.49 $5d^2 6s^2$	Ta 73 Tantalum 180.9479 $5d^3 6s^2$	W 74 Tungsten 183.85 $5d^4 6s^2$	Re 75 Rhenium 186.207 $5d^5 6s^2$	Os 76 Osmium 190.2 $5d^6 6s^2$	Ir 77 Iridium 192.22 $5d^7 6s^2$	Pt 78 Platinum 195.08 $5d^9 6s^1$	Au 79 Gold 196.96654 $5d^{10} 6s^1$	Hg 80 Mercury 200.59 $5d^{10} 6s^2$	Tl 81 Thallium 204.3833 $6p^1$	Pb 82 Lead 207.2 $6p^2$	Bi 83 Bismuth 208.98037 $6p^3$	Po 84 Polonium (209) $6p^4$	At 85 Astatine (210) $6p^5$	Rn 86 Radon (222) $6p^6$
Fr 87 Francium (223) $7s^1$	Ra 88 Radium (226) $7s^2$	*89 - 103 Actinide series*	Unq 104 Unnilquadium (261) $6d^2 7s^2$	Unp 105 Unnilpentium (262) $6d^3 7s^2$	Unh 106 Unnilhexium (263) $6d^4 7s^2$	Uns 107 Unnilseptium (262)	108	109									

Lanthanide series	La 57 Lanthanum 138.9055 $5d^1 6s^2$	Ce 58 Cerium 140.115 $4f^1 5d^1 6s^2$	Pr 59 Praseodymium 140.90765 $4f^3 6s^2$	Nd 60 Neodymium 144.24 $4f^4 6s^2$	Pm 61 Promethium (145) $4f^5 6s^2$	Sm 62 Samarium 150.36 $4f^6 6s^2$	Eu 63 Europium 151.965 $4f^7 6s^2$	Gd 64 Gadolinium 157.25 $4f^7 5d^1 6s^2$	Tb 65 Terbium 158.92534 $4f^9 6s^2$	Dy 66 Dysprosium 162.50 $4f^{10} 6s^2$	Ho 67 Holmium 164.93032 $4f^{11} 6s^2$	Er 68 Erbium 167.26 $4f^{12} 6s^2$	Tm 69 Thulium 168.93421 $4f^{13} 6s^2$	Yb 70 Ytterbium 173.04 $4f^{14} 6s^2$	Lu 71 Lutetium 174.967 $4f^{14} 5d^1 6s^2$
Actinide series	Ac 89 Actinium (227) $6d^1 7s^2$	Th 90 Thorium 232.0381 $6d^2 7s^2$	Pa 91 Protactinium 231.03588 $5f^2 6d^1 7s^2$	U 92 Uranium 238.0289 $5f^3 6d^1 7s^2$	Np 93 Neptunium (237) $5f^4 6d^1 7s^2$	Pu 94 Plutonium (244) $5f^6 6d^0 7s^2$	Am 95 Americium (243) $5f^7 6d^0 7s^2$	Cm 96 Curium (247) $5f^7 6d^1 7s^2$	Bk 97 Berkelium (247) $5f^9 6d^0 7s^2$	Cf 98 Californium (251) $5f^{10} 6d^0 7s^2$	Es 99 Einsteinium (252) $5f^{11} 6d^0 7s^2$	Fm 100 Fermium (257) $5f^{12} 6d^0 7s^2$	Md 101 Mendelevium (258) $5f^{13} 6d^0 7s^2$	No 102 Nobelium (259) $6d^0 7s^2$	Lr 103 Lawrencium (260) $6d^1 7s^2$

Dmitri Mendeleev, a Russian chemist, was the first to develop a periodic table of the elements. Although Mendeleev organized his table according to atomic mass rather than atomic number, and so produced a table that was not quite as useful as modern periodic tables, his development stands as an excellent example of scientific proof. Seeing the patterns of periodicity (similar chemical properties according to atomic mass), Mendeleev hypothesized that all elements would fit into this ordered scheme. When he discovered "empty" spots in the table, he followed the logic of the existing order and hypothesized the existence of heretofore undiscovered elements. The subsequent discovery of those elements granted scientific legitimacy to Mendeleev's hypothesis, further discoveries leading

to the form of the periodic table we use today.

This is how science *should* work: hypotheses followed to their logical conclusions, and accepted, modified, or rejected as determined by the agreement of experimental data to those conclusions. Any fool can formulate a hypothesis after-the-fact to explain existing experimental data, and many do. What sets a scientific hypothesis apart from *post hoc* speculation is the prediction of future experimental data yet uncollected, and the possibility of disproof as a result of that data. To boldly follow a hypothesis to its logical conclusion(s) and dare to predict the results of future experiments is not a dogmatic leap of faith, but rather a public test of that hypothesis, open to challenge from anyone able to produce contradictory data. In other words, scientific hypotheses are always "risky" in the sense that they claim to predict the results of experiments not yet conducted, and are therefore susceptible to disproof if the experiments do not turn out as predicted. Thus, if a hypothesis successfully predicts the results of repeated experiments, there is little probability of its falsehood.

Quantum mechanics, first as a hypothesis and later as a theory, has proven to be extremely successful in predicting experimental results, hence the high degree of scientific confidence placed in it. Many scientists have reason to believe that it is an incomplete theory, though, as its predictions hold true more so at very small physical scales than at *macro*scopic dimensions, but nevertheless it is a tremendously useful theory in explaining and predicting the interactions of particles and atoms.

As you have already seen in this chapter, quantum physics is essential in describing and predicting many different phenomena. In the next section, we will see its significance in the electrical conductivity of solid substances, including semiconductors. Simply put, nothing in chemistry or solid-state physics makes sense within the popular theoretical framework of electrons existing as discrete chunks of matter, whirling around atomic nuclei like miniature satellites. It is only when electrons are viewed as "wavefunctions" existing in definite, discrete states that the regular and periodic behavior of matter can be explained.

- **REVIEW:**
- Electrons in atoms exist in "clouds" of distributed probability, not as discrete chunks of matter orbiting the nucleus like tiny satellites, as common illustrations of atoms show.
- Individual electrons around an atomic nucleus seek unique "states," described by four *quantum numbers*: the *Principal Quantum Number*, otherwise known as the *shell*; the *Angular Momentum Quantum Number*, otherwise known as the *subshell*; the *Magnetic Quantum Number*, describing the *orbital* (subshell orientation); and the *Spin Quantum Number*, or simply *spin*. These states are quantized, meaning that there are no "in-between" conditions for an electron other than those states that fit into the quantum numbering scheme.
- The *Principal Quantum Number* (n) describes the basic level or shell that an electron resides in. The larger this number, the greater radius the electron cloud has from the atom's nucleus, and the greater than electron's energy. Principal quantum numbers are whole numbers (positive integers).
- The *Angular Momentum Quantum Number* (l) describes the shape of the electron cloud within a particular shell or level, and is often known as the "subshell." There are as many subshells (electron cloud shapes) in any given shell as that shell's principal quantum number. Angular momentum quantum numbers are positive integers beginning at zero and terminating at one less than the principal quantum number (n-1).

- The *Magnetic Quantum Number* (m_l) describes which orientation a subshell (electron cloud shape) has. There are as many different orientations for each subshell as the subshell number (l) plus 1, and each unique orientation is called an *orbital*. These numbers are integers ranging from the negative value of the subshell number (l) through 0 to the positive value of the subshell number.

- The *Spin Quantum Number* (m_s) describes another property of an electron, and can be a value of +1/2 or -1/2.

- *Pauli's Exclusion Principle* says that no two electrons in an atom may share the exact same set of quantum numbers. Therefore, there is room for two electrons in each orbital (spin=1/2 and spin=-1/2), $2l+1$ orbitals in every subshell, and n subshells in every shell, and no more.

- *Spectroscopic notation* is a convention for denoting the electron configuration of an atom. Shells are shown as whole numbers, followed by subshell letters (s,p,d,f), with superscripted numbers totaling the number of electrons residing in each respective subshell.

- An atom's chemical behavior is solely determined by the electrons in the unfilled shells. Low-level shells that are completely filled have little or no effect on the chemical bonding characteristics of elements.

- Elements with completely filled electron shells are almost entirely unreactive, and are called *noble* (formerly known as *inert*).

2.3 Band theory of solids

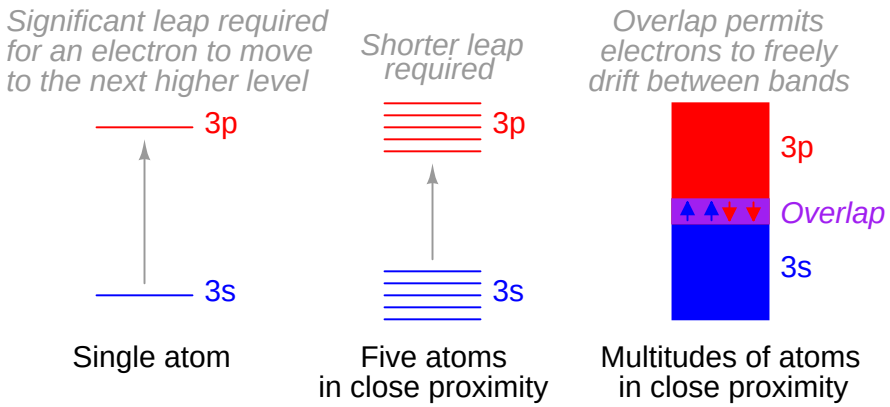
Quantum physics describes the states of electrons in an atom according to the four-fold scheme of *quantum numbers*. The quantum number system describes the *allowable states* electrons may assume in an atom. To use the analogy of an amphitheater, quantum numbers describe how many rows and seats there are. Individual electrons may be described by the combination of quantum numbers they possess, like a spectator in an amphitheater assigned to a particular row and seat.

Like spectators in an amphitheater moving between seats and/or rows, electrons may change their statuses, given the presence of available spaces for them to fit, and available energy. Since shell level is closely related to the amount of energy that an electron possesses, "leaps" between shell (and even subshell) levels requires transfers of energy. If an electron is to move into a higher-order shell, it requires that additional energy be given to the electron from an external source. Using the amphitheater analogy, it takes an increase in energy for a person to move into a higher row of seats, because that person must climb to a greater height against the force of gravity. Conversely, an electron "leaping" into a lower shell gives up some of its energy, like a person jumping down into a lower row of seats, the expended energy manifesting as heat and sound released upon impact.

Not all "leaps" are equal. Leaps between different shells requires a substantial exchange of energy, while leaps between subshells or between orbitals require lesser exchanges.

When atoms combine to form substances, the outermost shells, subshells, and orbitals merge, providing a greater number of available energy levels for electrons to assume. When large numbers of atoms exist in close proximity to each other, these available energy levels form a nearly continuous *band* wherein electrons may transition.

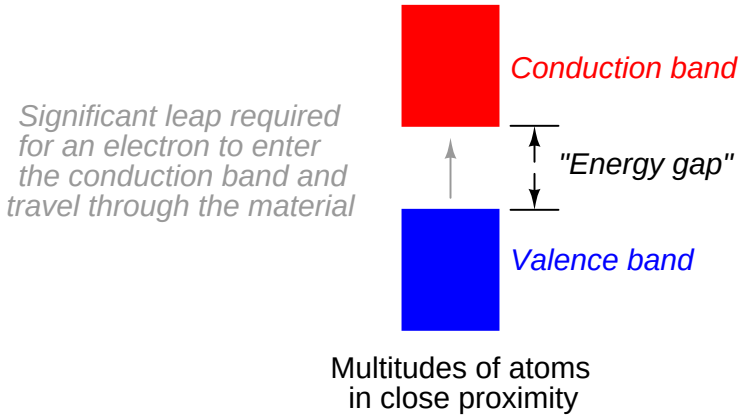
Electron band overlap in metallic elements



It is the width of these bands and their proximity to existing electrons that determines how mobile those electrons will be when exposed to an electric field. In metallic substances, empty bands overlap with bands containing electrons, meaning that electrons may move to what would normally be (in the case of a single atom) a higher-level state with little or no additional energy imparted. Thus, the outer electrons are said to be "free," and ready to move at the beckoning of an electric field.

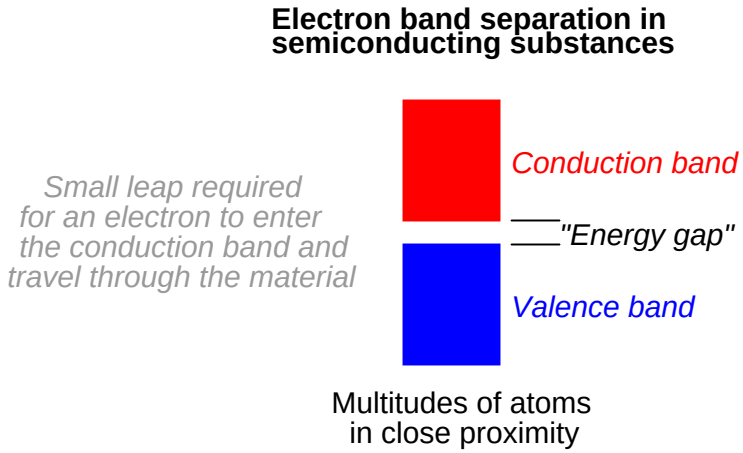
Band overlap will not occur in all substances, no matter how many atoms are in close proximity to each other. In some substances, a substantial gap remains between the highest band containing electrons (the so-called *valence band*) and the next band, which is empty (the so-called *conduction band*). As a result, valence electrons are "bound" to their constituent atoms and cannot become mobile within the substance without a significant amount of imparted energy. These substances are electrical insulators:

Electron band separation in insulating substances



Materials that fall within the category of *semiconductors* have a narrow gap between the valence and conduction bands. Thus, the amount of energy required to motivate a valence electron into the

conduction band where it becomes mobile is quite modest:



At low temperatures, there is little thermal energy available to push valence electrons across this gap, and the semiconducting material acts as an insulator. At higher temperatures, though, the ambient thermal energy becomes sufficient to force electrons across the gap, and the material will conduct electricity.

It is difficult to predict the conductive properties of a substance by examining the electron configurations of its constituent atoms. While it is true that the best metallic conductors of electricity (silver, copper, and gold) all have outer s subshells with a single electron, the relationship between conductivity and valence electron count is not necessarily consistent:

Element	Specific resistance (ρ) at 20° Celsius	Electron configuration
Silver (Ag)	9.546 Ω·cmil/ft	$4d^{10}5s^{1}$
Copper (Cu)	10.09 Ω·cmil/ft	$3d^{10}4s^{1}$
Gold (Au)	13.32 Ω·cmil/ft	$5d^{10}6s^{1}$
Aluminum (Al)	15.94 Ω·cmil/ft	$3p^{1}$
Tungsten (W)	31.76 Ω·cmil/ft	$5d^{4}6s^{2}$
Molybdenum (Mo)	32.12 Ω·cmil/ft	$4d^{5}5s^{1}$
Zinc (Zn)	35.49 Ω·cmil/ft	$3d^{10}4s^{2}$
Nickel (Ni)	41.69 Ω·cmil/ft	$3d^{8}4s^{2}$
Iron (Fe)	57.81 Ω·cmil/ft	$3d^{6}4s^{2}$
Platinum (Pt)	63.16 Ω·cmil/ft	$5d^{9}6s^{1}$

Likewise, the electron band configurations produced by compounds of different elements defies easy association with the electron configurations of its constituent elements.

- **REVIEW:**
-
-
-

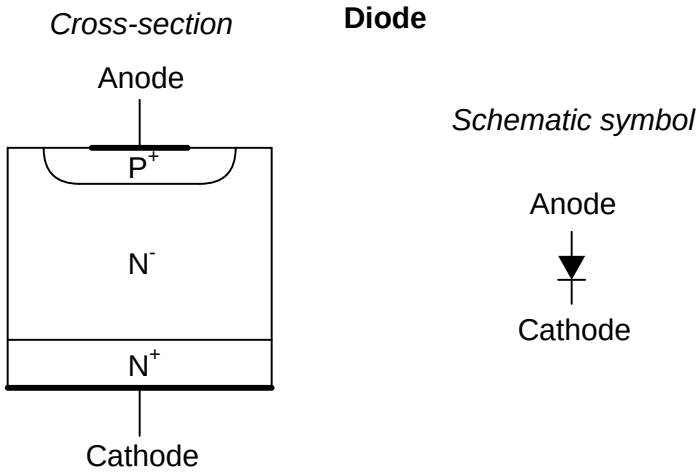
2.4 Electrons and "holes"

- **REVIEW:**
-
-
-

2.5 The P-N junction

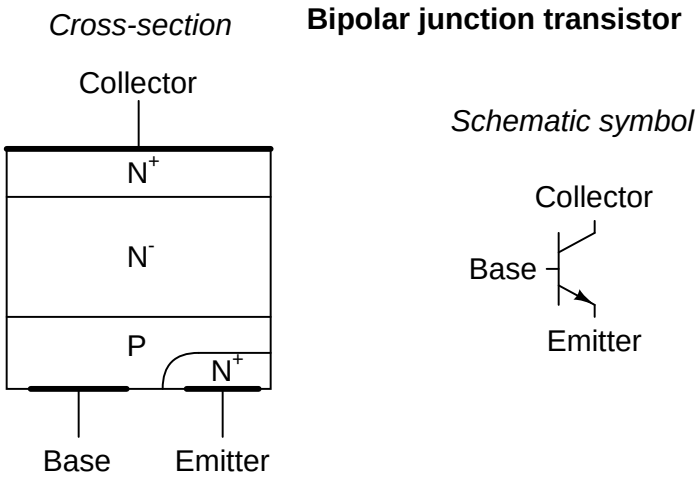
- **REVIEW:**
-
-
-

2.6 Junction diodes



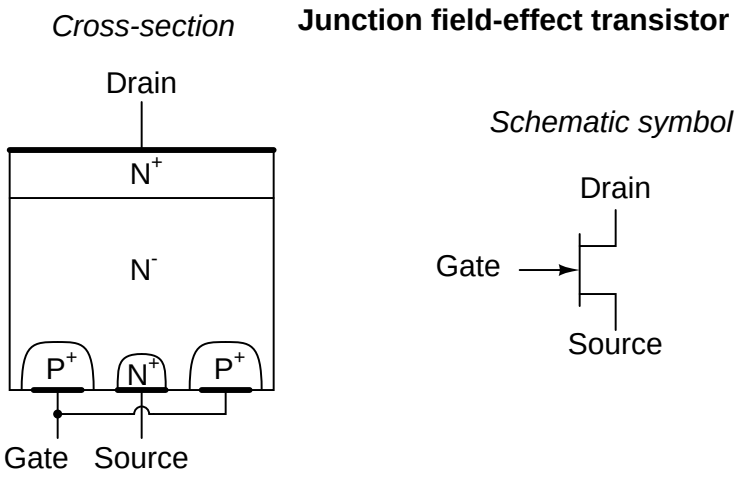
- **REVIEW:**
-
-
-

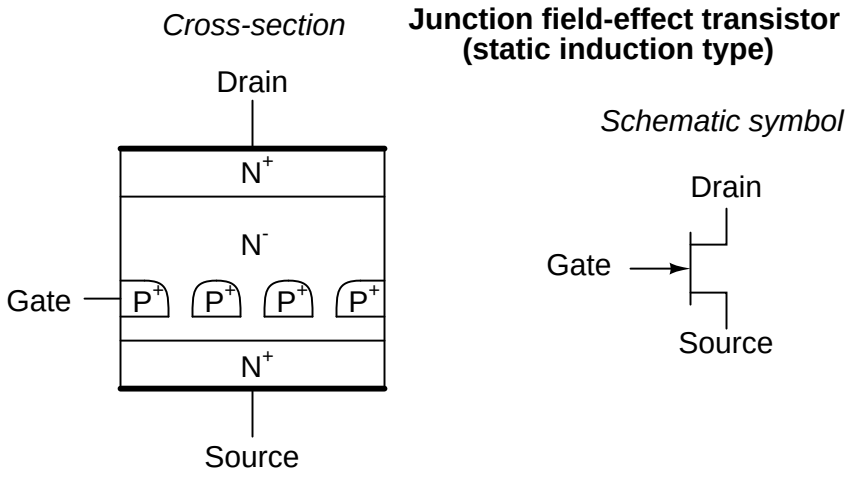
2.7 Bipolar junction transistors



- **REVIEW:**
-
-
-

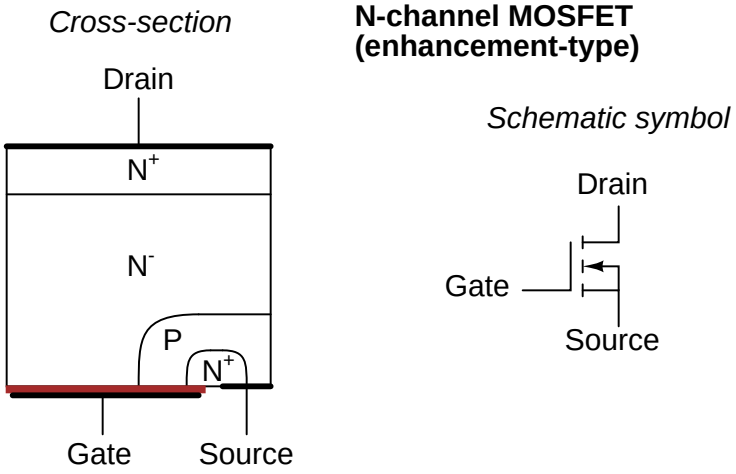
2.8 Junction field-effect transistors

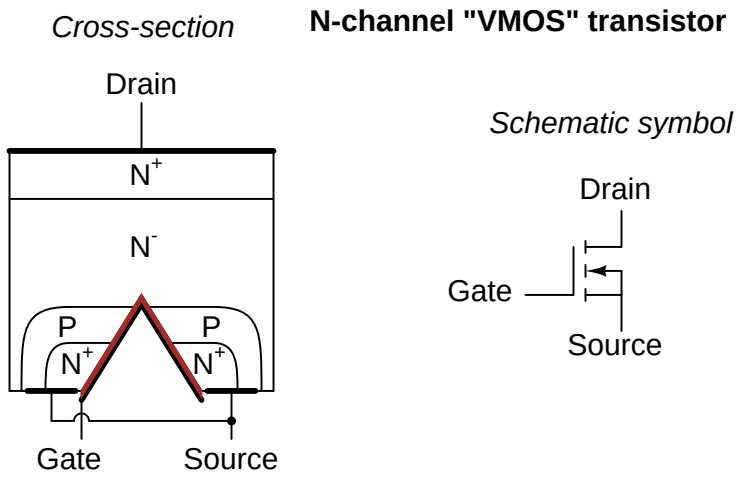




- **REVIEW:**
-
-
-

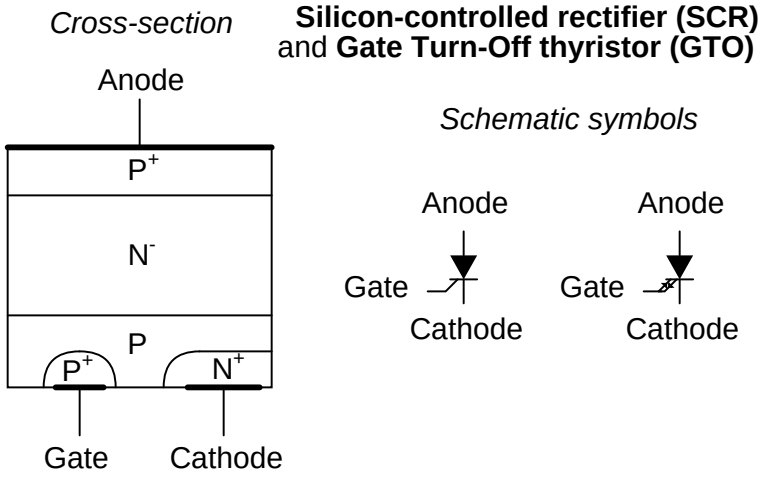
2.9 Insulated-gate field-effect transistors





- **REVIEW:**
-
-
-

2.10 Thyristors



- **REVIEW:**
-
-
-

2.11 Semiconductor manufacturing techniques

- **REVIEW:**
-
-
-

2.12 Superconducting devices

- **REVIEW:**
-
-
-

2.13 Quantum devices

- **REVIEW:**
-
-
-

2.14 Semiconductor devices in SPICE

- **REVIEW:**
-
-
-

2.15 Contributors

Contributors to this chapter are listed in chronological order of their contributions, from most recent to first. See Appendix 2 (Contributor List) for dates and contact information.

Maciej Noszczyski (December 2003): Corrected spelling of Niels Bohr's name.

Bill Heath (September 2002): Pointed out error in illustration of carbon atom – the nucleus was shown with seven protons instead of six.

Chapter 3

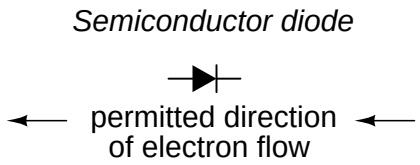
DIODES AND RECTIFIERS

Contents

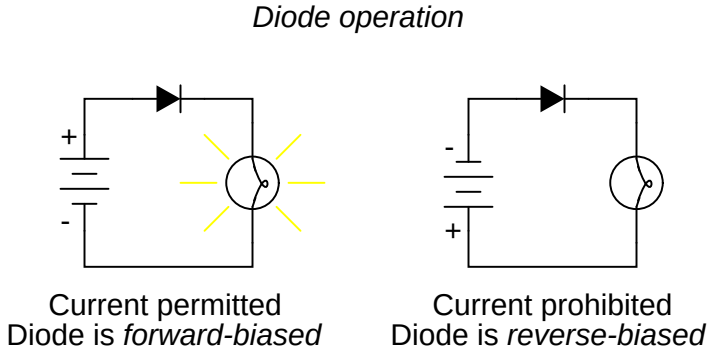
***** INCOMPLETE *****

3.1 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:



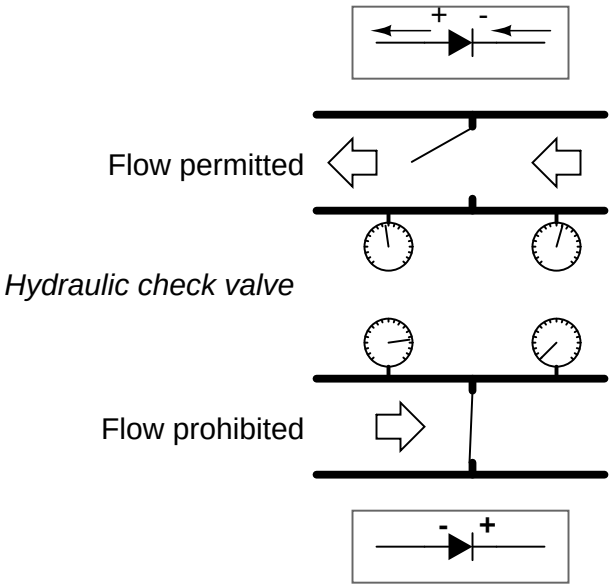
When placed in a simple battery-lamp circuit, the diode will either allow or prevent current through the lamp, depending on the polarity of the applied voltage:



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.

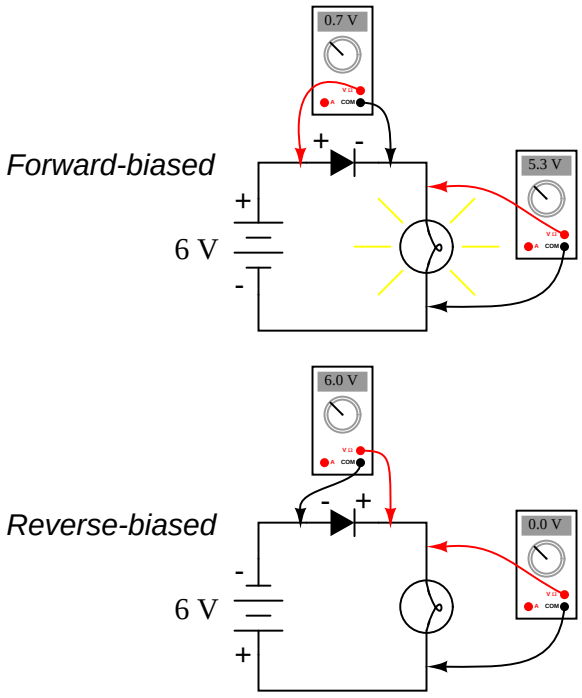
Oddly enough, the direction of the diode symbol's "arrowhead" points *against* the direction of electron flow. This is because the diode symbol was invented by engineers, who predominantly use *conventional flow* notation in their schematics, showing current as a flow of charge from the positive (+) 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:



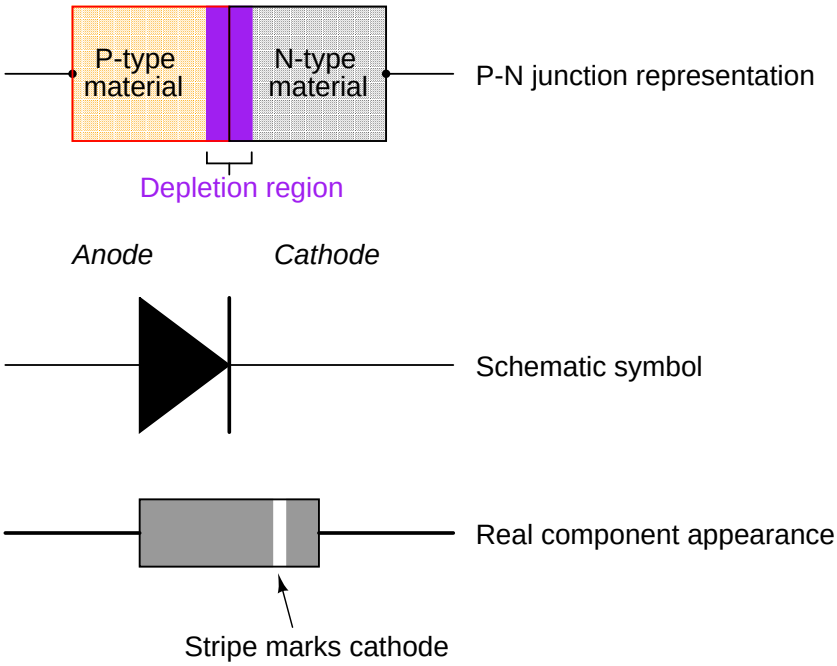
Check valves are essentially pressure-operated devices: they open and allow flow if the pressure across them is of the correct "polarity" to open the gate (in the analogy shown, greater fluid pressure on the right than on the left). If the pressure is of the opposite "polarity," the pressure difference 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:

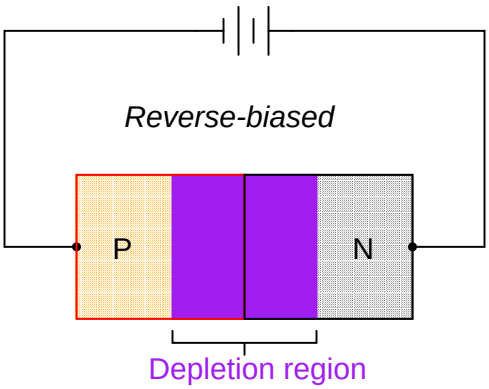


When the diode is forward-biased and conducting current, there is a small voltage dropped across it, leaving most of the battery voltage dropped across the lamp. When the battery's polarity is reversed and the diode becomes reverse-biased, it drops *all* of the battery's voltage and leaves none for the 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).

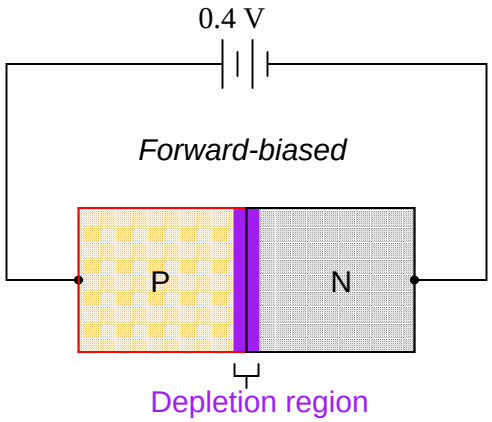
This forward-bias voltage drop exhibited by the diode is due to the action of the depletion region formed by the P-N junction under the influence of an applied voltage. When there is no voltage applied across a semiconductor diode, a thin depletion region exists around the region of the P-N junction,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

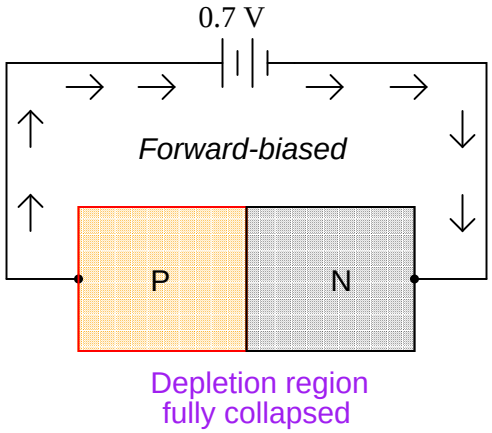


If a reverse-biasing voltage is applied across the P-N junction, this depletion region expands, further resisting any current through it:



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 x 10^{-12} amps)

e = Euler's constant (~ 2.718281828)

q = charge of electron (1.6 x 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 x 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 x 10^{-12} amps)

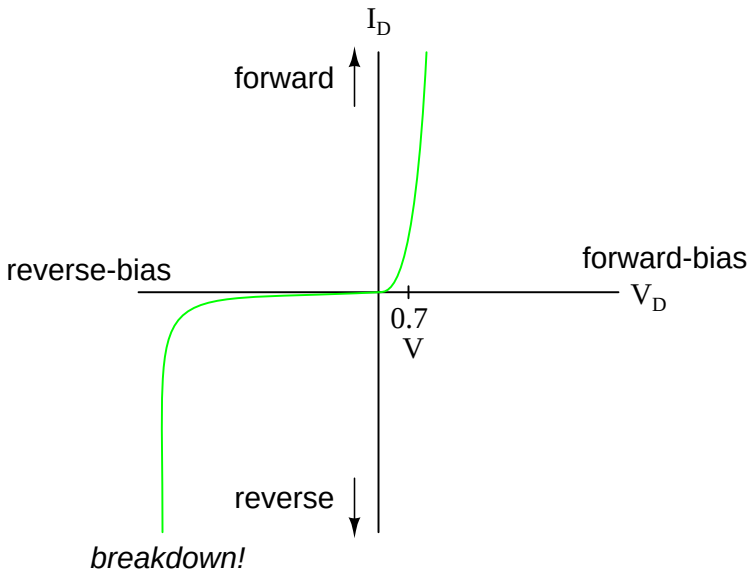
e = Euler's constant (~ 2.718281828)

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.

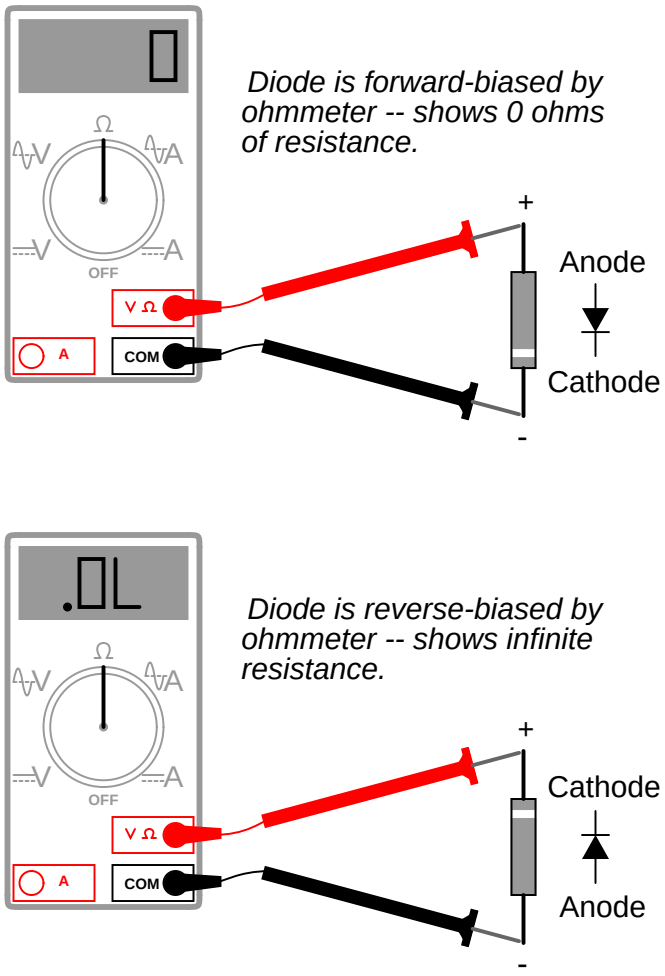


Typically, the PIV rating of a generic "rectifier" diode is at least 50 volts at room temperature. Diodes with PIV ratings in the many thousands of volts are available for modest prices.

- **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.

3.2 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):

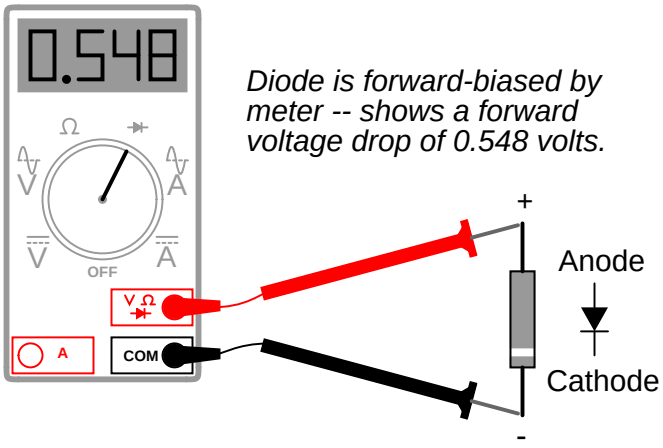


Of course, in order to determine which end of the diode is the cathode and which is the anode, you must know with certainty which test lead of the meter is positive (+) and which is negative (-) when set to the "resistance" or "Ω" function. With most digital multimeters I've seen, the red lead becomes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

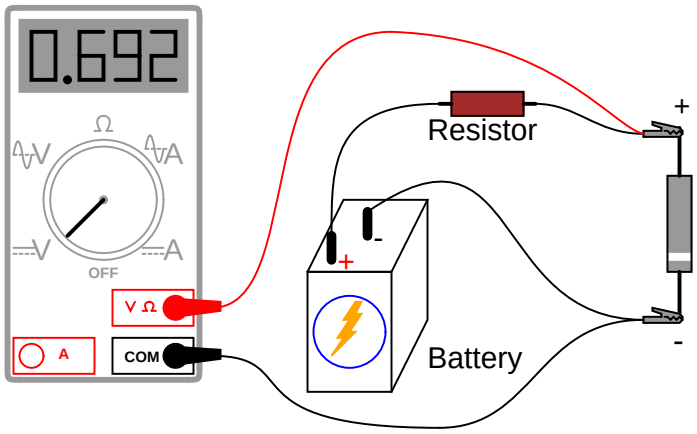
negative (-) when switched to the "resistance" function, because it is easier to manufacture it that way!

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.

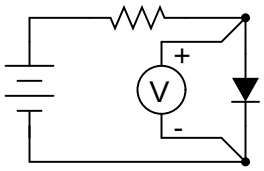
For this reason, some digital multimeter manufacturers equip their meters with a special "diode check" function which displays the actual forward voltage drop of the diode in volts, rather than a "resistance" figure in ohms. These meters work by forcing a small current through the diode and measuring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



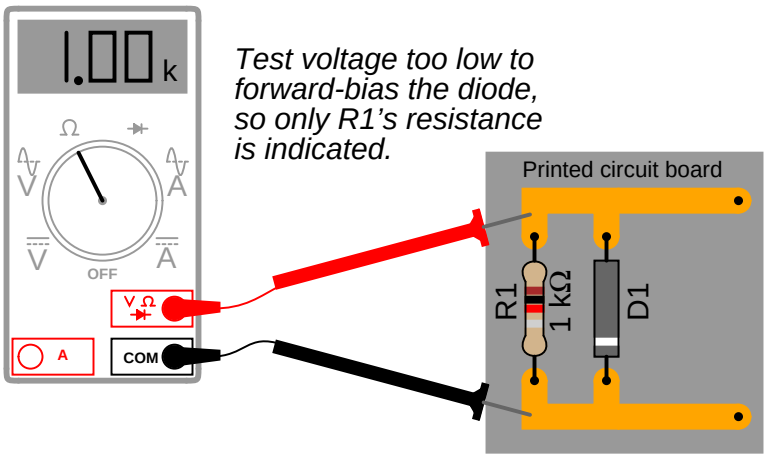
Resistor sized to obtain diode current of desired magnitude.

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

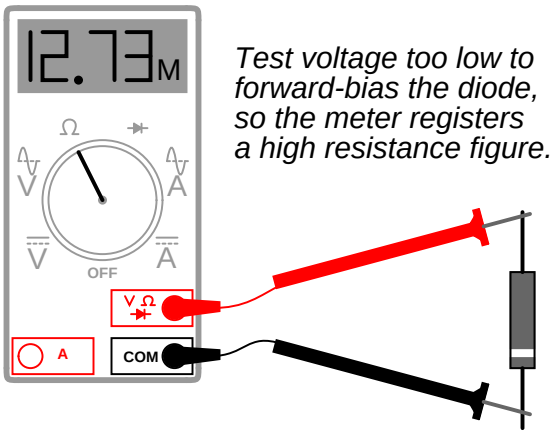
If this circuit were designed so as to provide a constant or nearly constant current through the diode despite changes in forward voltage drop, it could be used as the basis of a temperature-measurement instrument, the voltage measured across the diode being inversely proportional to diode junction temperature. 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.

3.3 Diode ratings

In addition to forward voltage drop (V_f) and peak inverse voltage (PIV), there are many other ratings of diodes important to circuit design and component selection. Semiconductor manufacturers provide detailed specifications on their products – diodes included – in publications known as *datasheets*. Datasheets for a wide variety of semiconductor components may be found in reference books and on the internet. I personally prefer the internet as a source of component specifications because all the data obtained from manufacturer websites are up-to-date.

A typical diode datasheet will contain figures for the following parameters:

Maximum repetitive reverse voltage = V_{RRM}, the maximum amount of voltage the diode can withstand in reverse-bias mode, in repeated pulses. Ideally, this figure would be infinite.

Maximum DC reverse voltage = V_R or V_{DC}, the maximum amount of voltage the diode can withstand in reverse-bias mode on a continual basis. Ideally, this figure would be infinite.

Maximum forward voltage = V_F, usually specified at the diode's rated forward current. Ideally, this figure would be zero: the diode providing no opposition whatsoever to forward current. In reality, the forward voltage is described by the "diode equation."

Maximum (average) forward current = $I_{F(AV)}$, the maximum average amount of current the diode is able to conduct in forward bias mode. This is fundamentally a thermal limitation: how much heat can the PN junction handle, given that dissipation power is equal to current (I) multiplied by voltage (V or E) and forward voltage is dependent upon both current and junction temperature. Ideally, this figure would be infinite.

Maximum (peak or surge) forward current = I_{FSM} or $i_{f(surge)}$, the maximum peak amount of current the diode is able to conduct in forward bias mode. Again, this rating is limited by the diode junction's thermal capacity, and is usually much higher than the average current rating due to thermal inertia (the fact that it takes a finite amount of time for the diode to reach maximum temperature for a given current). Ideally, this figure would be infinite.

Maximum total dissipation = P_D, the amount of power (in watts) allowable for the diode to dissipate, given the dissipation (P=IE) of diode current multiplied by diode voltage drop, and also the dissipation ($P=I^2R$) of diode current squared multiplied by bulk resistance. Fundamentally limited by the diode's thermal capacity (ability to tolerate high temperatures).

Operating junction temperature = T_J, the maximum allowable temperature for the diode's PN junction, usually given in degrees Celsius (^{o}C). Heat is the "Achilles' heel" of semiconductor devices: they *must* be kept cool to function properly and give long service life.

Storage temperature range = T_{STG}, the range of allowable temperatures for storing a diode (unpowered). Sometimes given in conjunction with operating junction temperature (T_J), because the maximum storage temperature and the maximum operating temperature ratings are often identical. If anything, though, maximum storage temperature rating will be greater than the maximum operating temperature rating.

Thermal resistance = $R(\Theta)$, the temperature difference between junction and outside air ($R(\Theta)_{JA}$) or between junction and leads ($R(\Theta)_{JL}$) for a given power dissipation. Expressed in units of degrees

Celsius per watt ($^oC/W$). Ideally, this figure would be zero, meaning that the diode package was a perfect thermal conductor and radiator, able to transfer all heat energy from the junction to the outside air (or to the leads) with no difference in temperature across the thickness of the diode package. A high thermal resistance means that the diode will build up excessive temperature at the junction (where it's critical) despite best efforts at cooling the outside of the diode, and thus will limit its maximum power dissipation.

Maximum reverse current = I_R, the amount of current through the diode in *reverse-bias* operation, with the maximum rated inverse voltage applied (V_{DC}). Sometimes referred to as *leakage current*. Ideally, this figure would be zero, as a perfect diode would block all current when reverse-biased. In reality, it is very small compared to the maximum forward current.

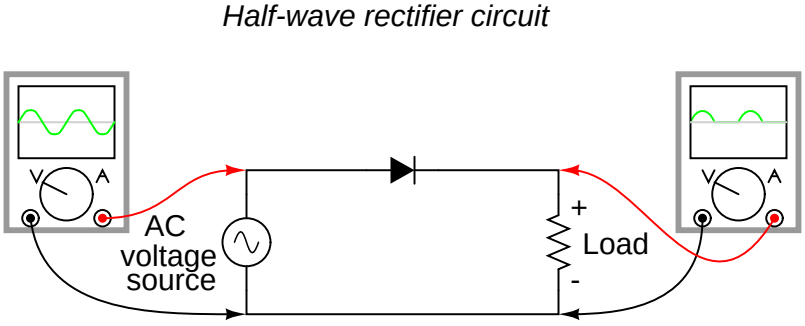
Typical junction capacitance = C_J, the typical amount of capacitance intrinsic to the junction, due to the depletion region acting as a dielectric separating the anode and cathode connections. This is usually a very small figure, measured in the range of picofarads (pF).

Reverse recovery time = t_{rr}, the amount of time it takes for a diode to "turn off" when the voltage across it alternates from forward-bias to reverse-bias polarity. Ideally, this figure would be zero: the diode halting conduction *immediately* upon polarity reversal. For a typical rectifier diode, reverse recovery time is in the range of tens of microseconds; for a "fast switching" diode, it may only be a few nanoseconds.

Most of these parameters vary with temperature or other operating conditions, and so a single figure fails to fully describe any given rating. Therefore, manufacturers provide graphs of component ratings plotted against other variables (such as temperature), so that the circuit designer has a better idea of what the device is capable of.

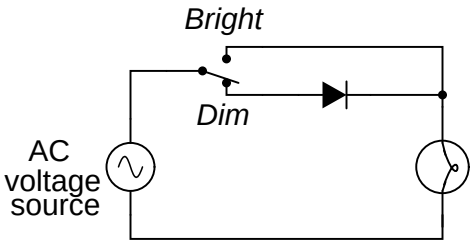
3.4 Rectifier circuits

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:



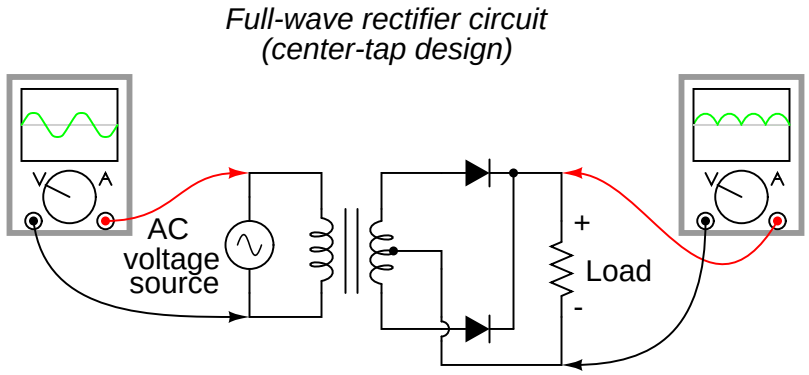
For most power applications, half-wave rectification is insufficient for the task. The harmonic content of the rectifier's output waveform is very large and consequently difficult to filter. Furthermore, AC power source only works to supply power to the load once every half-cycle, meaning that much of 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:

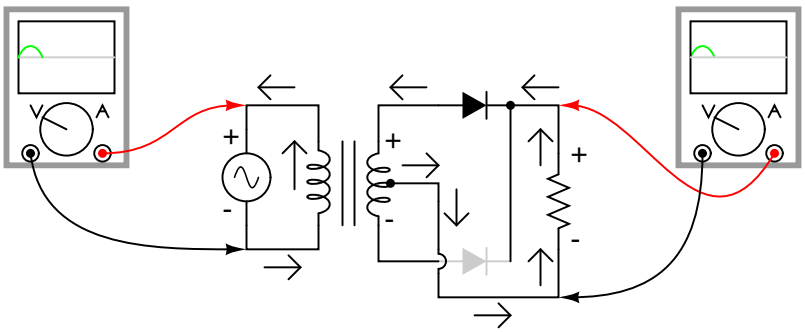


In the "Dim" switch position, the incandescent lamp receives approximately one-half the power it would normally receive operating on full-wave AC. Because the half-wave rectified power pulses far more rapidly than the filament has time to heat up and cool down, the lamp does not blink. Instead, its filament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

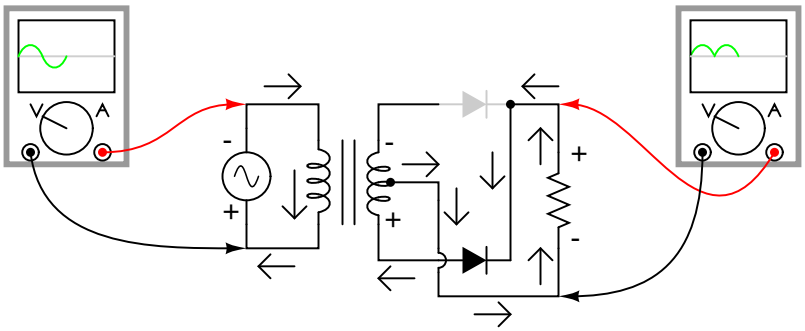
If we need to rectify AC power so as to obtain the full use of *both* half-cycles of the sine wave, a different rectifier circuit configuration must be used. Such a circuit is called a *full-wave* rectifier. One type of full-wave rectifier, called the *center-tap* design, uses a transformer with a center-tapped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

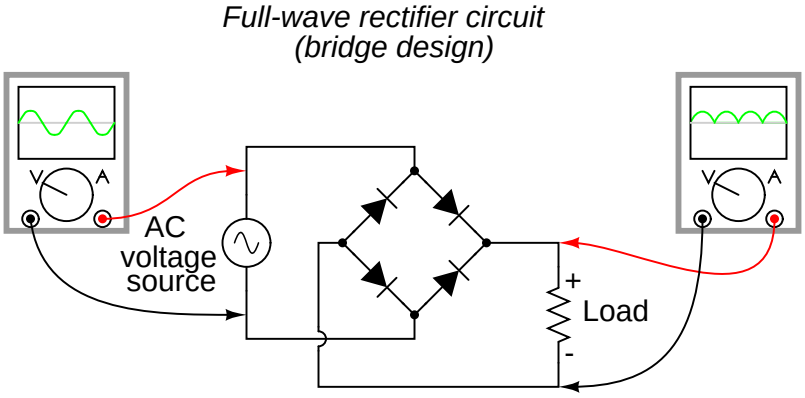


During the next half-cycle, the AC polarity reverses. Now, the other diode and the other half of the transformer's secondary winding carry current while the portions of the circuit formerly carrying current during the last half-cycle sit idle. The load still "sees" half of a sine wave, of the same polarity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

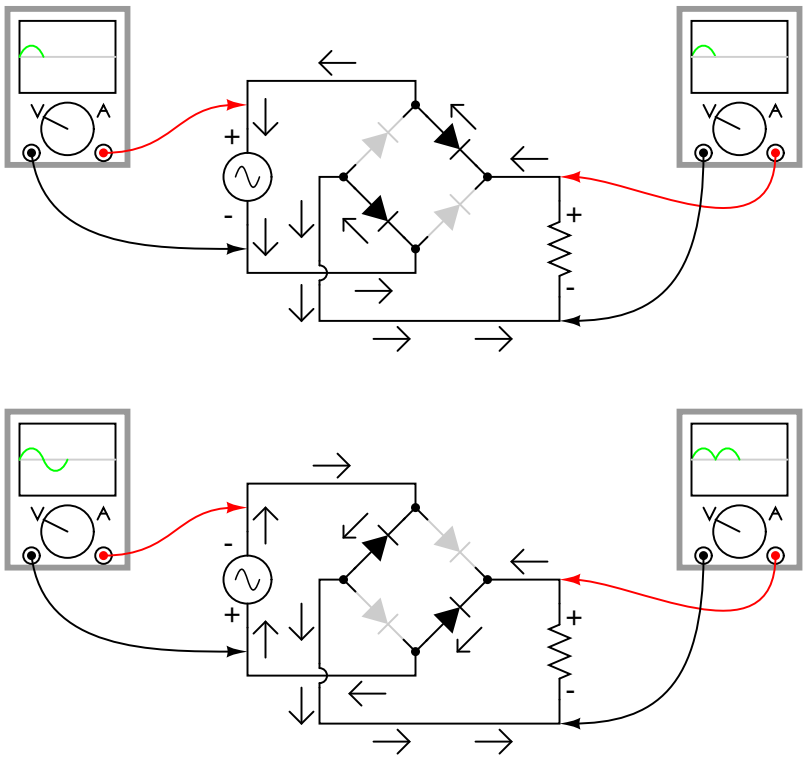


One disadvantage of this full-wave rectifier design is the necessity of a transformer with a center-tapped secondary winding. If the circuit in question is one of high power, the size and expense of a suitable transformer is significant. Consequently, the center-tap rectifier design is seen only in low-power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

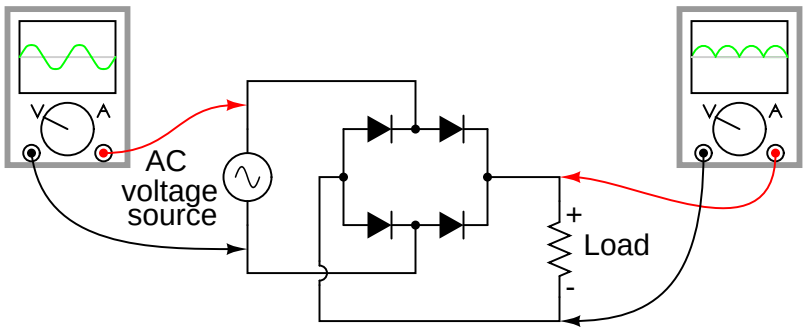


Current directions in the full-wave bridge rectifier circuit are as follows for each half-cycle of the AC waveform:

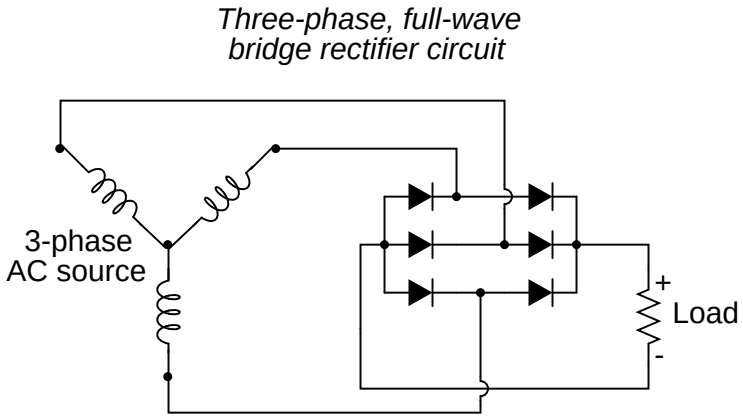


Remembering the proper layout of diodes in a full-wave bridge rectifier circuit can often be frustrating to the new student of electronics. I've found that an alternative representation of this circuit is easier both to remember and to comprehend. It's the exact same circuit, except all diodes are drawn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

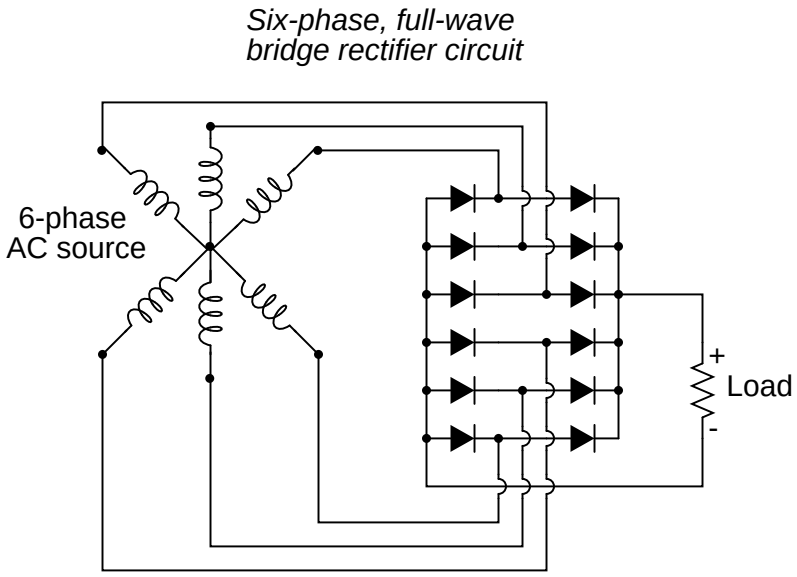
Full-wave bridge rectifier circuit
(alternative layout)



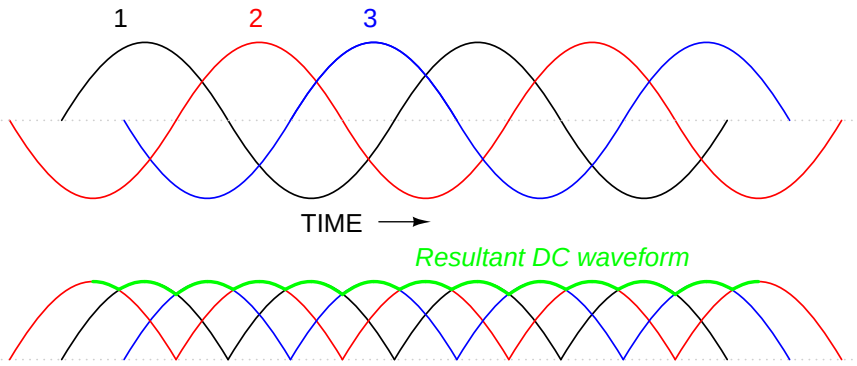
One advantage of remembering this layout for a bridge rectifier circuit is that it expands easily into a polyphase version:



Each three-phase line connects between a pair of diodes: one to route power to the positive (+) side of the load, and the other to route power to the negative (-) side of the load. Polyphase systems with more than three phases are easily accommodated into a bridge rectifier scheme. Take for instance this six-phase bridge rectifier circuit:



When polyphase AC is rectified, the phase-shifted pulses overlap each other to produce a DC output that is much "smoother" (has less AC content) than that produced by the rectification of single-phase AC. This is a decided advantage in high-power rectifier circuits, where the sheer physical size of filtering components would be prohibitive but low-noise DC power must be obtained. The following diagram shows the full-wave rectification of three-phase AC:



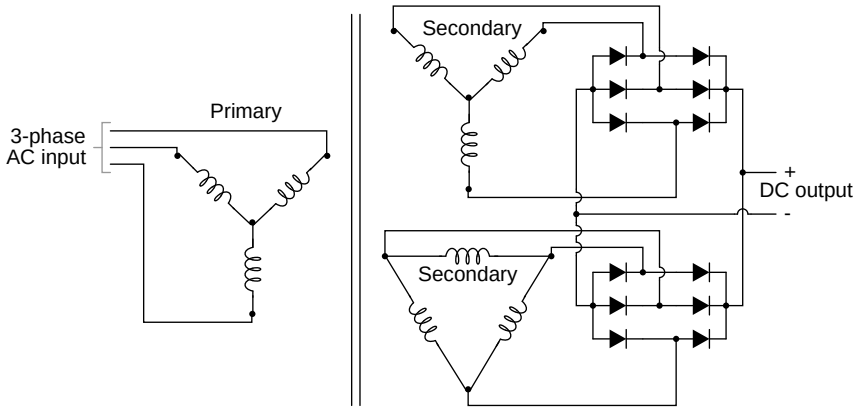
In any case of rectification – single-phase or polyphase – the amount of AC voltage mixed with the rectifier's DC output is called *ripple voltage*. In most cases, since "pure" DC is the desired goal, ripple voltage is undesirable. If the power levels are not too great, filtering networks may be employed to reduce the amount of ripple in the output voltage.

Sometimes, the method of rectification is referred to by counting the number of DC "pulses" output for every 360^o of electrical "rotation." A single-phase, half-wave rectifier circuit, then, would be called a *1-pulse* rectifier, because it produces a single pulse during the time of one complete cycle (360^o) of the AC waveform. A single-phase, full-wave rectifier (regardless of design, center-tap or bridge) would be called a *2-pulse* rectifier, because it outputs two pulses of DC during one AC cycle's worth of time. A three-phase full-wave rectifier would be called a *6-pulse* unit.

Modern electrical engineering convention further describes the function of a rectifier circuit by using a three-field notation of *phases*, *ways*, and number of *pulses*. A single-phase, half-wave rectifier circuit is given the somewhat cryptic designation of 1Ph1W1P (1 phase, 1 way, 1 pulse), meaning that the AC supply voltage is single-phase, that current on each phase of the AC supply lines moves in one direction (way) only, and that there is a single pulse of DC produced for every 360^o of electrical rotation. A single-phase, full-wave, center-tap rectifier circuit would be designated as 1Ph1W2P in this notational system: 1 phase, 1 way or direction of current in each winding half, and 2 pulses or output voltage per cycle. A single-phase, full-wave, bridge rectifier would be designated as 1Ph2W2P: the same as for the center-tap design, except current can go *both* ways through the AC lines instead of just one way. The three-phase bridge rectifier circuit shown earlier would be called a 3Ph2W6P rectifier.

Is it possible to obtain more pulses than twice the number of phases in a rectifier circuit? The answer to this question is yes: especially in polyphase circuits. Through the creative use of transformers, sets of full-wave rectifiers may be paralleled in such a way that more than six pulses of DC are produced for three phases of AC. A 30^o phase shift is introduced from primary to secondary of a three-phase transformer when the winding configurations are not of the same type. In other words, a transformer connected either Y-Δ or Δ-Y will exhibit this 30^o phase shift, while a transformer connected Y-Y or Δ-Δ will not. This phenomenon may be exploited by having one transformer connected Y-Y feed a bridge rectifier, and have another transformer connected Y-Δ feed a second bridge rectifier, then parallel the DC outputs of both rectifiers. Since the ripple voltage waveforms of the two rectifiers' outputs are phase-shifted 30^o from one another, their superposition results in less ripple than either rectifier output considered separately: 12 pulses per 360^o instead of just six:

3Ph2W12P rectifier circuit



- **REVIEW:**
- *Rectification* is the conversion of alternating current (AC) to direct current (DC).
- A *half-wave* rectifier is a circuit that allows only one half-cycle of the AC voltage waveform to be applied to the load, resulting in one non-alternating polarity across it. The resulting DC delivered to the load "pulsates" significantly.
- A *full-wave* rectifier is a circuit that converts both half-cycles of the AC voltage waveform to an unbroken series of voltage pulses of the same polarity. The resulting DC delivered to the load doesn't "pulsate" as much.
- Polyphase alternating current, when rectified, gives a much "smoother" DC waveform (less *ripple* voltage) than rectified single-phase AC.

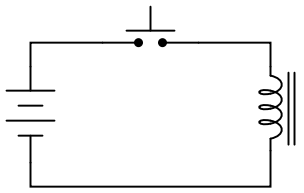
3.5 Clipper circuits

3.6 Clamper circuits

3.7 Voltage multipliers

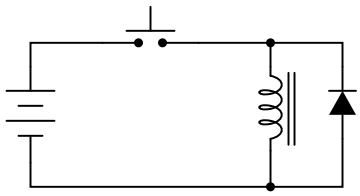
3.8 Inductor commutating circuits

A popular use of diodes is for the mitigation of inductive "kickback:" the pulses of high voltage produced when direct current through an inductor is interrupted. Take for example this simple circuit:

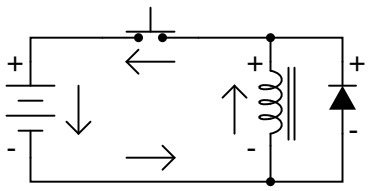


When the pushbutton switch is actuated, current goes through the inductor, producing a magnetic field around it. When the switch is de-actuated, its contacts open, interrupting current through the inductor, and causing the magnetic field to rapidly collapse. Because the voltage induced in a coil of wire is directly proportional to the *rate of change* over time of magnetic flux (Faraday's Law: $e = N d\Phi/dt$), this rapid collapse of magnetism around the coil produces a high voltage "spike."

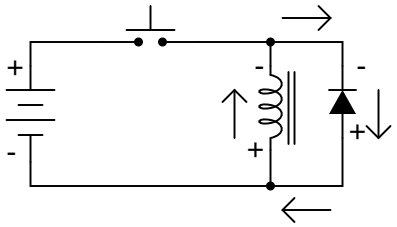
If the inductor in question is an electromagnet coil, such as might be seen in a solenoid or relay (constructed for the purpose of creating a physical force via its magnetic field when energized), the effect of inductive "kickback" serves no useful purpose at all. In fact, it is quite detrimental to the switch, as it will cause excessive arcing at the contacts, greatly reducing their service life. There are several practical methods of mitigating the high voltage transient created when the switch is opened, but none so simple as the so-called *commutating diode*:



In this circuit, the diode is placed in parallel with the coil, in such a way that it will be reverse-biased when DC voltage is applied to the coil through the switch. Thus, when the coil is energized, the diode conducts no current:



However, when the switch is opened, the coil's inductance responds to the decrease in current by inducing a voltage of reverse polarity, in an effort to maintain current at the same magnitude and in the same direction. This sudden reversal of voltage polarity across the coil forward-biases the diode, and the diode provides a current path for the inductor's current, so that its stored energy is dissipated slowly rather than suddenly:



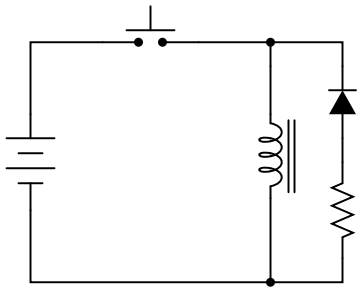
As a result, the voltage induced in the coil by its collapsing magnetic field is quite low: merely the forward voltage drop of the diode, rather than hundreds of volts as before. Thus, the switch contacts experience a voltage drop equal to the battery voltage plus about 0.7 volts (if the diode is silicon) during this discharge time.

In electronics parlance, *commutation* refers to the reversal of voltage polarity or current direction. Thus, the purpose of a *commutating diode* is to act whenever voltage reverses polarity, in this case, the voltage induced by the inductor coil when current through it is interrupted by the switch. A less formal term for a commutating diode is *snubber*, because it "snubs" or "squelches" the inductive kickback.

A noteworthy disadvantage of this method is the extra time it imparts to the coil's discharge. Because the induced voltage is clamped to a very low value, its rate of magnetic flux change over time is comparatively slow. Remember that Faraday's Law describes the magnetic flux rate-of-change ($d\Phi/dt$) as being proportional to the induced, instantaneous voltage (e or v). If the instantaneous voltage is limited to some low figure, then the rate of change of magnetic flux over time will likewise be limited to a low (slow) figure.

If an electromagnet coil is "snubbed" with a commutating diode, the magnetic field will dissipate at a relatively slow rate compared to the original scenario (no diode) where the field disappeared almost instantly upon switch release. The amount of time in question will most likely be less than one second, but it will be measurably slower than without a commutating diode in place. This may be an intolerable consequence if the coil is used to actuate an electromechanical relay, because the relay will possess a natural "time delay" upon coil de-energization, and an unwanted delay of even a fraction of a second may wreak havoc in some circuits.

Unfortunately, there is no way to eliminate the high-voltage transient of inductive kickback *and* maintain fast de-magnetization of the coil: Faraday's Law will not be violated. However, if slow de-magnetization is unacceptable, a compromise may be struck between transient voltage and time by allowing the coil's voltage to rise to some higher level (but not so high as without a commutating diode in place). The following schematic shows how this may be done:

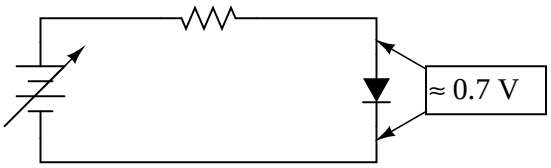


A resistor placed in series with the commutating diode allows the coil's induced voltage to rise to

a level greater than the diode's forward voltage drop, thus hastening the process of de-magnetization. This, of course, will place the switch contacts under greater stress, and so the resistor must be sized to limit that transient voltage at an acceptable maximum level.

3.9 Zener diodes

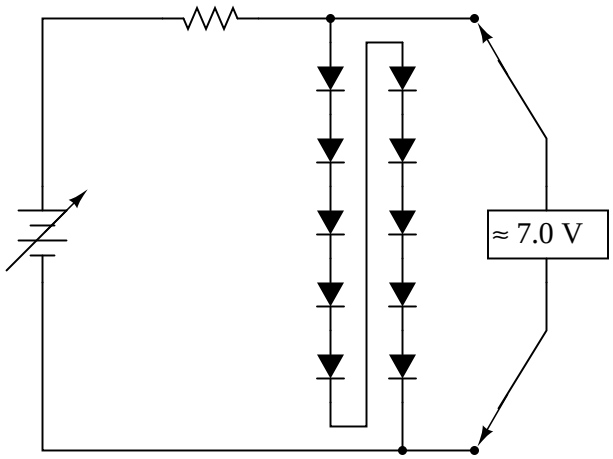
If we connect a diode and resistor in series with a DC voltage source so that the diode is forward-biased, the voltage drop across the diode will remain fairly constant over a wide range of power supply voltages:



According to the "diode equation," the current through a forward-biased PN junction is proportional to e raised to the power of the forward voltage drop. Because this is an exponential function, current rises quite rapidly for modest increases in voltage drop. Another way of considering this is to say that voltage dropped across a forward-biased diode changes little for large variations in diode current. In the circuit shown above, diode current is limited by the voltage of the power supply, the series resistor, and the diode's voltage drop, which as we know doesn't vary much from 0.7 volts. If the power supply voltage were to be increased, the resistor's voltage drop would increase almost the same amount, and the diode's voltage drop just a little. Conversely, a decrease in power supply voltage would result in an almost equal decrease in resistor voltage drop, with just a little decrease in diode voltage drop. In a word, we could summarize this behavior by saying that the diode is *regulating* the voltage drop at approximately 0.7 volts.

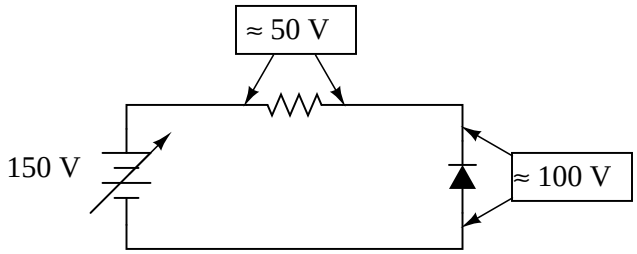
Voltage regulation is a useful diode property to exploit. Suppose we were building some kind of circuit which could not tolerate variations in power supply voltage, but needed to be powered by a chemical battery, whose voltage changes over its lifetime. We could form a circuit as shown and connect the circuit requiring steady voltage across the diode, where it would receive an unchanging 0.7 volts.

This would certainly work, but most practical circuits of any kind require a power supply voltage in excess of 0.7 volts to properly function. One way we could increase our voltage regulation point would be to connect multiple diodes in series, so that their individual forward voltage drops of 0.7 volts each would add to create a larger total. For instance, if we had ten diodes in series, the regulated voltage would be ten times 0.7, or 7 volts:



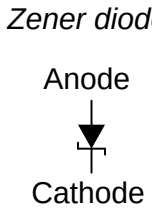
So long as the battery voltage never sagged below 7 volts, there would always be about 7 volts dropped across the ten-diode "stack."

If larger regulated voltages are required, we could either use more diodes in series (an inelegant option, in my opinion), or try a fundamentally different approach. We know that diode forward voltage is a fairly constant figure under a wide range of conditions, but so is *reverse breakdown voltage*, and breakdown voltage is typically much, much greater than forward voltage. If we reversed the polarity of the diode in our single-diode regulator circuit and increased the power supply voltage to the point where the diode "broke down" (could no longer withstand the reverse-bias voltage impressed across it), the diode would similarly regulate the voltage at that breakdown point, not allowing it to increase further:



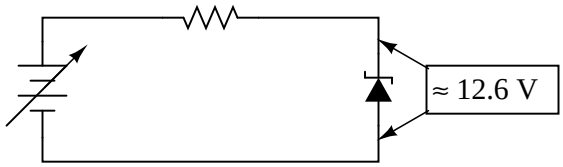
Diode $V_{breakdown}$ = 100 V

Unfortunately, when normal rectifying diodes "break down," they usually do so destructively. However, it is possible to build a special type of diode that can handle breakdown without failing completely. This type of diode is called a *zener diode*, and its symbol looks like this:



When forward-biased, zener diodes behave much the same as standard rectifying diodes: they have a forward voltage drop which follows the "diode equation" and is about 0.7 volts. In reverse-bias mode, they do not conduct until the applied voltage reaches or exceeds the so-called *zener voltage*, at which point the diode is able to conduct substantial current, and in doing so will try to limit the voltage dropped across it to that zener voltage point. So long as the power dissipated by this reverse current does not exceed the diode's thermal limits, the diode will not be harmed.

Zener diodes are manufactured with zener voltages ranging anywhere from a few volts to hundreds of volts. This zener voltage changes slightly with temperature, and like common carbon-composition resistor values, may be anywhere from 5 percent to 10 percent in error from the manufacturer's specifications. However, this stability and accuracy is generally good enough for the zener diode to be used as a voltage regulator device in common power supply circuit:

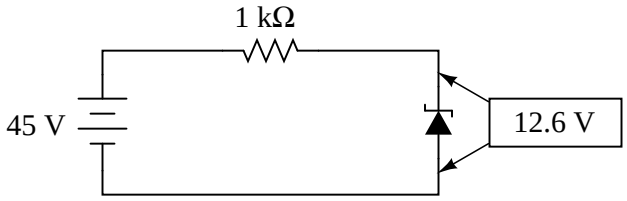


Diode V_{zener} = 12.6 V

Please take note of the zener diode's orientation in the above circuit: the diode is *reverse-biased*, and intentionally so. If we had oriented the diode in the "normal" way, so as to be forward-biased, it would only drop 0.7 volts, just like a regular rectifying diode. If we want to exploit this diode's reverse breakdown properties, we must operate it in its reverse-bias mode. So long as the power supply voltage remains above the zener voltage (12.6 volts, in this example), the voltage dropped across the zener diode will remain at approximately 12.6 volts.

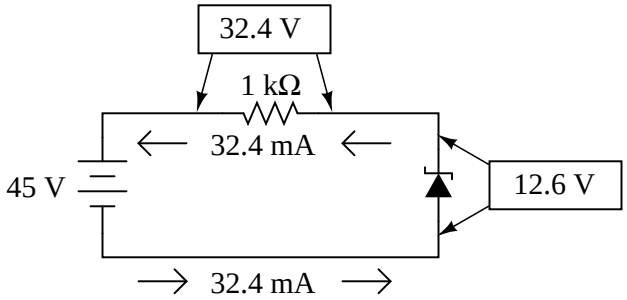
Like any semiconductor device, the zener diode is sensitive to temperature. Excessive temperature will destroy a zener diode, and because it both drops voltage and conducts current, it produces its own heat in accordance with Joule's Law (P=IE). Therefore, one must be careful to design the regulator circuit in such a way that the diode's power dissipation rating is not exceeded. Interestingly enough, when zener diodes fail due to excessive power dissipation, they usually fail *shorted* rather than open. A diode failed in this manner is easy to detect: it drops almost zero voltage when biased either way, like a piece of wire.

Let's examine a zener diode regulating circuit mathematically, determining all voltages, currents, and power dissipations. Taking the same form of circuit shown earlier, we'll perform calculations assuming a zener voltage of 12.6 volts, a power supply voltage of 45 volts, and a series resistor value of 1000 Ω (we'll regard the zener voltage to be *exactly* 12.6 volts so as to avoid having to qualify all figures as "approximate"):



Diode V_{zener} = 12.6 V

If the zener diode's voltage is 12.6 volts and the power supply's voltage is 45 volts, there will be 32.4 volts dropped across the resistor (45 volts - 12.6 volts = 32.4 volts). 32.4 volts dropped across 1000 Ω gives 32.4 mA of current in the circuit:



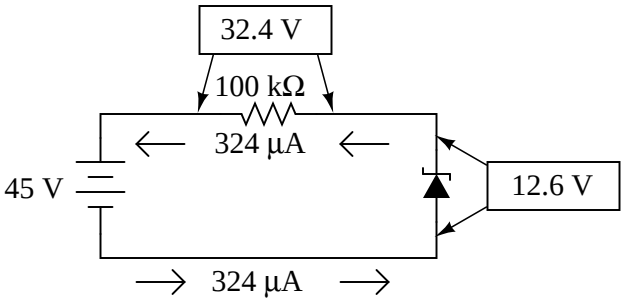
Power is calculated by multiplying current by voltage (P=IE), so we can calculate power dissipations for both the resistor and the zener diode quite easily:

$$P_{resistor} = (32.4\ mA)(32.4\ V)$$
$$P_{resistor} = 1.0498\ W$$

$$P_{diode} = (32.4\ mA)(12.6\ V)$$
$$P_{diode} = 408.24\ mW$$

A zener diode with a power rating of 0.5 watt would be adequate, as would a resistor rated for 1.5 or 2 watts of dissipation.

If excessive power dissipation is detrimental, then why not design the circuit for the least amount of dissipation possible? Why not just size the resistor for a very high value of resistance, thus severely limiting current and keeping power dissipation figures very low? Take this circuit, for example, with a 100 kΩ resistor instead of a 1 kΩ resistor. Note that both the power supply voltage and the diode's zener voltage are identical to the last example:



With only 1/100 of the current we had before (324 μA instead of 32.4 mA), both power dissipation figures should be 100 times smaller:

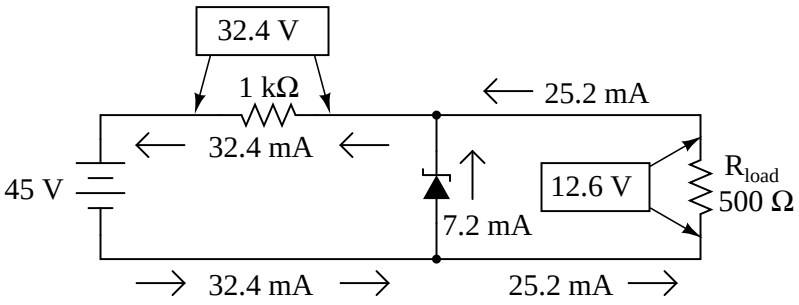
$$P_{resistor} = (324\ \mu A)(32.4\ V)$$

$$P_{resistor} = 10.498\ mW$$

$$P_{diode} = (324\ \mu A)(12.6\ V)$$

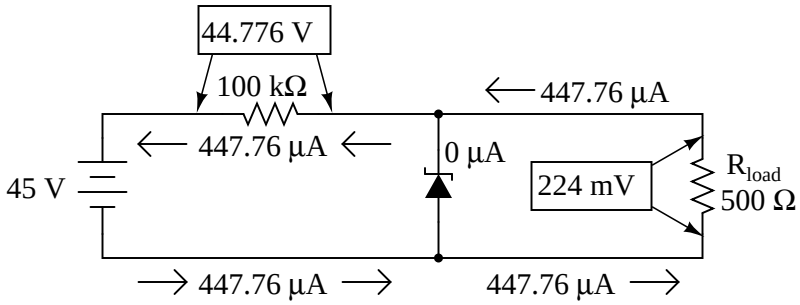
$$P_{diode} = 4.0824\ mW$$

Seems ideal, doesn't it? Less power dissipation means lower operating temperatures for both the diode and the resistor, and also less wasted energy in the system, right? A higher resistance value *does* reduce power dissipation levels in the circuit, but it unfortunately introduces another problem. Remember that the purpose of a regulator circuit is to provide a stable voltage *for another circuit*. In other words, we're eventually going to power something with 12.6 volts, and this something will have a current draw of its own. Consider our first regulator circuit, this time with a 500 Ω load connected in parallel with the zener diode:



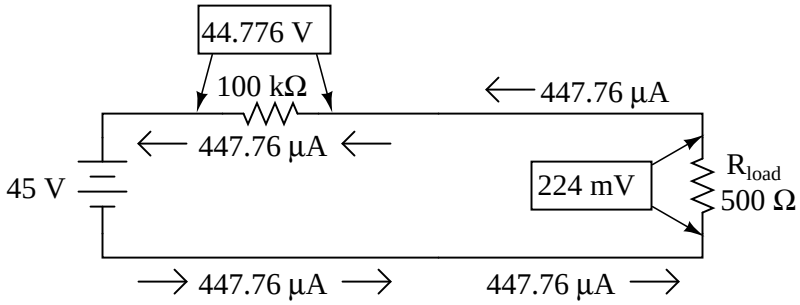
If 12.6 volts is maintained across a 500 Ω load, the load will draw 25.2 mA of current. In order for the 1 kΩ series "dropping" resistor to drop 32.4 volts (reducing the power supply's voltage of 45 volts down to 12.6 across the zener), it still must conduct 32.4 mA of current. This leaves 7.2 mA of current through the zener diode.

Now consider our "power-saving" regulator circuit with the 100 kΩ dropping resistor, delivering power to the same 500 Ω load. What it is supposed to do is maintain 12.6 volts across the load, just like the last circuit. However, as we will see, it *cannot* accomplish this task:



With the larger value of dropping resistor in place, there will only be about 224 mV of voltage across the 500 Ω load, far less than the expected value of 12.6 volts! Why is this? If we actually had 12.6 volts across the load, it would draw 25.2 mA of current, as before. This load current would have to go through the series dropping resistor as it did before, but with a new (much larger!) dropping resistor in place, the voltage dropped across that resistor with 25.2 mA of current going through it would be 2,520 volts! Since we obviously don't have that much voltage supplied by the battery, this cannot happen.

The situation is easier to comprehend if we temporarily remove the zener diode from the circuit and analyze the behavior of the two resistors alone:



Both the 100 kΩ dropping resistor and the 500 Ω load resistance are in series with each other, giving a total circuit resistance of 100.5 kΩ. With a total voltage of 45 volts and a total resistance of 100.5 kΩ, Ohm's Law (I=E/R) tells us that the current will be 447.76 μA. Figuring voltage drops across both resistors (E=IR), we arrive at 44.776 volts and 224 mV, respectively. If we were to re-install the zener diode at this point, it would "see" 224 mV across it as well, being in parallel with the load resistance. This is far below the zener breakdown voltage of the diode and so it will not "break down" and conduct current. For that matter, at this low voltage the diode wouldn't conduct even if it were forward-biased! Thus, the diode ceases to regulate voltage, for it can do so only when there is at least 12.6 volts dropped across to "activate" it.

The analytical technique of removing a zener diode from a circuit and seeing whether or not there is enough voltage present to make it conduct is a sound one. Just because a zener diode happens to be connected in a circuit doesn't guarantee that the full zener voltage will always be dropped across it! Remember that zener diodes work by *limiting* voltage to some maximum level; they cannot *make up* for a lack of voltage.

In summary, any zener diode regulating circuit will function so long as the load's resistance is equal to or greater than some minimum value. If the load resistance is too low, it will draw too much

current, dropping too much voltage across the series dropping resistor, leaving insufficient voltage across the zener diode to make it conduct. When the zener diode stops conducting current, it can no longer regulate voltage, and the load voltage will fall below the regulation point.

Our regulator circuit with the 100 kΩ dropping resistor must be good for some value of load resistance, though. To find this acceptable load resistance value, we can use a table to calculate resistance in the two-resistor series circuit (no diode), inserting the known values of total voltage and dropping resistor resistance, and calculating for an expected load voltage of 12.6 volts:

	$R_{dropping}$	R_{load}	Total	
E		12.6	45	Volts
I				Amps
R	100 k			Ohms

With 45 volts of total voltage and 12.6 volts across the load, we should have 32.4 volts across $R_{dropping}$:

	$R_{dropping}$	R_{load}	Total	
E	**32.4**	12.6	45	Volts
I				Amps
R	100 k			Ohms

With 32.4 volts across the dropping resistor, and 100 kΩ worth of resistance in it, the current through it will be 324 μA:

	$R_{dropping}$	R_{load}	Total	
E	32.4	12.6	45	Volts
I	**324** μ			Amps
R	100 k			Ohms

Ohm's Law

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

Being a series circuit, the current is equal through all components at any given time:

	$R_{dropping}$	R_{load}	Total	
E	32.4	12.6	45	Volts
I	324 µ	**324** µ	**324** µ	Amps
R	100 k			Ohms

Rule of series circuits:

$I_{Total} = I_1 = I_2 = \ldots I_n$

Calculating load resistance is now a simple matter of Ohm's Law (R = E/I), giving us 38.889 kΩ:

	$R_{dropping}$	R_{load}	Total	
E	32.4	12.6	45	Volts
I	324 µ	324 µ	324 µ	Amps
R	100 k	**38.889 k**		Ohms

Ohm's Law

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

Thus, if the load resistance is exactly 38.889 kΩ, there will be 12.6 volts across it, diode or no diode. Any load resistance smaller than 38.889 kΩ will result in a load voltage less than 12.6 volts, diode or no diode. With the diode in place, the load voltage will be regulated to a maximum of 12.6 volts for any load resistance *greater* than 38.889 kΩ.

With the original value of 1 kΩ for the dropping resistor, our regulator circuit was able to adequately regulate voltage even for a load resistance as low as 500 Ω. What we see is a tradeoff between power dissipation and acceptable load resistance. The higher-value dropping resistor gave us less power dissipation, at the expense of raising the acceptable minimum load resistance value. If we wish to regulate voltage for low-value load resistances, the circuit must be prepared to handle higher power dissipation.

Zener diodes regulate voltage by acting as complementary loads, drawing more or less current as necessary to ensure a constant voltage drop across the load. This is analogous to regulating the speed of an automobile by braking rather than by varying the throttle position: not only is it wasteful, but the brakes must be built to handle all the engine's power when the driving conditions don't demand it. Despite this fundamental inefficiency of design, zener diode regulator circuits are widely employed due to their sheer simplicity. In high-power applications where the inefficiencies would be unacceptable, other voltage-regulating techniques are applied. But even then, small zener-based circuits are often used to provide a "reference" voltage to drive a more efficient amplifier-type of circuit controlling the main power.

- **REVIEW:**

- Zener diodes are designed to be operated in reverse-bias mode, providing a relatively low, stable breakdown, or *zener* voltage at which they being to conduct substantial reverse current.

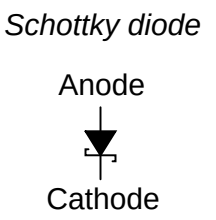
- A zener diode may function as a voltage regulator by acting as an accessory load, drawing more current from the source if the voltage is too high, and less if it is too low.

3.10 Special-purpose diodes

3.10.1 Schottky diodes

Schottky diodes are constructed of a *metal*-to-N junction rather than a P-N semiconductor junction. Also known as *hot-carrier* diodes, Schottky diodes are characterized by fast switching times (low reverse-recovery time), low forward voltage drop (typically 0.25 to 0.4 volts for a metal-silicon junction), and low junction capacitance.

The schematic symbol for a Schottky diode is shown here:

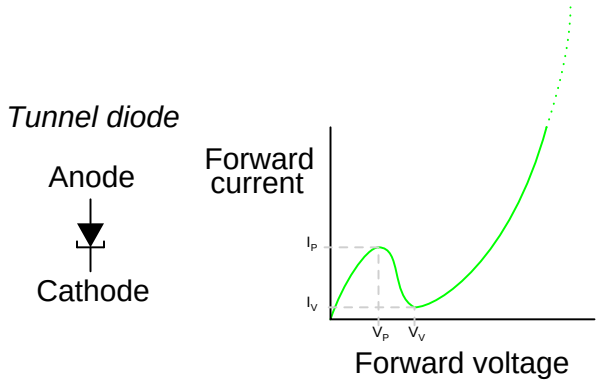


In terms of forward voltage drop (V_F), reverse-recovery time (t_{rr}), and junction capacitance (C_J), Schottky diodes are closer to ideal than the average "rectifying" diode. This makes them well suited for high-frequency applications. Unfortunately, though, Schottky diodes typically have lower forward current (I_F) and reverse voltage (V_{RRM} and V_{DC}) ratings than rectifying diodes and are thus unsuitable for applications involving substantial amounts of power.

Schottky diode technology finds broad application in high-speed computer circuits, where the fast switching time equates to high speed capability, and the low forward voltage drop equates to less power dissipation when conducting.

3.10.2 Tunnel diodes

Tunnel diodes exploit a strange quantum phenomenon called *resonant tunneling* to provide interesting forward-bias characteristics. When a small forward-bias voltage is applied across a tunnel diode, it begins to conduct current. As the voltage is increased, the current increases and reaches a peak value called the *peak current* (I_P). If the voltage is increased a little more, the current actually begins to *decrease* until it reaches a low point called the *valley current* (I_V). If the voltage is increased further yet, the current begins to increase again, this time without decreasing into another "valley." Both the schematic symbol and a current/voltage plot for the tunnel diode are shown in the following illustration:



The forward voltages necessary to drive a tunnel diode to its peak and valley currents are known as peak voltage (V_P) and valley voltage (V_V), respectively. The region on the graph where current is decreasing while applied voltage is increasing (between V_P and V_V on the horizontal scale) is known as the region of *negative resistance*.

Tunnel diodes, also known as *Esaki diodes* in honor of their Japanese inventor Leo Esaki, are able to transition between peak and valley current levels very quickly, "switching" between high and low states of conduction much faster than even Schottky diodes. Tunnel diode characteristics are also relatively unaffected by changes in temperature.

Unfortunately, tunnel diodes are not good rectifiers, as they have relatively high "leakage" current when reverse-biased. Consequently, they find application only in special circuits where their unique tunnel effect has value. In order to exploit the tunnel effect, these diodes are maintained at a bias voltage somewhere between the peak and valley voltage levels, always in a forward-biased polarity (anode positive, and cathode negative).

Perhaps the most common application of a tunnel diode is in simple high-frequency oscillator circuits, where they allow a DC voltage source to contribute power to an LC "tank" circuit, the diode conducting when the voltage across it reaches the peak (tunnel) level and effectively insulating at all other voltages.

3.10.3 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.

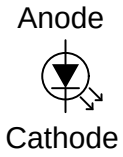
Electrons flowing through a PN junction experience similar transitions in energy level, and emit radiant energy as they do so. The frequency of this radiant energy is determined by the crystal

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.

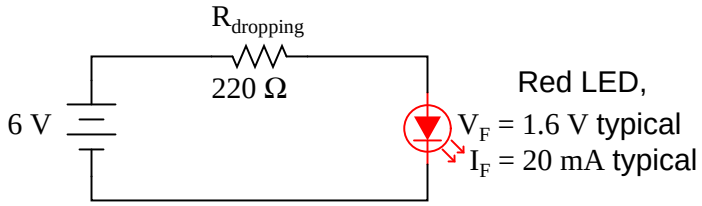
The schematic symbol for an LED is a regular diode shape inside of a circle, with two small arrows pointing away (indicating emitted light):

Light-emitting diode (LED)



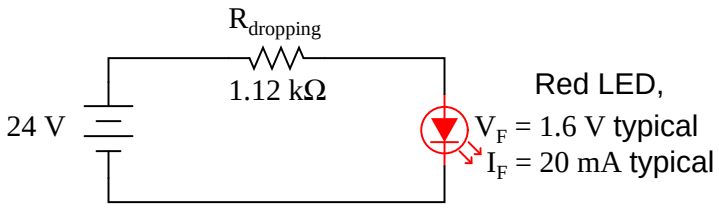
This notation of having two small arrows pointing away from the device is common to the schematic symbols of all light-emitting semiconductor devices. Conversely, if a device is light-*activated* (meaning that incoming light stimulates it), then the symbol will have two small arrows pointing *toward* 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

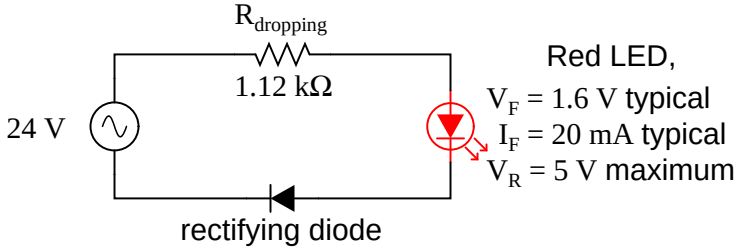
voltages will require larger-value dropping resistors, and possibly higher-power rating resistors as well. Consider this example for a supply voltage of 24 volts:



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:



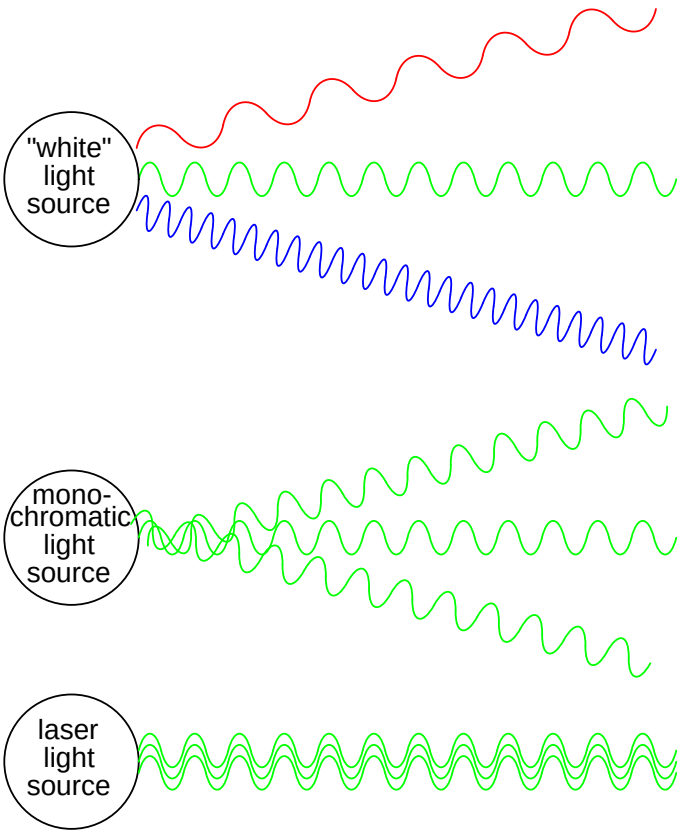
As lamps, LEDs are superior to incandescent bulbs in many ways. First and foremost is efficiency: LEDs output far more light power per watt than an incandescent lamp. This is a significant advantage if the circuit in question is battery-powered, efficiency translating to longer battery life. Second is 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.

3.10.4 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.

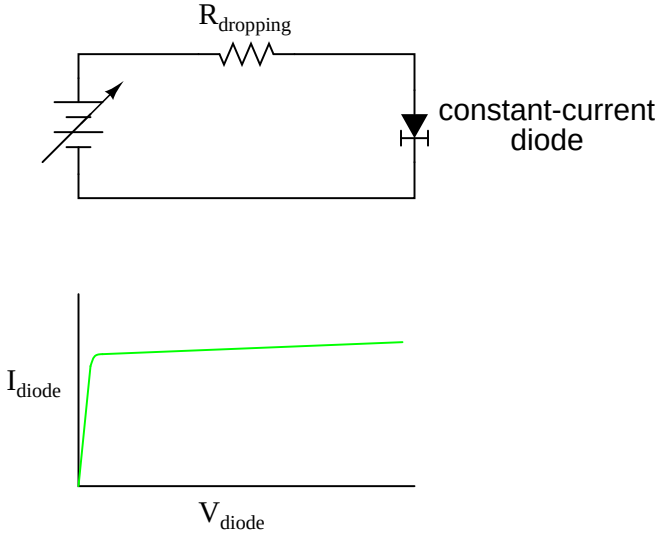
Some laser diodes require special high-power "pulsing" circuits to deliver large quantities of voltage and current in short bursts. Other laser diodes may be operated continuously at lower power. In the latter case, laser action occurs only within a certain range of diode current, necessitating some 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.

3.10.5 Photodiodes

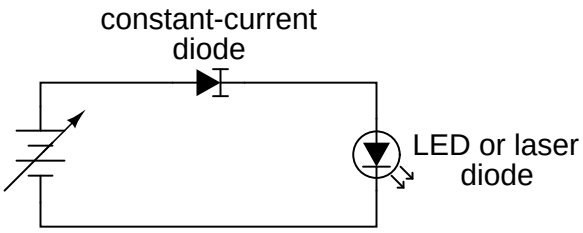
3.10.6 Varactor diodes

3.10.7 Constant-current diodes

A *constant-current diode*, also known as a *current-limiting diode*, or *current-regulating diode*, does exactly what its name implies: it regulates current through it to some maximum level. If you try to force more current through a constant-current diode than its current-regulation point, it simply "fights back" by dropping more voltage. If we were to build the following circuit and plot diode current over diode current, we'd get a graph that rises normally at first and then levels off at the current regulation point:



One interesting application for a constant-current diode is to automatically limit current through an LED or laser diode over a wide range of power supply voltages, like this:



Of course, the constant-current diode's regulation point should be chosen to match the LED or laser diode's optimum forward current. This is especially important for the laser diode, not so much for the LED, as regular LEDs tend to be more tolerant of forward current variations.

Another application is in the charging of small secondary-cell batteries, where a constant charging current leads to very predictable charging times. Of course, large secondary-cell battery banks might also benefit from constant-current charging, but constant-current diodes tend to be very small devices, limited to regulating currents in the milliamp range.

3.11 Other diode technologies

3.12 Contributors

Contributors to this chapter are listed in chronological order of their contributions, from most recent to first. See Appendix 2 (Contributor List) for dates and contact information.

Jered Wierzbicki (December 2002): Pointed out error in diode equation – Boltzmann's constant shown incorrectly.

Chapter 4

BIPOLAR JUNCTION TRANSISTORS

Contents

*** INCOMPLETE ***

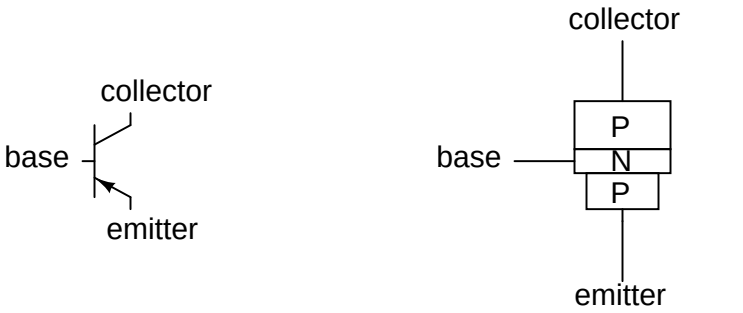
4.1 Introduction

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.

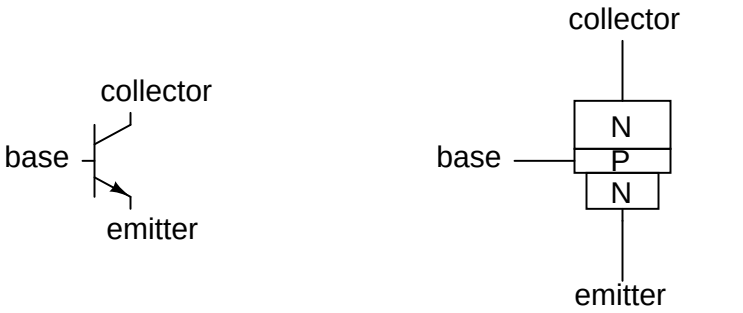
A bipolar transistor consists of a three-layer "sandwich" of doped (extrinsic) semiconductor materials, either P-N-P or N-P-N. Each layer forming the transistor has a specific name, and each layer is provided with a wire contact for connection to a circuit. Shown here are schematic symbols and physical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

PNP transistor



schematic symbol *physical diagram*

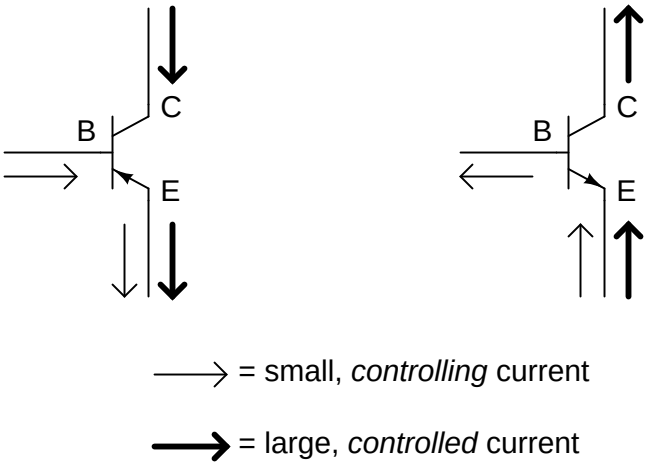
NPN transistor



schematic symbol *physical diagram*

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

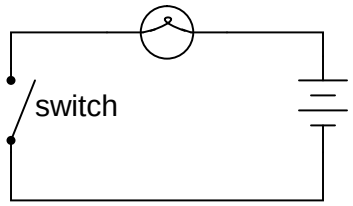
of semiconductor material: P and N. The current consists of both electron and hole flow, in different parts of the transistor.

- 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.

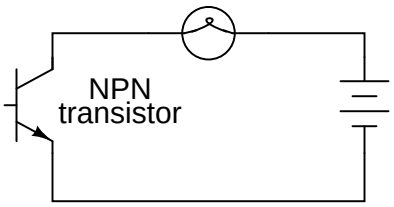
4.2 The transistor as a switch

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.

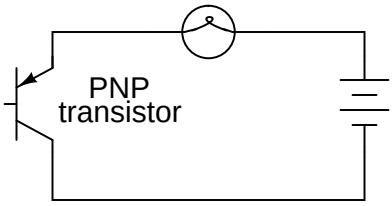
Suppose we had a lamp that we wanted to turn on and off by means of a switch. Such a circuit would be extremely simple:



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:

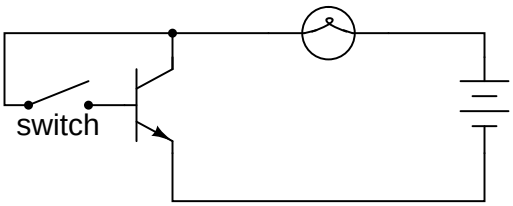


In this example I happened to choose an NPN transistor. A PNP transistor could also have been chosen for the job, and its application would look like this:

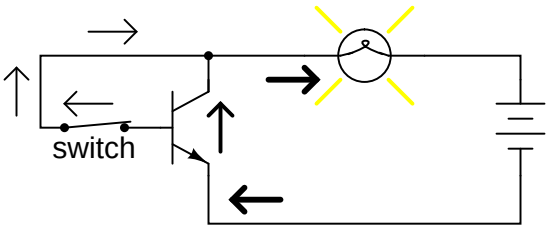


The choice between NPN and PNP is really arbitrary. All that matters is that the proper current directions are maintained for the sake of correct junction biasing (electron flow going *against* the transistor symbol's arrow).

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:



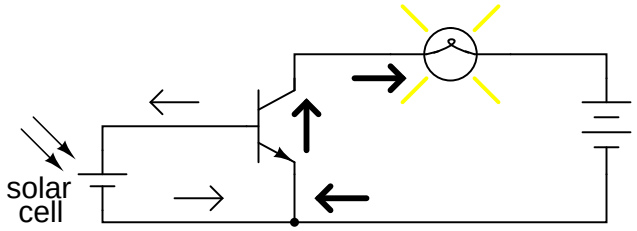
If the switch is open, the base wire of the transistor will be left "floating" (not connected to anything) and there will be no current through it. In this state, the transistor is said to be *cutoff*. If the switch is closed, however, electrons will be able to flow from the emitter through to the base 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*.



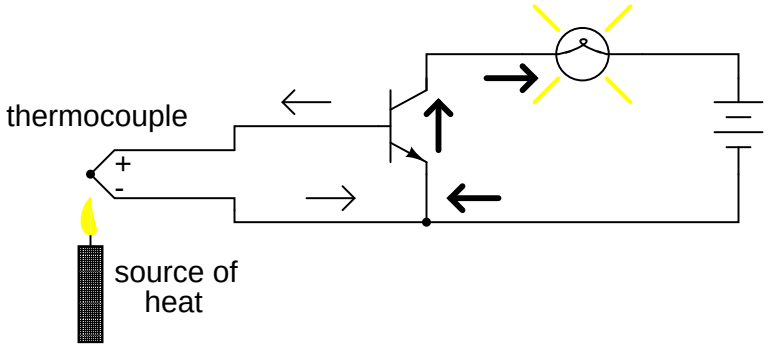
Of course, it may seem pointless to use a transistor in this capacity to control the lamp. After all, we're still using a switch in the circuit, aren't we? If we're still using a switch to control the lamp – if only indirectly – then what's the point of having a transistor to control the current? Why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

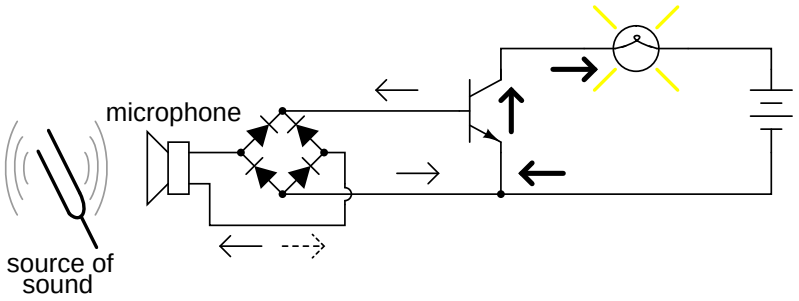
be an important advantage if the switch has a low current rating: a small switch may be used to control a relatively high-current load. Perhaps more importantly, though, is the fact that the current-controlling behavior of the transistor enables us to use something completely different to turn the lamp 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:



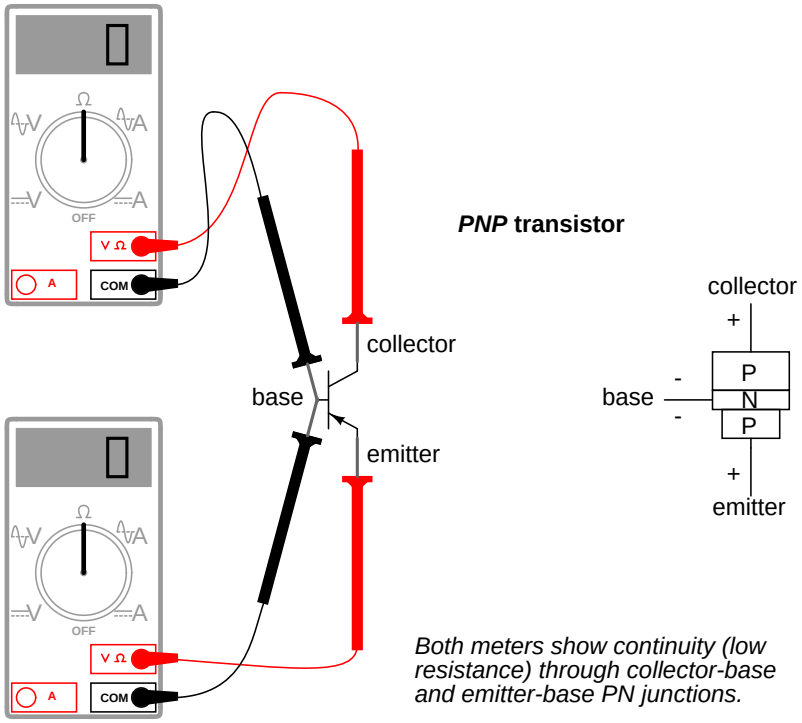
The point should be quite apparent by now: *any* sufficient source of DC current may be used to turn the transistor on, and that source of current need only be a fraction of the amount of current needed to energize the lamp. Here we see the transistor functioning not only as a switch, but as a true *amplifier*: 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).

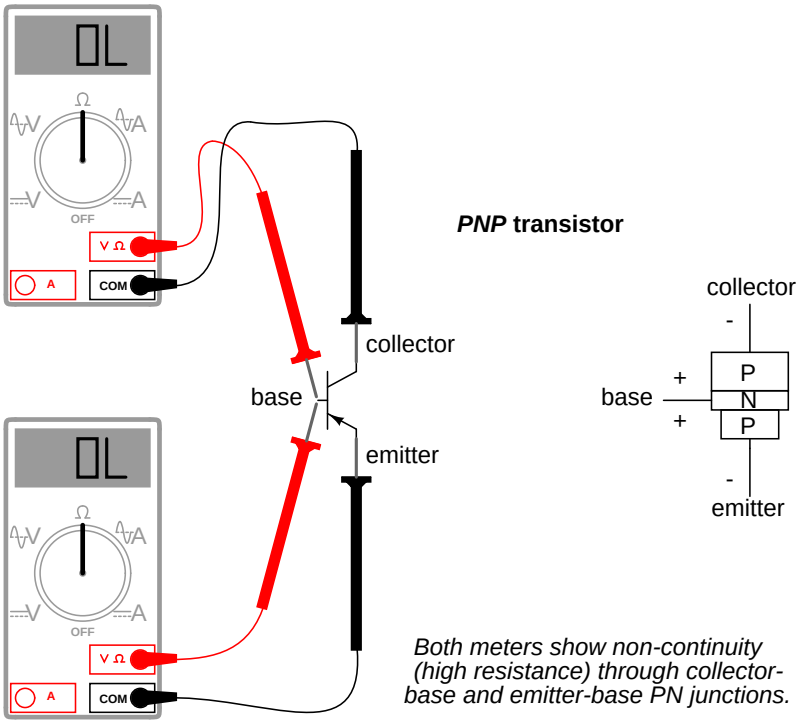
4.3 Meter check of a transistor

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

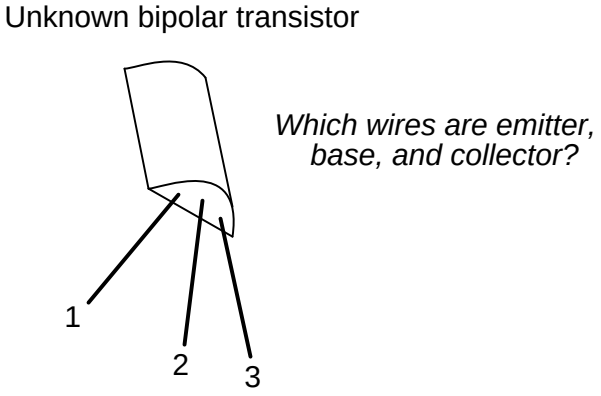
designated "diode check" function, use that rather than the "resistance" range, and the meter will display the actual forward voltage of the PN junction and not just whether or not it conducts current.



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.

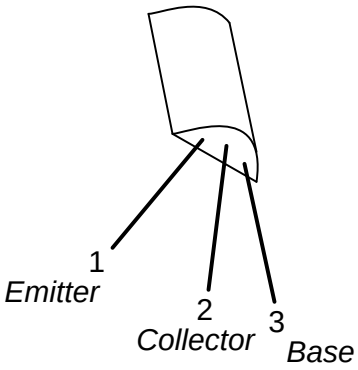
Suppose a technician finds a bipolar transistor and proceeds to measure continuity with a multimeter set in the "diode check" mode. Measuring between pairs of wires and recording the values displayed by the meter, the technician obtains the following data:



- 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).

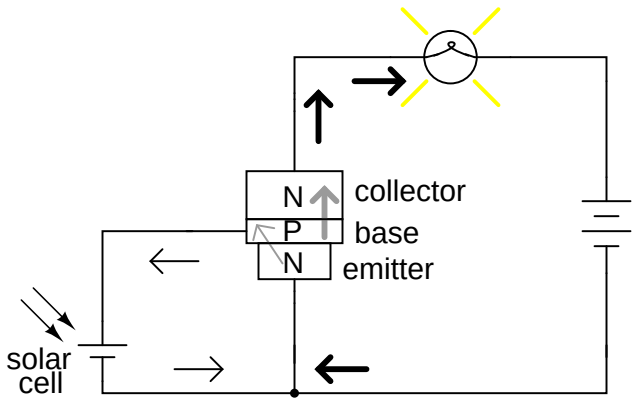
Now we look for the one wire common to both sets of conductive readings. It must be the base connection of the transistor, because the base is the only layer of the three-layer device common to both sets of PN junctions (emitter-base and collector-base). In this example, that wire is number 3, being 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.

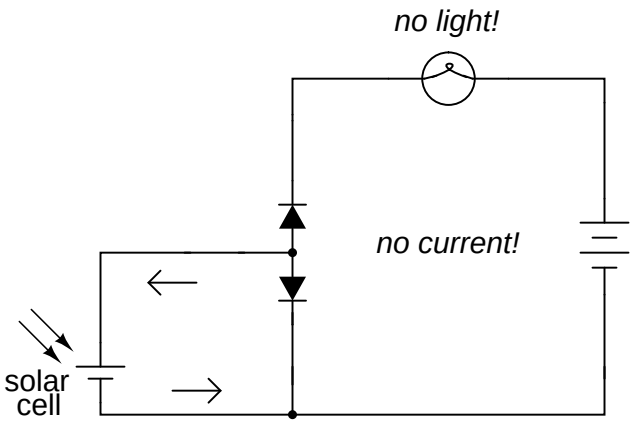
Knowing that a bipolar transistor behaves as two back-to-back diodes when tested with a conductivity meter is helpful for identifying an unknown transistor purely by meter readings. It is also helpful for a quick functional check of the transistor. If the technician were to measure continuity in any 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.

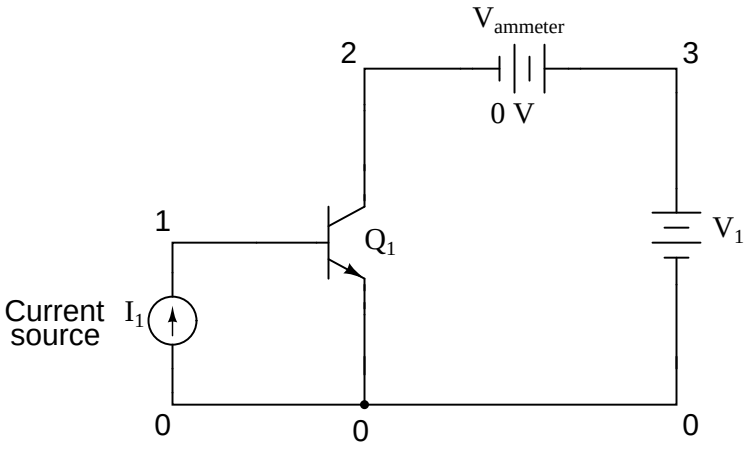
4.4 Active mode operation

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:

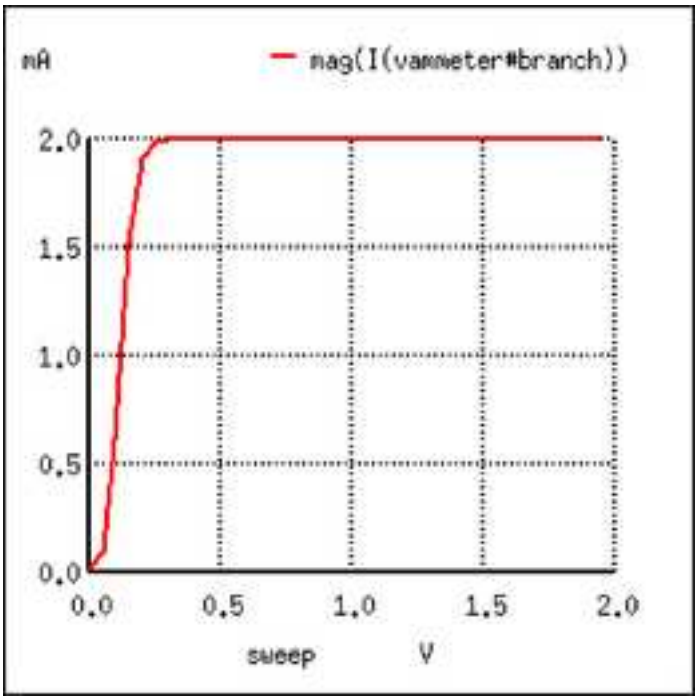


"Q" is the standard letter designation for a transistor in a schematic diagram, just as "R" is for resistor and "C" is for capacitor. In this circuit, we have an NPN transistor powered by a battery (V_1) and controlled by current through a *current source* (I_1). A current source is a device that outputs 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.

In the SPICE simulation, I'll set the current source at a constant value of 20 μA, then vary the voltage source (V_1) over a range of 0 to 2 volts and monitor how much current goes through it. The "dummy" battery ($V_{ammeter}$) with its output of 0 volts serves merely to provide SPICE with a circuit element for current measurement.

```
bipolar transistor simulation
i1 0 1 dc 20u
q1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 npn
.dc v1 0 2 0.05
.plot dc i(vammeter)
.end
```

```
type        npn
is        1.00E-16
bf         100.000
nf           1.000
br           1.000
nr           1.000
```

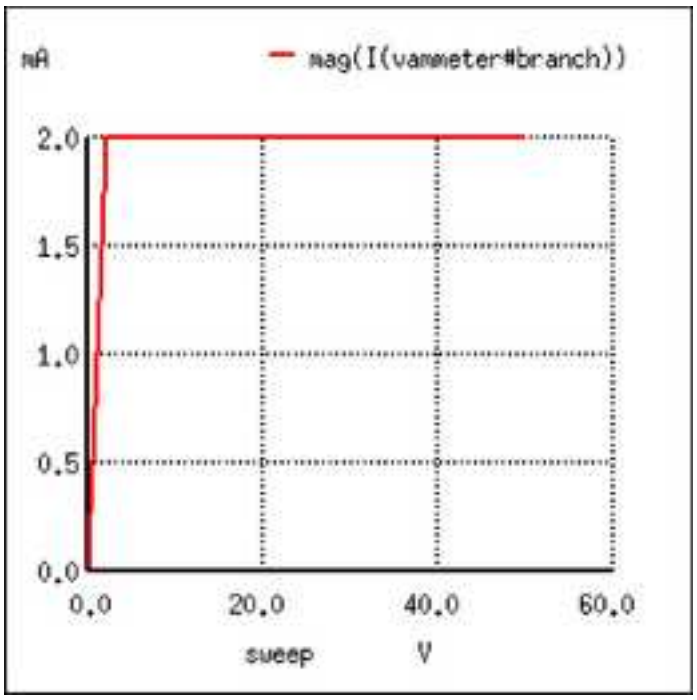


The constant base current of 20 μA sets a collector current limit of 2 mA, exactly 100 times as much. Notice how flat the curve is for collector current over the range of battery voltage from 0 to 2 volts. The only exception to this featureless plot is at the very beginning, where the battery increases from 0 volts to 0.25 volts. There, the collector current increases rapidly from 0 amps to its limit of 2 mA.

Let's see what happens if we vary the battery voltage over a wider range, this time from 0 to 50 volts. We'll keep the base current steady at 20 μA:

```
bipolar transistor simulation
i1 0 1 dc 20u
q1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 npn
.dc v1 0 50 2
.plot dc i(vammeter)
.end

type        npn
is        1.00E-16
bf         100.000
nf           1.000
br           1.000
```

```
nr              1.000
```

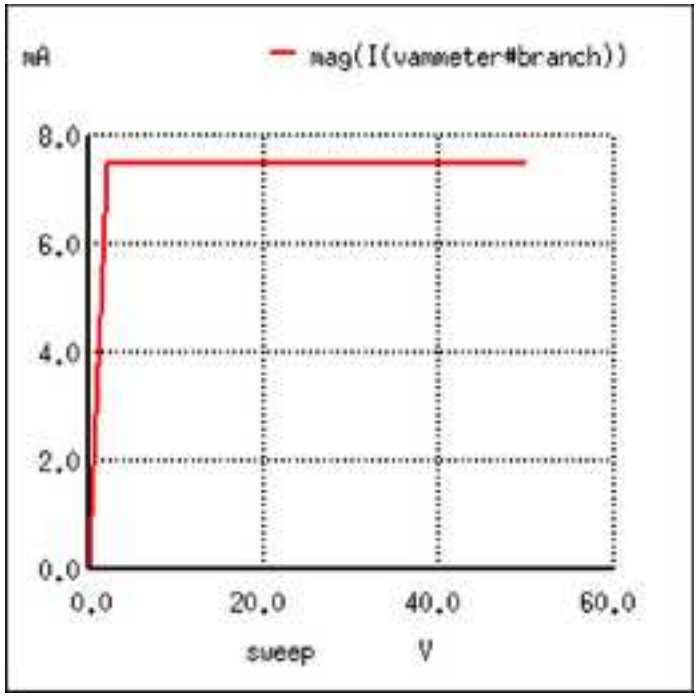


Same result! The collector current holds absolutely steady at 2 mA despite the fact that the battery (v1) voltage varies all the way from 0 to 50 volts. It would appear from our simulation that collector-to-emitter voltage has little effect over collector current, except at very low levels (just above 0 volts). The transistor is acting as a current regulator, allowing exactly 2 mA through the collector and no more.

Now let's see what happens if we increase the controlling (I_1) current from 20 μA to 75 μA, once again sweeping the battery (V_1) voltage from 0 to 50 volts and graphing the collector current:

```
bipolar transistor simulation
i1 0 1 dc 75u
q1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 npn
.dc v1 0 50 2
.plot dc i(vammeter)
.end

type        npn
is        1.00E-16
```

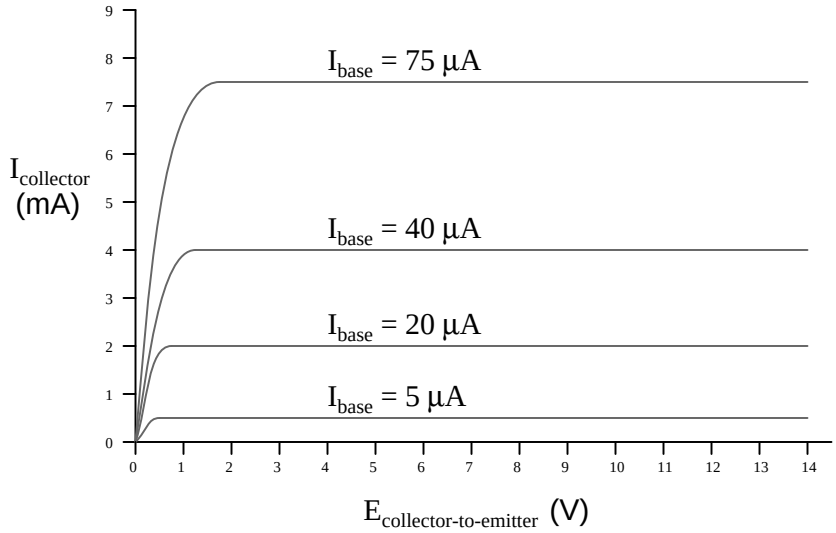
```
bf          100.000
nf            1.000
br            1.000
nr            1.000
```



Not surprisingly, SPICE gives us a similar plot: a flat line, holding steady this time at 7.5 mA – exactly 100 times the base current – over the range of battery voltages from just above 0 volts to 50 volts. It appears that the base current is the deciding factor for collector current, the V_1 battery voltage being irrelevant so long as it's above a certain minimum level.

This voltage/current relationship is entirely different from what we're used to seeing across a resistor. With a resistor, current increases linearly as the voltage across it increases. Here, with a transistor, current from emitter to collector stays limited at a fixed, maximum value no matter how high the voltage across emitter and collector increases.

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}

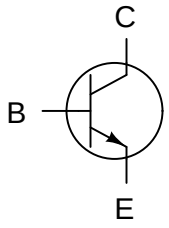
Sometimes the β ratio is designated as "h_{fe}," a label used in a branch of mathematical semiconductor analysis known as "hybrid parameters" which strives to achieve very precise predictions of transistor performance with detailed equations. Hybrid parameter variables are many, but they are all labeled 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.

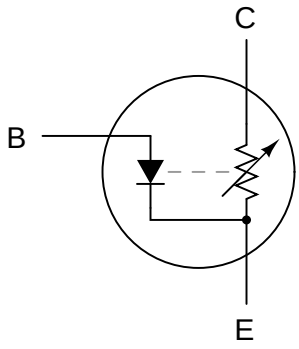
It would be nice if the β of a transistor remained stable for all operating conditions, but this is not true in real life. For an actual transistor, the β ratio may vary by a factor of over 3 within its operating current limits. For example, a transistor with advertised β of 50 may actually test with 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:

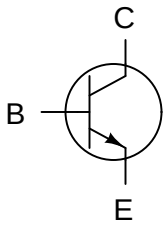


NPN diode-rheostat model

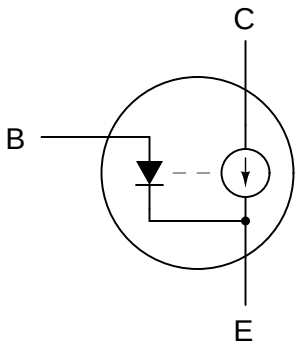


This model casts the transistor as a combination of diode and rheostat (variable resistor). Current through the base-emitter diode controls the resistance of the collector-emitter rheostat (as implied by the dashed line connecting the two components), thus controlling collector current. An NPN transistor 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:

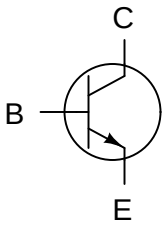


NPN diode-current source model

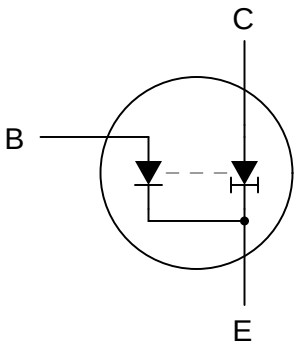


It casts the transistor as a combination of diode and current source, the output of the current source being set at a multiple (β ratio) of the base current. This model is far more accurate in depicting the true input/output characteristics of a transistor: base current establishes a certain amount of 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.

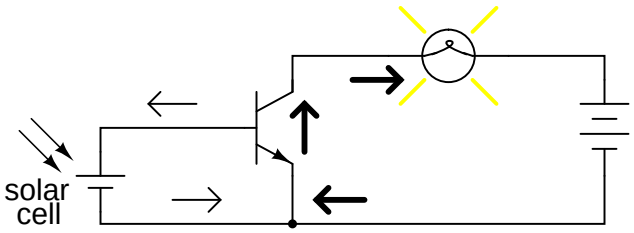
- **REVIEW:**
- A transistor is said to be in its *active* mode if it is operating somewhere between fully on (saturated) and fully off (cutoff).
- Base current tends to regulate collector current. By *regulate*, we mean that no more collector current may exist than what is allowed by the base current.
- 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.

4.5 The common-emitter amplifier

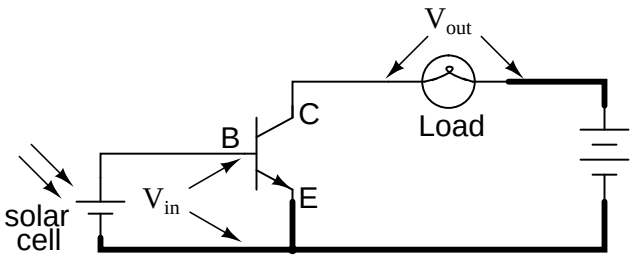
At the beginning of this chapter we saw how transistors could be used as switches, operating in either their "saturation" or "cutoff" modes. In the last section we saw how transistors behave

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:

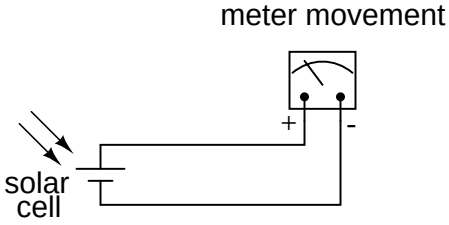


It is called the *common-emitter* configuration because (ignoring the power supply battery) both the signal source and the load share the emitter lead as a common connection point. This is not the only way in which a transistor may be used as an amplifier, as we will see in later sections of this chapter:



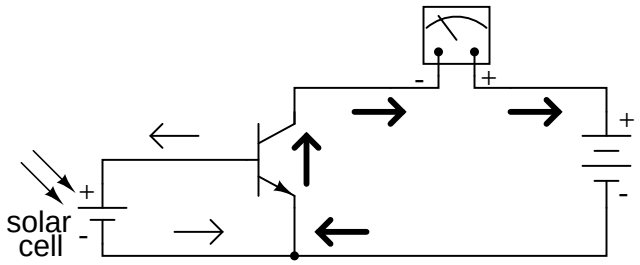
Before, this circuit was shown to illustrate how a relatively small current from a solar cell could be used to saturate a transistor, resulting in the illumination of a lamp. Knowing now that transistors are able to "throttle" their collector currents according to the amount of base current supplied 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.

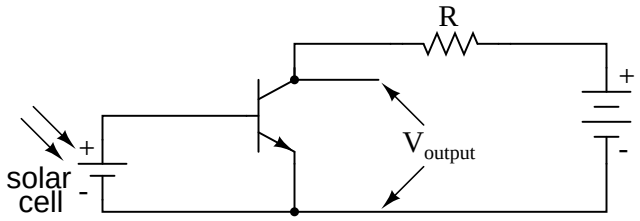
Perhaps the most direct solution to this measurement problem is to use a transistor to *amplify* the solar cell's current so that more meter movement needle deflection may be obtained for less incident light. Consider this approach:



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:



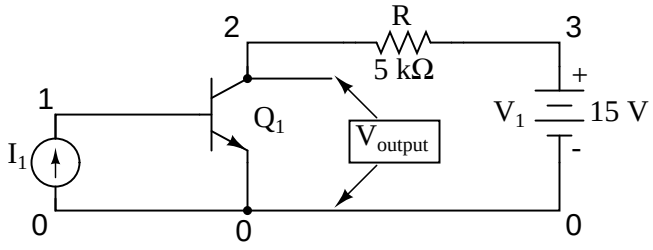
With the solar cell darkened (no current), the transistor will be in cutoff mode and behave as an open switch between collector and emitter. This will produce maximum voltage drop between collector and emitter for maximum V_{output}, equal to the full voltage of the battery.

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.

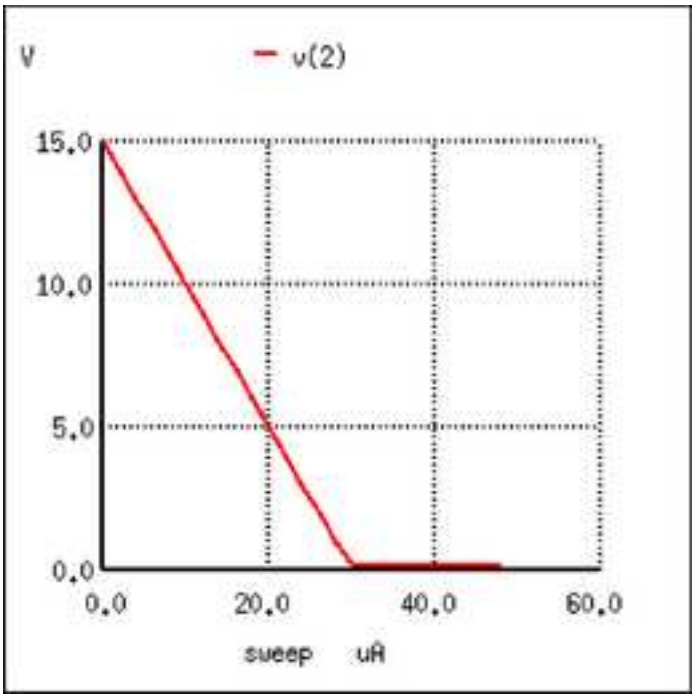
For light exposure levels somewhere between zero and maximum solar cell output, the transistor will be in its active mode, and the output voltage will be somewhere between zero and full battery voltage. An important quality to note here about the common-emitter configuration is that the output voltage 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.

A quick SPICE simulation will verify our qualitative conclusions about this amplifier circuit:



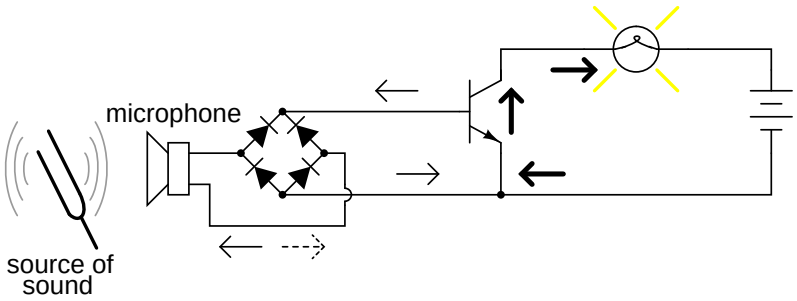

```
common-emitter amplifier
i1 0 1 dc
q1 2 1 0 mod1
r 3 2 5000
v1 3 0 dc 15
.model mod1 npn
.dc i1 0 50u 2u
.plot dc v(2,0)
.end
```

```
type        npn
is        1.00E-16
bf         100.000
nf           1.000
br           1.000
nr           1.000
```

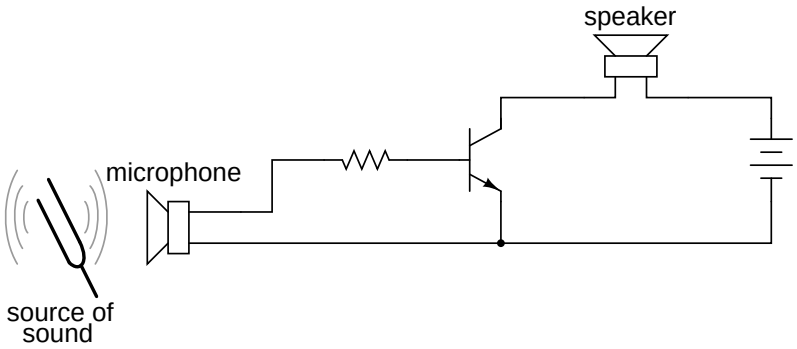


At the beginning of the simulation where the current source (solar cell) is outputting zero current, the transistor is in cutoff mode and the full 15 volts from the battery is shown at the amplifier output (between nodes 2 and 0). As the solar cell's current begins to increase, the output voltage proportionally decreases, until the transistor reaches saturation at 30 μA of base current (3 mA of collector current). Notice how the output voltage trace on the graph is perfectly linear (1 volt steps from 15 volts to 1 volt) until the point of saturation, where it never quite reaches zero. This is the effect mentioned earlier, where a saturated transistor can never achieve exactly zero voltage drop between collector and emitter due to internal junction effects. What we do see is a sharp output voltage decrease from 1 volt to 0.2261 volts as the input current increases from 28 μA to 30 μA, and then a continuing decrease in output voltage from then on (albeit in progressively smaller steps). The lowest the output voltage ever gets in this simulation is 0.1299 volts, asymptotically approaching zero.

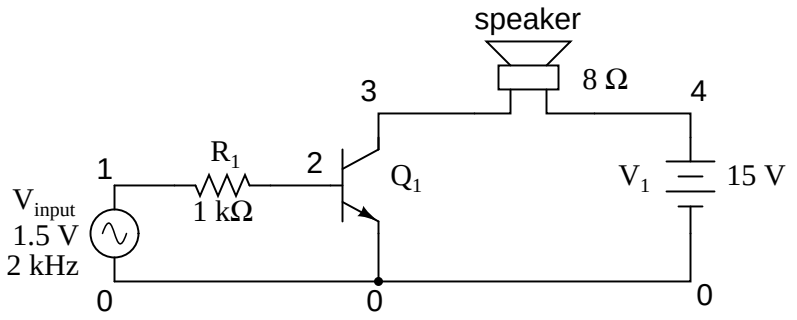
So far, we've seen the transistor used as an amplifier for DC signals. In the solar cell light meter example, we were interested in amplifying the DC output of the solar cell to drive a DC meter movement, or to produce a DC output voltage. However, this is not the only way in which a transistor may be employed as an amplifier. In many cases, what is desired is an AC amplifier for amplifying *alternating* current and voltage signals. One common application of this is in audio electronics (radios, televisions, and public-address systems). Earlier, we saw an example where the audio output of a tuning fork could be used to activate a transistor as a switch. Let's see if we can modify that circuit to send power to a speaker rather than to a lamp:



In the original circuit, a full-wave bridge rectifier was used to convert the microphone's AC output signal into a DC voltage to drive the input of the transistor. All we cared about here was turning the lamp on with a sound signal from the microphone, and this arrangement sufficed for that purpose. But now we want to actually reproduce the AC signal and drive a speaker. This means we cannot rectify the microphone's output anymore, because we need undistorted AC signal to drive the transistor! Let's remove the bridge rectifier and replace the lamp with a speaker:

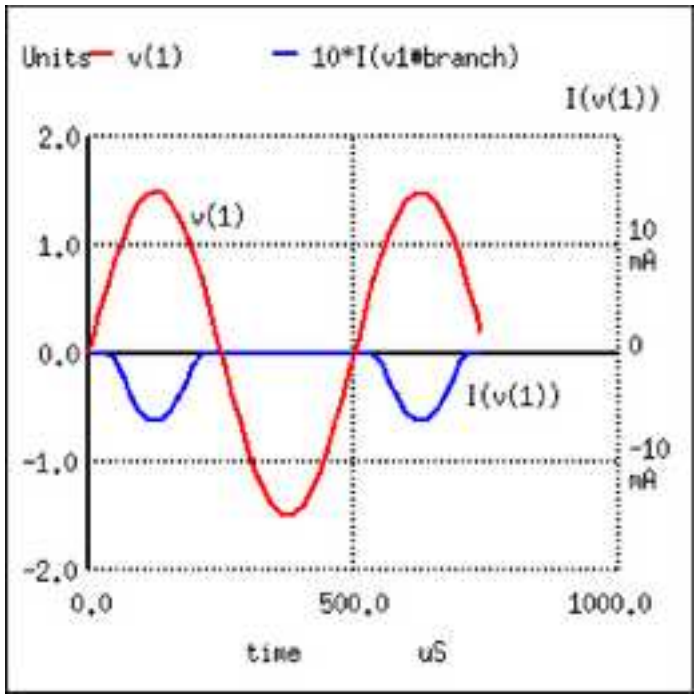


Since the microphone may produce voltages exceeding the forward voltage drop of the base-emitter PN (diode) junction, I've placed a resistor in series with the microphone. Let's simulate this circuit now in SPICE and see what happens:



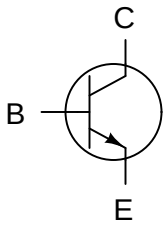

```
common-emitter amplifier
vinput 1 0 sin (0 1.5 2000 0 0)
r1 1 2 1k
```

```
q1 3 2 0 mod1
rspkr 3 4 8
v1 4 0 dc 15
.model mod1 npn
.tran 0.02m 0.74m
.plot tran v(1,0) i(v1)
.end
```

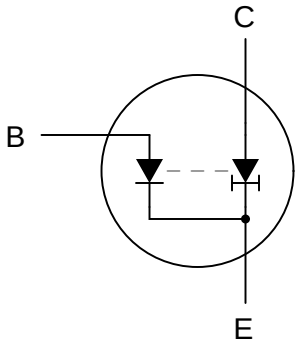


The simulation plots both the input voltage (an AC signal of 1.5 volt peak amplitude and 2000 Hz frequency) and the current through the 15 volt battery, which is the same as the current through the speaker. What we see here is a full AC sine wave alternating in both positive and negative directions, and a half-wave output current waveform that only pulses in one direction. If we were actually driving a speaker with this waveform, the sound produced would be horribly distorted.

What's wrong with the circuit? Why won't it faithfully reproduce the entire AC waveform from the microphone? The answer to this question is found by close inspection of the transistor diode-regulating diode model:

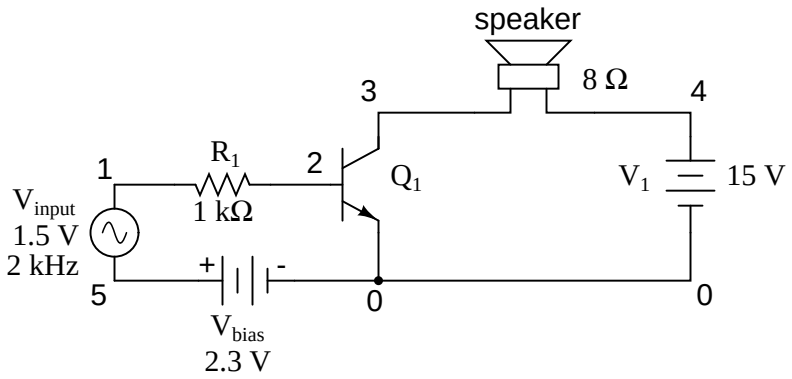


NPN diode-regulating diode model

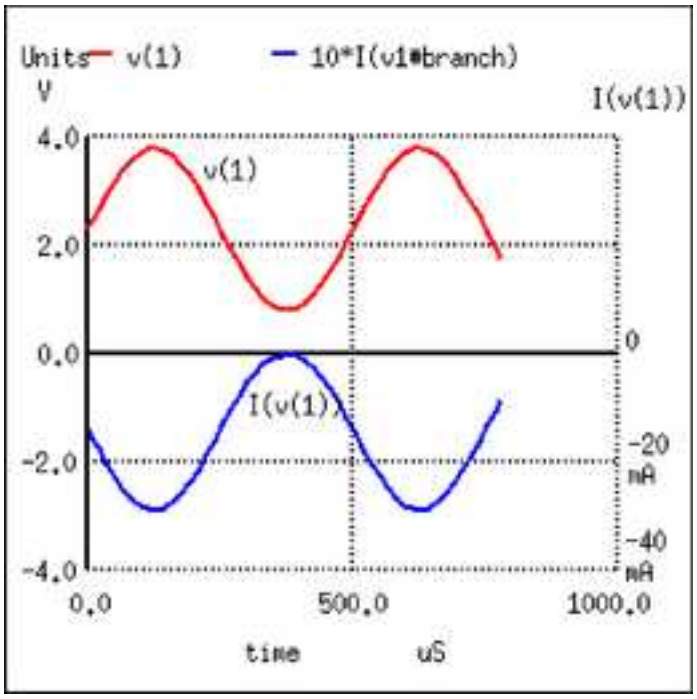


Collector current is controlled, or regulated, through the constant-current mechanism according to the pace set by the current through the base-emitter diode. Note that both current paths through the transistor are monodirectional: *one way only!* Despite our intent to use the transistor to amplify an AC signal, it is essentially a DC device, capable of handling currents in a single direction only. We may apply an AC voltage input signal between the base and emitter, but electrons cannot flow in that circuit during the part of the cycle that reverse-biases the base-emitter diode junction. Therefore, the transistor will remain in cutoff mode throughout that portion of the cycle. It will "turn on" in its active mode only when the input voltage is of the correct polarity to forward-bias the base-emitter diode, and only when that voltage is sufficiently high to overcome the diode's forward voltage drop. Remember that bipolar transistors are *current-controlled devices*: they regulate collector current based on the existence of base-to-emitter *current*, not base-to-emitter *voltage*.

The only way we can get the transistor to reproduce the entire waveform as current through the speaker is to keep the transistor in its active mode the entire time. This means we must maintain current through the base during the entire input waveform cycle. Consequently, the base-emitter diode junction must be kept forward-biased at all times. Fortunately, this can be accomplished with the aid of a DC *bias voltage* added to the input signal. By connecting a sufficient DC voltage in series with the AC signal source, forward-bias can be maintained at all points throughout the wave cycle:

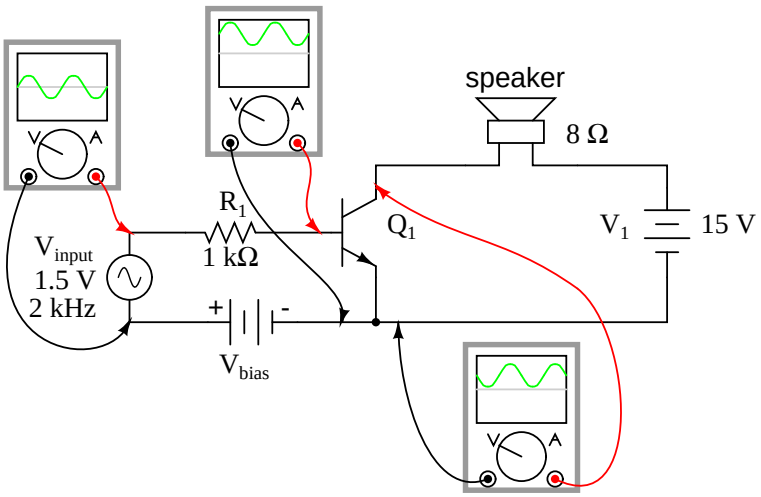



```
common-emitter amplifier
vinput 1 5 sin (0 1.5 2000 0 0)
vbias 5 0 dc 2.3
r1 1 2 1k
q1 3 2 0 mod1
rspkr 3 4 8
v1 4 0 dc 15
.model mod1 npn
.tran 0.02m 0.78m
.plot tran v(1,0) i(v1)
.end
```



With the bias voltage source of 2.3 volts in place, the transistor remains in its active mode throughout the entire cycle of the wave, faithfully reproducing the waveform at the speaker. Notice that the input voltage (measured between nodes 1 and 0) fluctuates between about 0.8 volts and 3.8 volts, a peak-to-peak voltage of 3 volts just as expected (source voltage = 1.5 volts peak). The output (speaker) current varies between zero and almost 300 mA, 180^o out of phase with the input (microphone) signal.

The following illustration is another view of the same circuit, this time with a few oscilloscopes ("scopemeters") connected at crucial points to display all the pertinent signals:



The need for biasing a transistor amplifier circuit to obtain full waveform reproduction is an important consideration. A separate section of this chapter will be devoted entirely to the subject biasing and biasing techniques. For now, it is enough to understand that biasing may be necessary for proper voltage and current output from the amplifier.

Now that we have a functioning amplifier circuit, we can investigate its voltage, current, and power gains. The generic transistor used in these SPICE analyses has a β of 100, as indicated by the short transistor statistics printout included in the text output (these statistics were cut from the last two analyses for brevity's sake):

```
type        npn
is        1.00E-16
bf         100.000
nf           1.000
br           1.000
nr           1.000
```

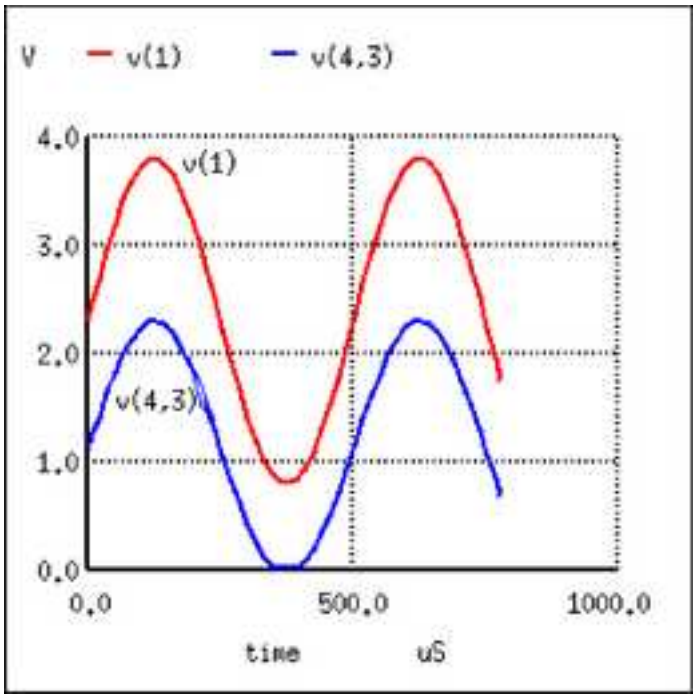
β is listed under the abbreviation "bf," which actually stands for "**b**eta, **f**orward". If we wanted to insert our own β ratio for an analysis, we could have done so on the `.model` line of the SPICE netlist.

Since β is the ratio of collector current to base current, and we have our load connected in series with the collector terminal of the transistor and our source connected in series with the base, the ratio of output current to input current is equal to beta. Thus, our current gain for this example amplifier is 100, or 40 dB.

Voltage gain is a little more complicated to figure than current gain for this circuit. As always, voltage gain is defined as the ratio of output voltage divided by input voltage. In order to experimentally determine this, we need to modify our last SPICE analysis to plot output voltage rather than output current so we have two voltage plots to compare:

```
common-emitter amplifier
vinput 1 5 sin (0 1.5 2000 0 0)
```

```
vbias 5 0 dc 2.3
r1 1 2 1k
q1 3 2 0 mod1
rspkr 3 4 8
v1 4 0 dc 15
.model mod1 npn
.tran 0.02m 0.78m
.plot tran v(1,0) v(4,3)
.end
```

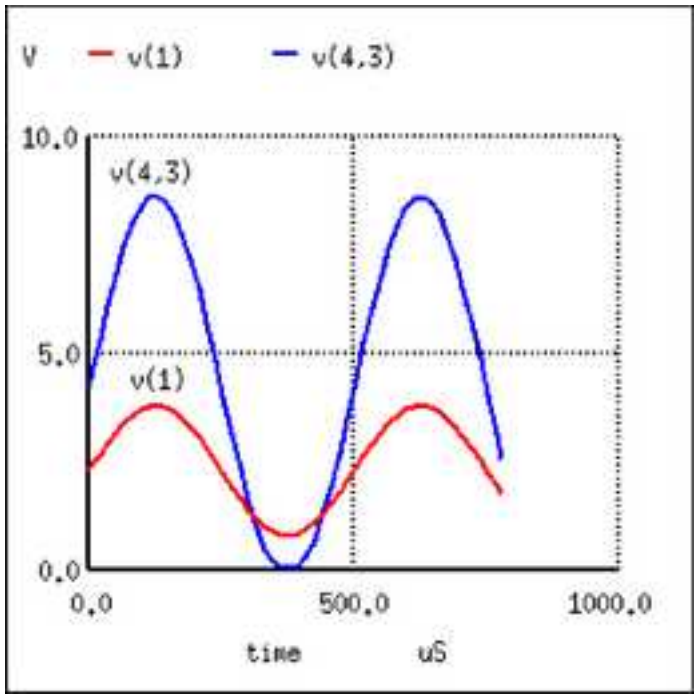


Plotted on the same scale (from 0 to 4 volts), we see that the output waveform ("+") has a smaller peak-to-peak amplitude than the input waveform ("*"), in addition to being at a lower bias voltage, not elevated up from 0 volts like the input. Since voltage gain for an AC amplifier is defined by the ratio of AC amplitudes, we can ignore any DC bias separating the two waveforms. Even so, the input waveform is still larger than the output, which tells us that the voltage gain is less than 1 (a negative dB figure).

To be honest, this low voltage gain is not characteristic to *all* common-emitter amplifiers. In this case it is a consequence of the great disparity between the input and load resistances. Our input resistance (R_1) here is 1000 Ω, while the load (speaker) is only 8 Ω. Because the current gain of this amplifier is determined solely by the β of the transistor, and because that β figure is fixed, the current gain for this amplifier won't change with variations in either of these resistances. However, voltage gain *is* dependent on these resistances. If we alter the load resistance, making it a larger value, it will drop a proportionately greater voltage for its range of load currents, resulting in a

larger output waveform. Let's try another simulation, only this time with a 30 Ω load instead of an 8 Ω load:

```
common-emitter amplifier
vinput 1 5 sin (0 1.5 2000 0 0)
vbias 5 0 dc 2.3
r1 1 2 1k
q1 3 2 0 mod1
rspkr 3 4 30
v1 4 0 dc 15
.model mod1 npn
.tran 0.02m 0.78m
.plot tran v(1,0) v(4,3)
.end
```



This time the output voltage waveform is significantly greater in amplitude than the input waveform. Looking closely, we can see that the output waveform ("+") crests between 0 and about 9 volts: approximately 3 times the amplitude of the input voltage.

We can perform another computer analysis of this circuit, only this time instructing SPICE to analyze it from an AC point of view, giving us peak voltage figures for input and output instead of a time-based plot of the waveforms:

```
common-emitter amplifier
vinput 1 5 ac 1.5
vbias 5 0 dc 2.3
r1 1 2 1k
q1 3 2 0 mod1
rspkr 3 4 30
v1 4 0 dc 15
.model mod1 npn
.ac lin 1 2000 2000
.print ac v(1,0) v(4,3)
.end
```

```
freq          v(1)        v(4,3)
2.000E+03     1.500E+00   4.418E+00
```

Peak voltage measurements of input and output show an input of 1.5 volts and an output of 4.418 volts. This gives us a voltage gain ratio of 2.9453 (4.418 V / 1.5 V), or 9.3827 dB.

$$A_V = \frac{V_{out}}{V_{in}}$$

$$A_V = \frac{4.418\ V}{1.5\ V}$$

$$A_V = 2.9453$$

$$A_{V(dB)} = 20 \log A_{V(ratio)}$$

$$A_{V(dB)} = 20 \log 2.9453$$

$$A_{V(dB)} = 9.3827\ dB$$

Because the current gain of the common-emitter amplifier is fixed by β, and since the input and output voltages will be equal to the input and output currents multiplied by their respective resistors, we can derive an equation for approximate voltage gain:

$$A_V = \beta \frac{R_{out}}{R_{in}}$$

$$A_V = (100) \frac{30\ \Omega}{1000\ \Omega}$$

$$A_V = 3$$

$$A_{V(dB)} = 20 \log A_{V(ratio)}$$

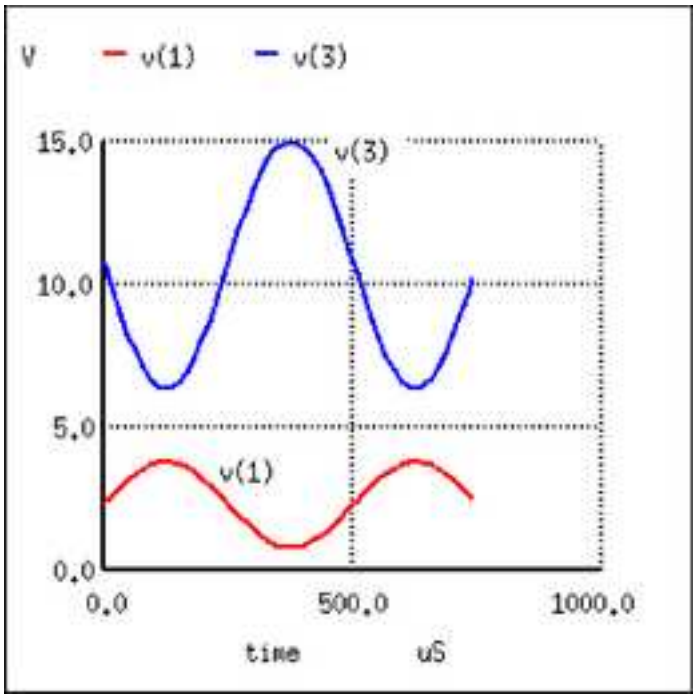
$$A_{V(dB)} = 20 \log 3$$

$$A_{V(dB)} = 9.5424\ dB$$

As you can see, the predicted results for voltage gain are quite close to the simulated results. With perfectly linear transistor behavior, the two sets of figures would exactly match. SPICE does a reasonable job of accounting for the many "quirks" of bipolar transistor function in its analysis, hence the slight mismatch in voltage gain based on SPICE's output.

These voltage gains remain the same regardless of where we measure output voltage in the circuit: across collector and emitter, or across the series load resistor as we did in the last analysis. The amount of output voltage *change* for any given amount of input voltage will remain the same. Consider the two following SPICE analyses as proof of this. The first simulation is time-based, to provide a plot of input and output voltages. You will notice that the two signals are 180^o out of phase with each other. The second simulation is an AC analysis, to provide simple, peak voltage readings for input and output:

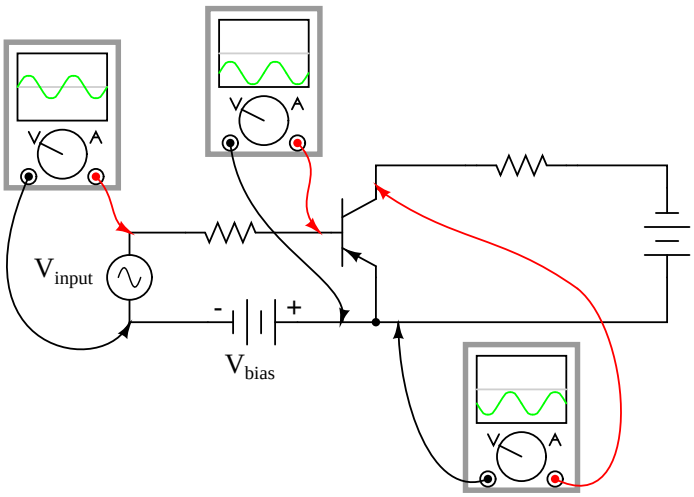
```
common-emitter amplifier
vinput 1 5 sin (0 1.5 2000 0 0)
vbias 5 0 dc 2.3
r1 1 2 1k
q1 3 2 0 mod1
rspkr 3 4 30
v1 4 0 dc 15
.model mod1 npn
.tran 0.02m 0.74m
.plot tran v(1,0) v(3,0)
.end
```




```
common-emitter amplifier
vinput 1 5 ac 1.5
vbias 5 0 dc 2.3
r1 1 2 1k
q1 3 2 0 mod1
rspkr 3 4 30
v1 4 0 dc 15
.model mod1 npn
.ac lin 1 2000 2000
.print ac v(1,0) v(3,0)
.end

freq          v(1)        v(3)
2.000E+03     1.500E+00   4.418E+00
```

We still have a peak output voltage of 4.418 volts with a peak input voltage of 1.5 volts. The only difference from the last set of simulations is the *phase* of the output voltage.

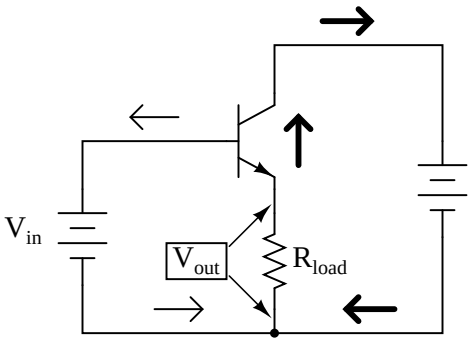
So far, the example circuits shown in this section have all used NPN transistors. PNP transistors are just as valid to use as NPN in *any* amplifier configuration, so long as the proper polarity and current directions are maintained, and the common-emitter amplifier is no exception. The inverting behavior and gain properties of a PNP transistor amplifier are the same as its NPN counterpart, just the polarities are different:



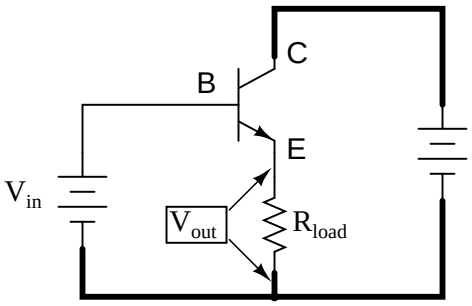
- **REVIEW:**
- *Common-emitter* transistor amplifiers are so-called because the input and output voltage points share the emitter lead of the transistor in common with each other, not considering any power supplies.
- Transistors are essentially DC devices: they cannot directly handle voltages or currents that reverse direction. In order to make them work for amplifying AC signals, the input signal must be offset with a DC voltage to keep the transistor in its active mode throughout the entire cycle of the wave. This is called *biasing*.
- If the output voltage is measured between emitter and collector on a common-emitter amplifier, it will be 180^o out of phase with the input voltage waveform. For this reason, the common-emitter amplifier is called an *inverting* amplifier circuit.
- The current gain of a common-emitter transistor amplifier with the load connected in series with the collector is equal to β. The voltage gain of a common-emitter transistor amplifier is approximately given here:
- $$A_V = \beta \frac{R_{out}}{R_{in}}$$
- Where "R_{out}" is the resistor connected in series with the collector and "R_{in}" is the resistor connected in series with the base.

4.6 The common-collector amplifier

Our next transistor configuration to study is a bit simpler in terms of gain calculations. Called the *common-collector* configuration, its schematic diagram looks like this:

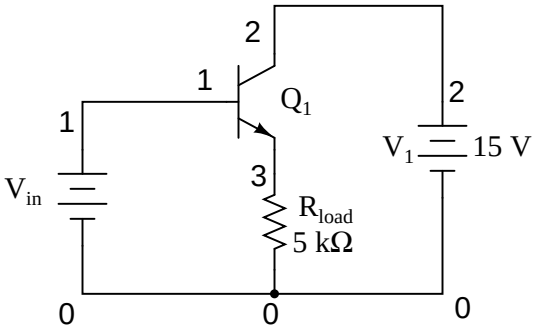


It is called the *common-collector* configuration because (ignoring the power supply battery) both the signal source and the load share the collector lead as a common connection point:

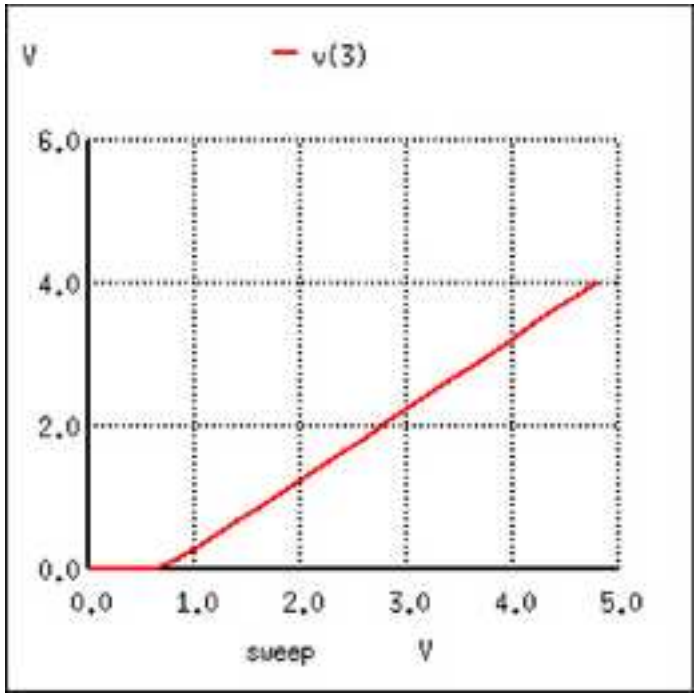


It should be apparent that the load resistor in the common-collector amplifier circuit receives both the base and collector currents, being placed in series with the emitter. Since the emitter lead of a transistor is the one handling the most current (the sum of base and collector currents, since base and collector currents always mesh together to form the emitter current), it would be reasonable to presume that this amplifier will have a very large current gain (maximum output current for minimum input current). This presumption is indeed correct: the current gain for a common-collector amplifier is quite large, larger than any other transistor amplifier configuration. However, this is not necessarily what sets it apart from other amplifier designs.

Let's proceed immediately to a SPICE analysis of this amplifier circuit, and you will be able to immediately see what is unique about this amplifier:



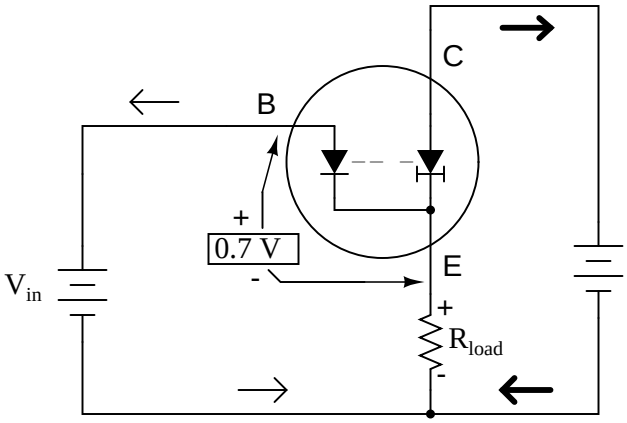
```
common-collector amplifier
vin 1 0
q1 2 1 3 mod1
v1 2 0 dc 15
rload 3 0 5k
.model mod1 npn
.dc vin 0 5 0.2
.plot dc v(3,0)
.end

type        npn
is        1.00E-16
bf         100.000
nf           1.000
br           1.000
nr           1.000
```



Unlike the common-emitter amplifier from the previous section, the common-collector produces an output voltage in *direct* rather than *inverse* proportion to the rising input voltage. As the input voltage increases, so does the output voltage. More than that, a close examination reveals that the output voltage is nearly *identical* to the input voltage, lagging behind only about 0.77 volts.

This is the unique quality of the common-collector amplifier: an output voltage that is nearly equal to the input voltage. Examined from the perspective of output voltage *change* for a given amount of input voltage *change*, this amplifier has a voltage gain of almost exactly unity (1), or 0 dB. This holds true for transistors of any β value, and for load resistors of any resistance value.

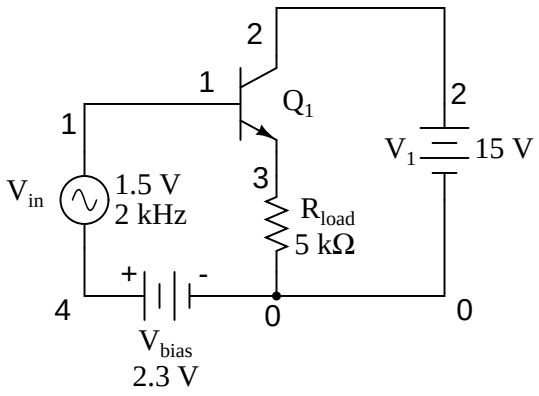
It is simple to understand why the output voltage of a common-collector amplifier is always nearly equal to the input voltage. Referring back to the diode-regulating diode transistor model, we see that the base current must go through the base-emitter PN junction, which is equivalent to a normal rectifying diode. So long as this junction is forward-biased (the transistor conducting current in either its active or saturated modes), it will have a voltage drop of approximately 0.7 volts, assuming silicon construction. This 0.7 volt drop is largely irrespective of the actual magnitude of base current, so we can regard it as being constant:



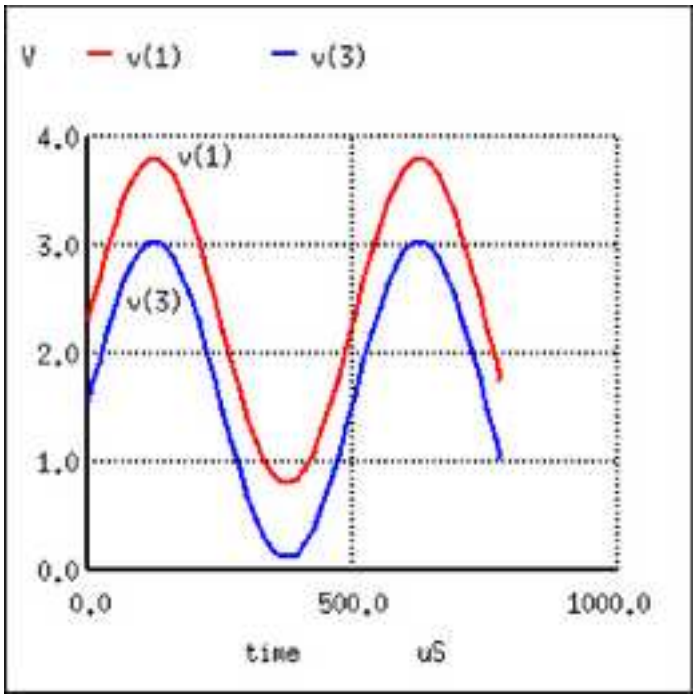
Given the voltage polarities across the base-emitter PN junction and the load resistor, we see that they *must* add together to equal the input voltage, in accordance with Kirchhoff's Voltage Law. In other words, the load voltage will always be about 0.7 volts less than the input voltage for all conditions where the transistor is conducting. Cutoff occurs at input voltages below 0.7 volts, and saturation at input voltages in excess of battery (supply) voltage plus 0.7 volts.

Because of this behavior, the common-collector amplifier circuit is also known as the *voltage-follower* or *emitter-follower* amplifier, in reference to the fact that the input and load voltages follow each other so closely.

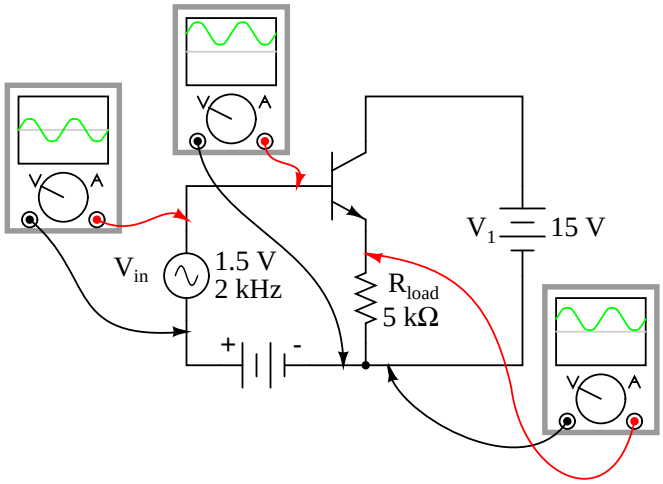
Applying the common-collector circuit to the amplification of AC signals requires the same input "biasing" used in the common-emitter circuit: a DC voltage must be added to the AC input signal to keep the transistor in its active mode during the entire cycle. When this is done, the result is a non-inverting amplifier:




```
common-collector amplifier
vin 1 4 sin(0 1.5 2000 0 0)
vbias 4 0 dc 2.3
q1 2 1 3 mod1
v1 2 0 dc 15
rload 3 0 5k
.model mod1 npn
.tran .02m .78m
.plot tran v(1,0) v(3,0)
.end
```



Here's another view of the circuit, this time with oscilloscopes connected to several points of interest:



Since this amplifier configuration doesn't provide any voltage gain (in fact, in practice it actually has a voltage gain of slightly *less* than 1), its only amplifying factor is current. The common-emitter amplifier configuration examined in the previous section had a current gain equal to the β of the transistor, being that the input current went through the base and the output (load) current went through the collector, and β by definition is the ratio between the collector and base currents. In

the common-collector configuration, though, the load is situated in series with the emitter, and thus its current is equal to the emitter current. With the emitter carrying collector current *and* base current, the load in this type of amplifier has all the current of the collector running through it *plus* the input current of the base. This yields a current gain of β plus 1:

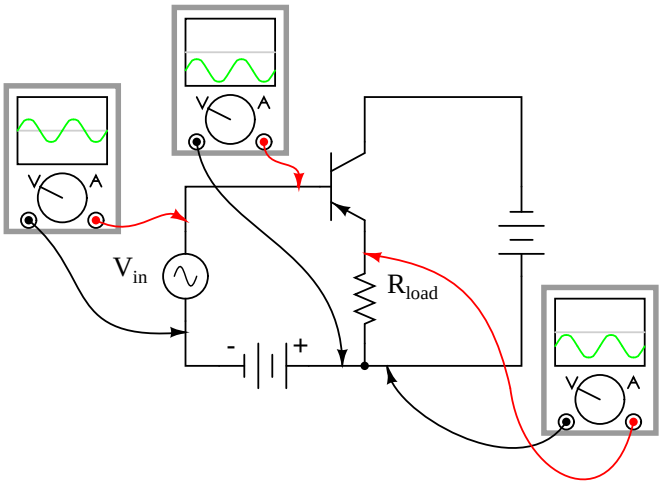
$$A_I = \frac{I_{emitter}}{I_{base}}$$

$$A_I = \frac{I_{collector} + I_{base}}{I_{base}}$$

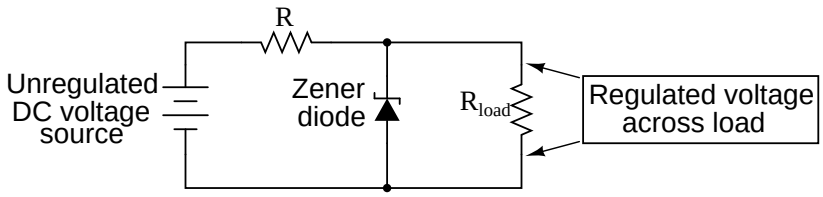
$$A_I = \frac{I_{collector}}{I_{base}} + 1$$

$$A_I = \beta + 1$$

Once again, PNP transistors are just as valid to use in the common-collector configuration as NPN transistors. The gain calculations are all the same, as is the non-inverting behavior of the amplifier. The only difference is in voltage polarities and current directions:

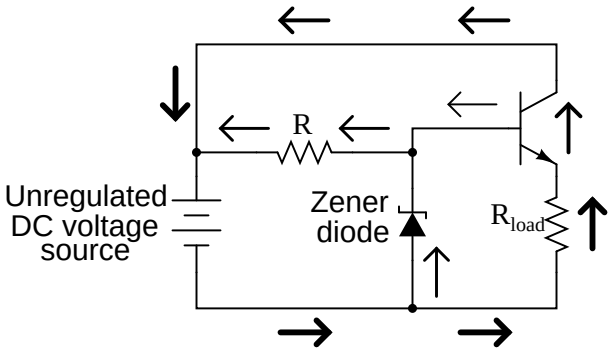


A popular application of the common-collector amplifier is for regulated DC power supplies, where an unregulated (varying) source of DC voltage is clipped at a specified level to supply regulated (steady) voltage to a load. Of course, zener diodes already provide this function of voltage regulation:



However, when used in this direct fashion, the amount of current that may be supplied to the load is usually quite limited. In essence, this circuit regulates voltage across the load by keeping current through the series resistor at a high enough level to drop all the excess power source voltage across it, the zener diode drawing more or less current as necessary to keep the voltage across itself steady. For high-current loads, an plain zener diode voltage regulator would have to be capable of shunting a lot of current through the diode in order to be effective at regulating load voltage in the event of large load resistance or voltage source changes.

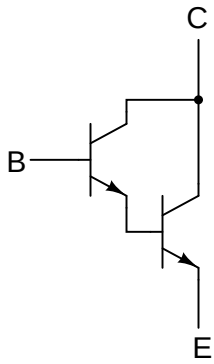
One popular way to increase the current-handling ability of a regulator circuit like this is to use a common-collector transistor to amplify current to the load, so that the zener diode circuit only has to handle the amount of current necessary to drive the base of the transistor:



There's really only one caveat to this approach: the load voltage will be approximately 0.7 volts less than the zener diode voltage, due to the transistor's 0.7 volt base-emitter drop. However, since this 0.7 volt difference is fairly constant over a wide range of load currents, a zener diode with a 0.7 volt higher rating can be chosen for the application.

Sometimes the high current gain of a single-transistor, common-collector configuration isn't enough for a particular application. If this is the case, multiple transistors may be staged together in a popular configuration known as a *Darlington pair*, just an extension of the common-collector concept:

An NPN "Darlington pair"



Darlington pairs essentially place one transistor as the common-collector load for another transistor, thus multiplying their individual current gains. Base current through the upper-left transistor is amplified through that transistor's emitter, which is directly connected to the base of the lower-right transistor, where the current is again amplified. The overall current gain is as follows:

Darlington pair current gain

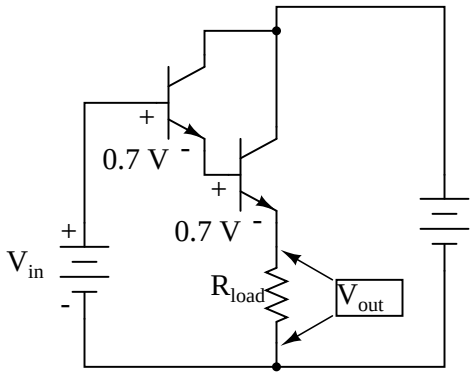
$$A_I = (\beta_1 + 1)(\beta_2 + 1)$$

Where,

β_1 = Beta of first transistor

β_2 = Beta of second transistor

Voltage gain is still nearly equal to 1 if the entire assembly is connected to a load in common-collector fashion, although the load voltage will be a full 1.4 volts less than the input voltage:



$$V_{out} = V_{in} - 1.4$$

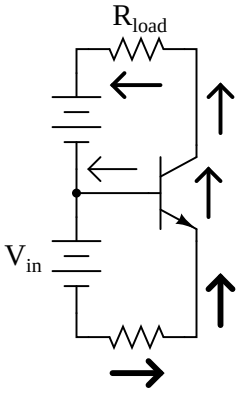
Darlington pairs may be purchased as discrete units (two transistors in the same package), or may be built up from a pair of individual transistors. Of course, if even more current gain is desired than what may be obtained with a pair, Darlington triplet or quadruplet assemblies may be constructed.

- **REVIEW:**
- *Common-collector* transistor amplifiers are so-called because the input and output voltage points share the collector lead of the transistor in common with each other, not considering any power supplies.
- The output voltage on a common-collector amplifier will be in phase with the input voltage, making the common-collector a *non-inverting* amplifier circuit.
- The current gain of a common-collector amplifier is equal to β plus 1. The voltage gain is approximately equal to 1 (in practice, just a little bit less).
- A *Darlington pair* is a pair of transistors "piggybacked" on one another so that the emitter of one feeds current to the base of the other in common-collector form. The result is an overall

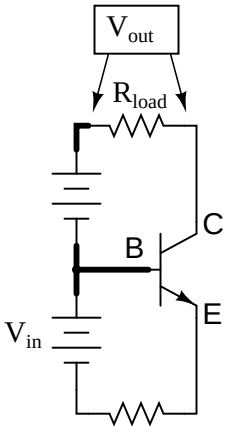
current gain equal to the product (multiplication) of their individual common-collector current gains (β plus 1).

4.7 The common-base amplifier

The final transistor amplifier configuration we need to study is the *common-base*. This configuration is more complex than the other two, and is less common due to its strange operating characteristics.



It is called the *common-base* configuration because (DC power source aside), the signal source and the load share the base of the transistor as a common connection point:

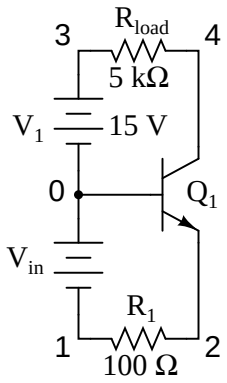


Perhaps the most striking characteristic of this configuration is that the input signal source must carry the full emitter current of the transistor, as indicated by the heavy arrows in the first illustration. As we know, the emitter current is greater than any other current in the transistor, being the sum of base and collector currents. In the last two amplifier configurations, the signal source was connected to the base lead of the transistor, thus handling the *least* current possible.

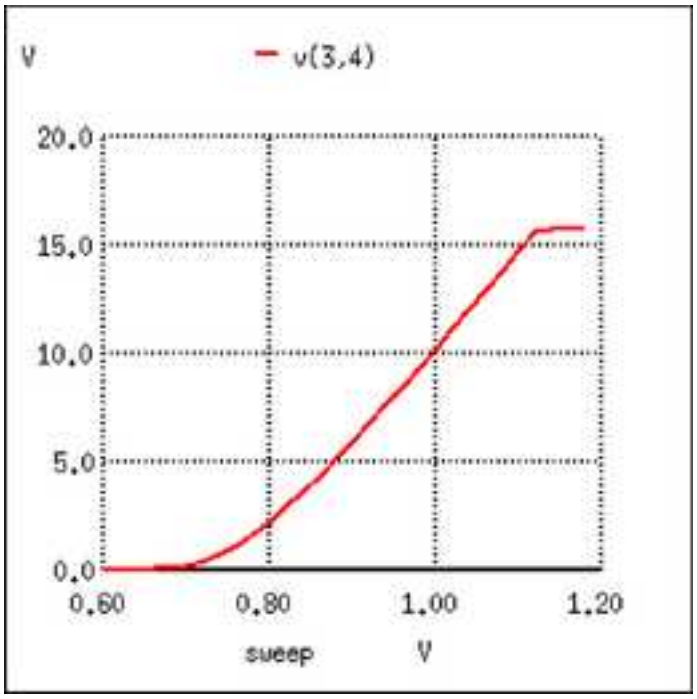
Because the input current exceeds all other currents in the circuit, including the output current, the current gain of this amplifier is actually *less than 1* (notice how R_{load} is connected to the collector, thus carrying slightly less current than the signal source). In other words, it *attenuates* current

rather than *amplifying* it. With common-emitter and common-collector amplifier configurations, the transistor parameter most closely associated with gain was β. In the common-base circuit, we follow another basic transistor parameter: the ratio between collector current and emitter current, which is a fraction always less than 1. This fractional value for any transistor is called the *alpha* ratio, or α ratio.

Since it obviously can't boost signal current, it only seems reasonable to expect it to boost signal voltage. A SPICE simulation will vindicate that assumption:

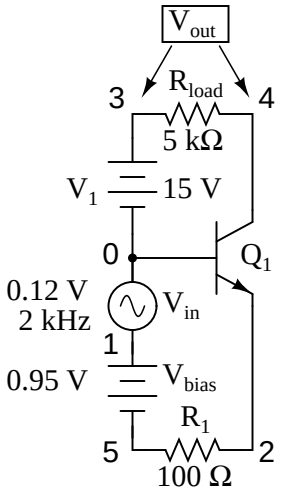



```
common-base amplifier
vin 0 1
r1 1 2 100
q1 4 0 2 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
.dc vin 0.6 1.2 .02
.plot dc v(3,4)
.end
```

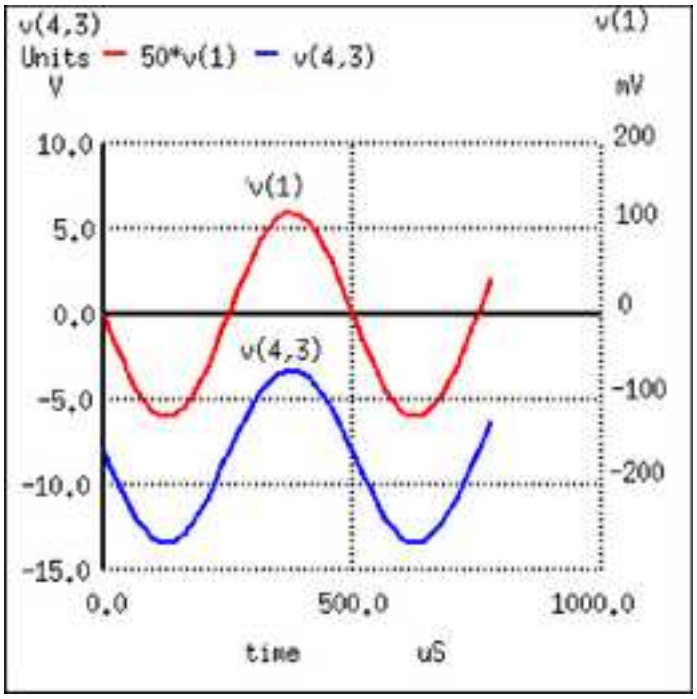


Notice how in this simulation the output voltage goes from practically nothing (cutoff) to 15.75 volts (saturation) with the input voltage being swept over a range of 0.6 volts to 1.2 volts. In fact, the output voltage plot doesn't show a rise until about 0.7 volts at the input, and cuts off (flattens) at about 1.12 volts input. This represents a rather large voltage gain with an output voltage span of 15.75 volts and an input voltage span of only 0.42 volts: a gain ratio of 37.5, or 31.48 dB. Notice also how the output voltage (measured across R_{load}) actually exceeds the power supply (15 volts) at saturation, due to the series-aiding effect of the the input voltage source.

A second set of SPICE analyses with an AC signal source (and DC bias voltage) tells the same story: a high voltage gain.




```
common-base amplifier
vin 0 1 sin (0 0.12 2000 0 0)
vbias 1 5 dc 0.95
r1 5 2 100
q1 4 0 2 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
.tran 0.02m 0.78m
.plot tran v(1,0) v(4,3)
.end
```



As you can see, the input and output waveforms are in phase with each other. This tells us that the common-base amplifier is non-inverting.

```
common-base amplifier
vin 0 1 ac 0.12
vbias 1 5 dc 0.95
r1 5 2 100
q1 4 0 2 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
.ac lin 1 2000 2000
.print ac v(1,0) v(3,4)
.end
```

```
freq          v(1)        v(3,4)
2.000E+03     1.200E-01   5.129E+00
```

Voltage figures from the second analysis (AC mode) show a voltage gain of 42.742 (5.129 V / 0.12 V), or 32.617 dB:

$$A_V = \frac{V_{out}}{V_{in}}$$

$$A_V = \frac{5.129 \text{ V}}{0.12 \text{ V}}$$

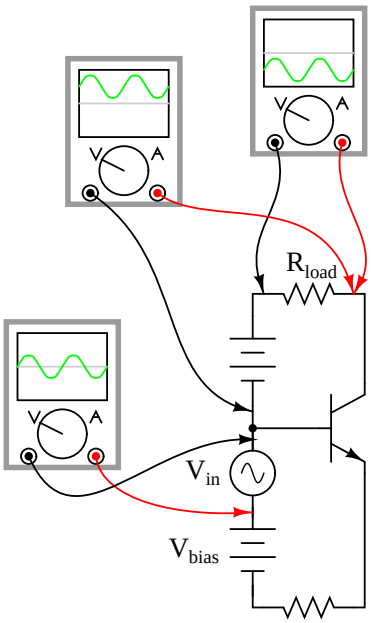
$$A_V = 42.742$$

$$A_{V(dB)} = 20 \log A_{V(ratio)}$$

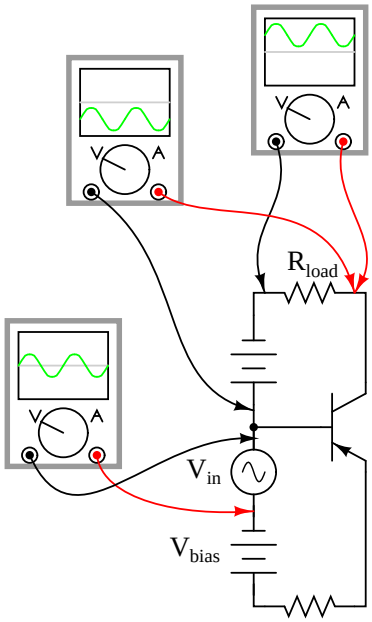
$$A_{V(dB)} = 20 \log 42.742$$

$$A_{V(dB)} = 32.617 \text{ dB}$$

Here's another view of the circuit, showing the phase relations and DC offsets of various signals in the circuit just simulated:



. . . and for a PNP transistor:



Predicting voltage gain for the common-base amplifier configuration is quite difficult, and involves approximations of transistor behavior that are difficult to measure directly. Unlike the other amplifier configurations, where voltage gain was either set by the ratio of two resistors (common-emitter), or fixed at an unchangeable value (common-collector), the voltage gain of the common-base amplifier depends largely on the amount of DC bias on the input signal. As it turns out, the internal transistor resistance between emitter and base plays a major role in determining voltage gain, and this resistance changes with different levels of current through the emitter.

While this phenomenon is difficult to explain, it is rather easy to demonstrate through the use of computer simulation. What I'm going to do here is run several SPICE simulations on a common-base amplifier circuit, changing the DC bias voltage slightly while keeping the AC signal amplitude and all other circuit parameters constant. As the voltage gain changes from one simulation to another, different output voltage amplitudes will be noticed as a result.

Although these analyses will all be conducted in the AC mode, they were first "proofed" in the transient analysis mode (voltage plotted over time) to ensure that the entire wave was being faithfully reproduced and not "clipped" due to improper biasing. No meaningful calculations of gain can be based on waveforms that are distorted:

```
common-base amplifier DC bias = 0.85 volts
vin 0 1 ac 0.08
vbias 1 5 dc 0.85
r1 5 2 100
q1 4 0 2 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
.ac lin 1 2000 2000
```

```
.print ac v(1,0) v(3,4)
.end

freq          v(1)        v(3,4)
2.000E+03     8.000E-02   3.005E+00


common-base amplifier dc bias = 0.9 volts
vin 0 1 ac 0.08
vbias 1 5 dc 0.90
r1 5 2 100
q1 4 0 2 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
.ac lin 1 2000 2000
.print ac v(1,0) v(3,4)
.end

freq          v(1)        v(3,4)
2.000E+03     8.000E-02   3.264E+00


common-base amplifier dc bias = 0.95 volts
vin 0 1 ac 0.08
vbias 1 5 dc 0.95
r1 5 2 100
q1 4 0 2 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
.ac lin 1 2000 2000
.print ac v(1,0) v(3,4)
.end

freq          v(1)        v(3,4)
2.000E+03     8.000E-02   3.419E+00
```

A trend should be evident here: with increases in DC bias voltage, voltage gain increases as well. We can see that the voltage gain is increasing because each subsequent simulation produces greater output voltage for the exact same input signal voltage (0.08 volts). As you can see, the changes are quite large, and they are caused by miniscule variations in bias voltage!

The combination of very low current gain (always less than 1) and somewhat unpredictable voltage gain conspire against the common-base design, relegating it to few practical applications.

- **REVIEW:**

- *Common-base* transistor amplifiers are so-called because the input and output voltage points share the base lead of the transistor in common with each other, not considering any power supplies.

- The current gain of a common-base amplifier is always less than 1. The voltage gain is a function of input and output resistances, and also the internal resistance of the emitter-base junction, which is subject to change with variations in DC bias voltage. Suffice to say that the voltage gain of a common-base amplifier can be very high.

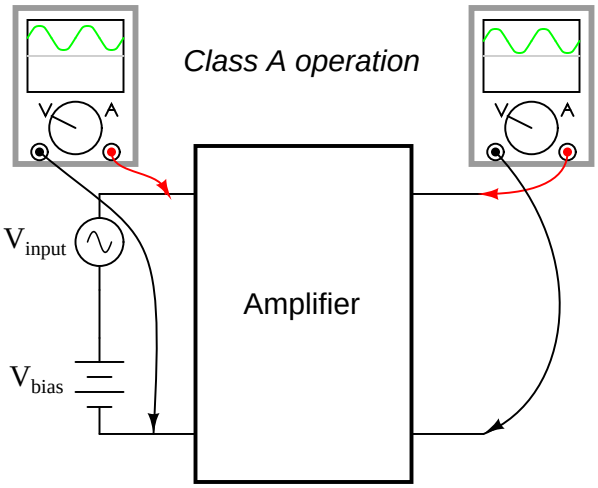
- The ratio of a transistor's collector current to emitter current is called α. The α value for any transistor is always less than unity, or in other words, less than 1.

4.8 Biasing techniques

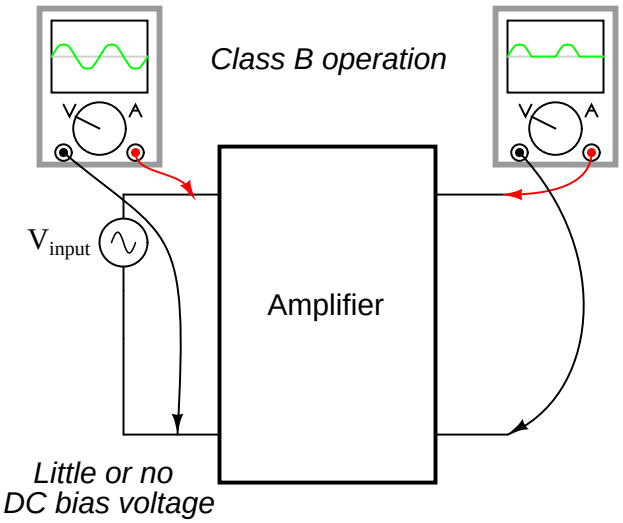
In the common-emitter section of this chapter, we saw a SPICE analysis where the output waveform resembled a half-wave rectified shape: only half of the input waveform was reproduced, with the other half being completely cut off. Since our purpose at that time was to reproduce the entire waveshape, this constituted a problem. The solution to this problem was to add a small bias voltage to the amplifier input so that the transistor stayed in active mode throughout the entire wave cycle. This addition was called a *bias voltage*.

There are applications, though, where a half-wave output is not problematic. In fact, some applications may *necessitate* this very type of amplification. Because it is possible to operate an amplifier in modes other than full-wave reproduction, and because there are specific applications requiring different ranges of reproduction, it is useful to describe the degree to which an amplifier reproduces the input waveform by designating it according to *class*. Amplifier class operation is categorized by means of alphabetical letters: A, B, C, and AB.

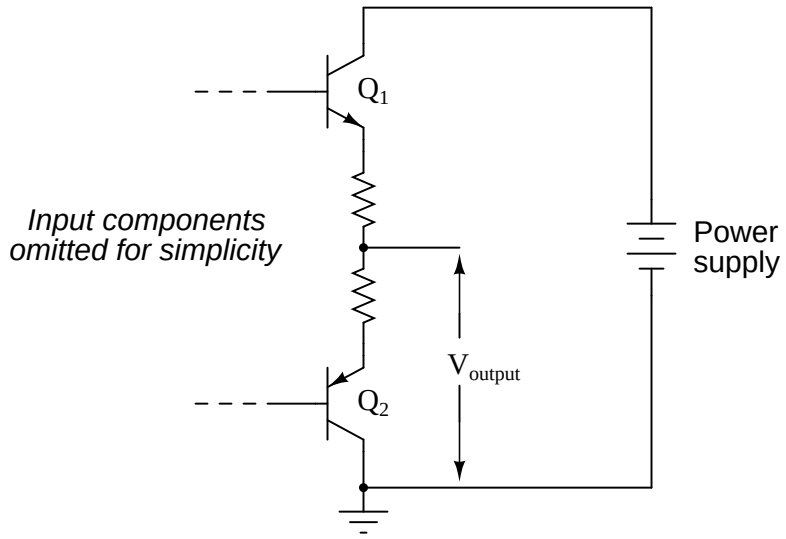
Class A operation is where the entire input waveform is faithfully reproduced. Although I didn't introduce this concept back in the common-emitter section, this is what we were hoping to attain in our simulations. Class A operation can only be obtained when the transistor spends its entire time in the active mode, never reaching either cutoff or saturation. To achieve this, sufficient DC bias voltage is usually set at the level necessary to drive the transistor exactly halfway between cutoff and saturation. This way, the AC input signal will be perfectly "centered" between the amplifier's high and low signal limit levels.



Class B operation is what we had the first time an AC signal was applied to the common-emitter amplifier with no DC bias voltage. The transistor spent half its time in active mode and the other half in cutoff with the input voltage too low (or even of the wrong polarity!) to forward-bias its base-emitter junction.



By itself, an amplifier operating in class B mode is not very useful. In most circumstances, the severe distortion introduced into the waveshape by eliminating half of it would be unacceptable. However, class B operation is a useful mode of biasing if two amplifiers are operated as a *push-pull* pair, each amplifier handling only half of the waveform at a time:

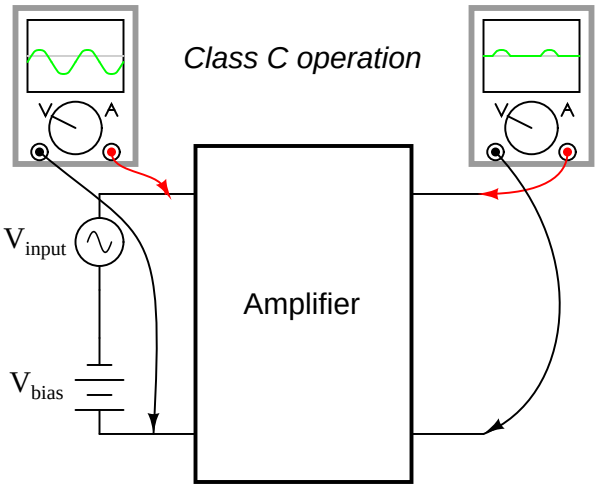


Transistor Q_1 "pushes" (drives the output voltage in a positive direction with respect to ground), while transistor Q_2 "pulls" the output voltage (in a negative direction, toward 0 volts with respect to ground). Individually, each of these transistors is operating in class B mode, active only for one-half of the input waveform cycle. Together, however, they function as a team to produce an output waveform identical in shape to the input waveform.

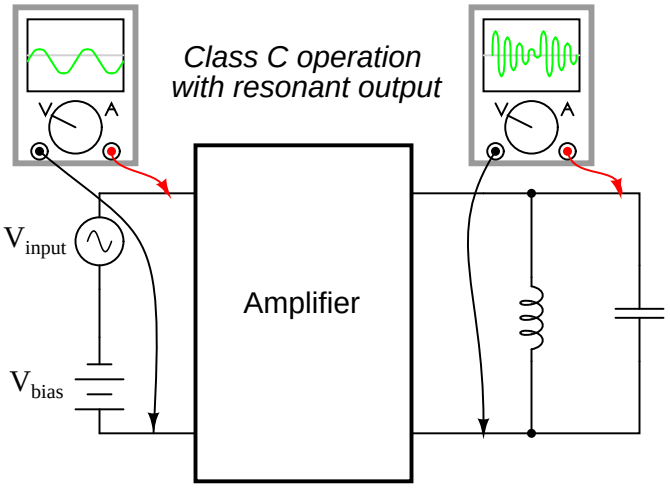
A decided advantage of the class B (push-pull) amplifier design over the class A design is greater output power capability. With a class A design, the transistor dissipates a lot of energy in the form of heat because it never stops conducting current. At all points in the wave cycle it is in the active (conducting) mode, conducting substantial current and dropping substantial voltage. This means there is substantial power dissipated by the transistor throughout the cycle. In a class B design, each transistor spends half the time in cutoff mode, where it dissipates zero power (zero current = zero power dissipation). This gives each transistor a time to "rest" and cool while the other transistor carries the burden of the load. Class A amplifiers are simpler in design, but tend to be limited to low-power signal applications for the simple reason of transistor heat dissipation.

There is another class of amplifier operation known as *class AB*, which is somewhere between class A and class B: the transistor spends more than 50% but less than 100% of the time conducting current.

If the input signal bias for an amplifier is slightly negative (opposite of the bias polarity for class A operation), the output waveform will be further "clipped" than it was with class B biasing, resulting in an operation where the transistor spends the majority of the time in cutoff mode:



At first, this scheme may seem utterly pointless. After all, how useful could an amplifier be if it clips the waveform as badly as this? If the output is used directly with no conditioning of any kind, it would indeed be of questionable utility. However, with the application of a tank circuit (parallel resonant inductor-capacitor combination) to the output, the occasional output surge produced by the amplifier can set in motion a higher-frequency oscillation maintained by the tank circuit. This may be likened to a machine where a heavy flywheel is given an occasional "kick" to keep it spinning:

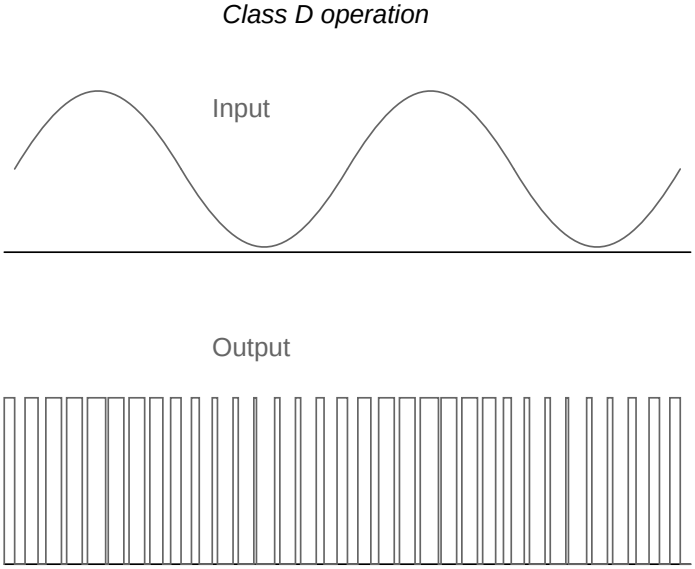


Called *class C* operation, this scheme also enjoys high power efficiency due to the fact that the transistor(s) spend the vast majority of time in the cutoff mode, where they dissipate zero power. The rate of output waveform decay (decreasing oscillation amplitude between "kicks" from the amplifier) is exaggerated here for the benefit of illustration. Because of the tuned tank circuit on the output, this type of circuit is usable only for amplifying signals of definite, fixed frequency.

Another type of amplifier operation, significantly different from Class A, B, AB, or C, is called *Class D*. It is not obtained by applying a specific measure of bias voltage as are the other classes of operation, but requires a radical re-design of the amplifier circuit itself. It's a little too early in this

chapter to investigate exactly how a class D amplifier is built, but not too early to discuss its basic principle of operation.

A class D amplifier reproduces the profile of the input voltage waveform by generating a rapidly-pulsing squarewave output. The duty cycle of this output waveform (time "on" versus total cycle time) varies with the instantaneous amplitude of the input signal. The following plots demonstrate this principle:

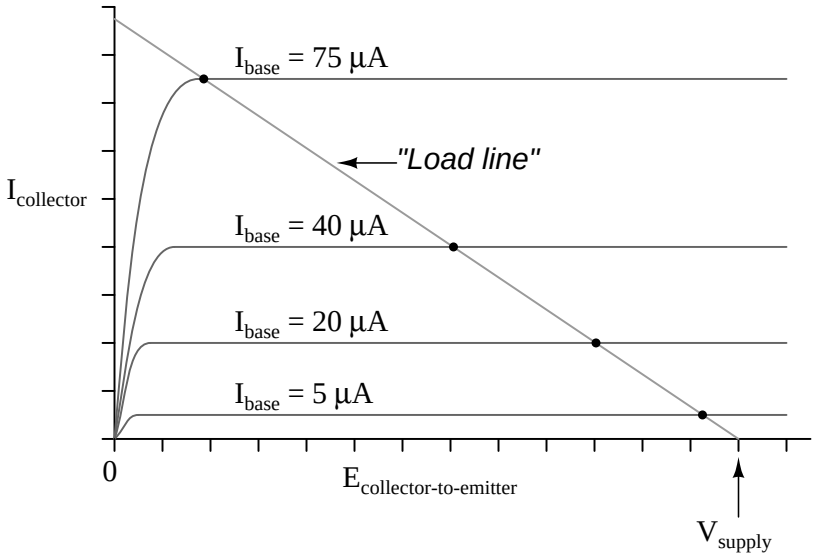


The greater the instantaneous voltage of the input signal, the greater the duty cycle of the output squarewave pulse. If there can be any goal stated of the class D design, it is to avoid active-mode transistor operation. Since the output transistor of a class D amplifier is never in the active mode, only cutoff or saturated, there will be little heat energy dissipated by it. This results in very high power efficiency for the amplifier. Of course, the disadvantage of this strategy is the overwhelming presence of harmonics on the output. Fortunately, since these harmonic frequencies are typically much greater than the frequency of the input signal, they can be filtered out by a low-pass filter with relative ease, resulting in an output more closely resembling the original input signal waveform. Class D technology is typically seen where extremely high power levels and relatively low frequencies are encountered, such as in industrial inverters (devices converting DC into AC power to run motors and other large devices) and high-performance audio amplifiers.

A term you will likely come across in your studies of electronics is something called *quiescent*, which is a modifier designating the normal, or zero input signal, condition of a circuit. Quiescent current, for example, is the amount of current in a circuit with zero input signal voltage applied. Bias voltage in a transistor circuit forces the transistor to operate at a different level of collector current with zero input signal voltage than it would without that bias voltage. Therefore, the amount of bias in an amplifier circuit determines its quiescent values.

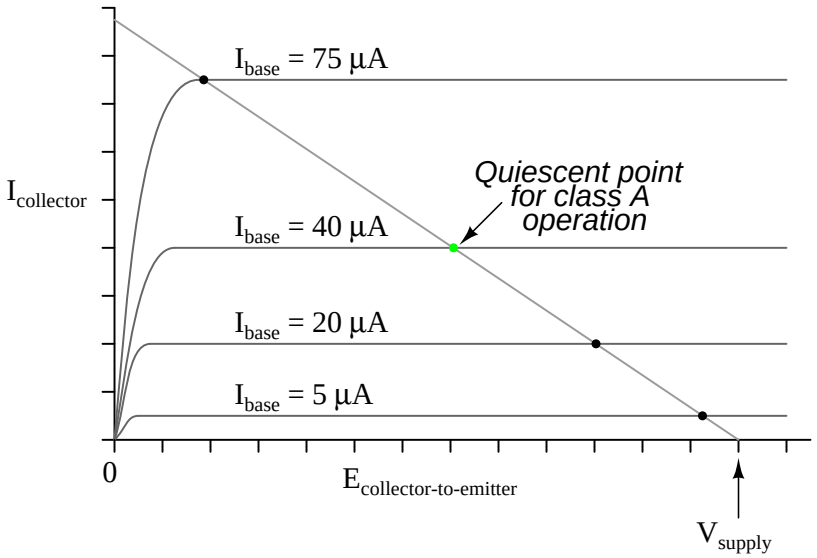
In a class A amplifier, the quiescent current should be exactly half of its saturation value (halfway between saturation and cutoff, cutoff by definition being zero). Class B and class C amplifiers have quiescent current values of zero, since they are supposed to be cutoff with no signal applied. Class AB amplifiers have very low quiescent current values, just above cutoff. To illustrate this graphically,

a "load line" is sometimes plotted over a transistor's characteristic curves to illustrate its range of operation while connected to a load resistance of specific value:

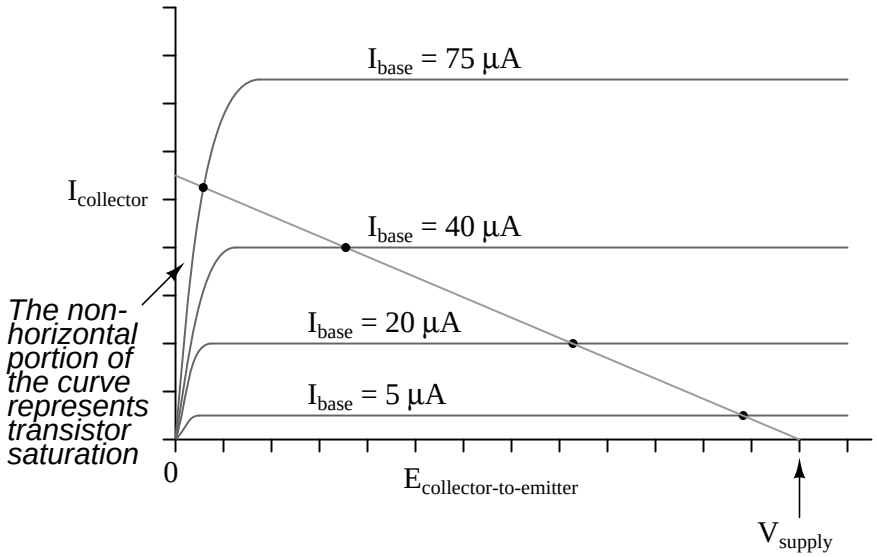


A load line is a plot of collector-to-emitter voltage over a range of base currents. At the lower-right corner of the load line, voltage is at maximum and current is at zero, representing a condition of cutoff. At the upper-left corner of the line, voltage is at zero while current is at a maximum, representing a condition of saturation. Dots marking where the load line intersects the various transistor curves represent realistic operating conditions for those base currents given.

Quiescent operating conditions may be shown on this type of graph in the form of a single dot along the load line. For a class A amplifier, the quiescent point will be in the middle of the load line, like this:



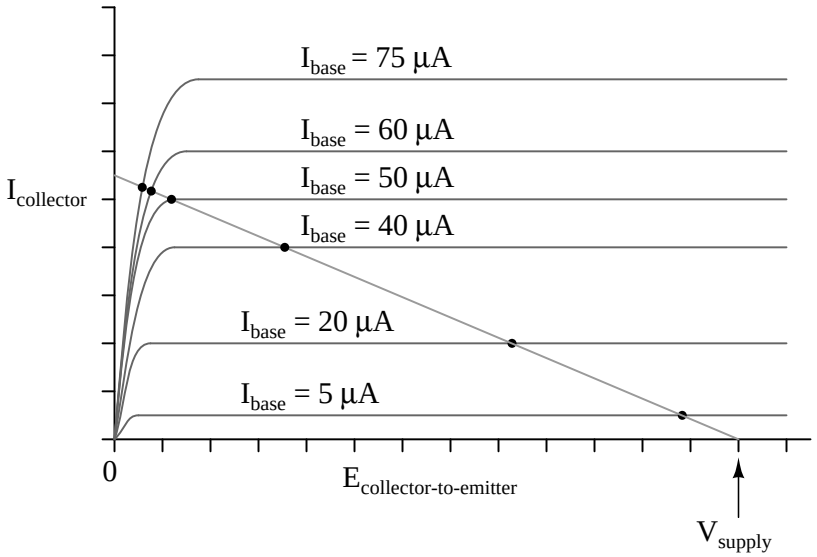
In this illustration, the quiescent point happens to fall on the curve representing a base current of 40 μA. If we were to change the load resistance in this circuit to a greater value, it would affect the slope of the load line, since a greater load resistance would limit the maximum collector current at saturation, but would not change the collector-emitter voltage at cutoff. Graphically, the result is a load line with a different upper-left point and the same lower-right point:



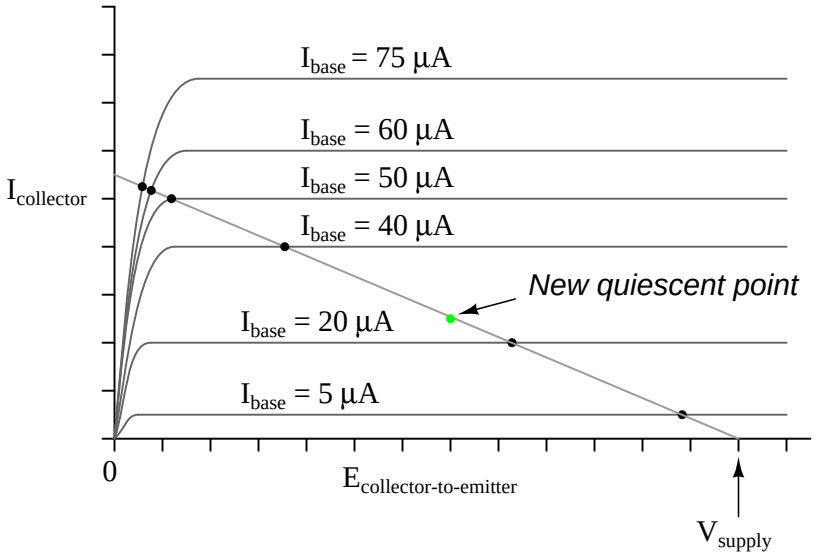
Note how the new load line doesn't intercept the 75 μA curve along its flat portion as before. This is very important to realize because the non-horizontal portion of a characteristic curve represents a condition of saturation. Having the load line intercept the 75 μA curve outside of the curve's horizontal range means that the amplifier will be saturated at that amount of base current. Increasing the load resistor value is what caused the load line to intercept the 75 μA curve at this new point, and it indicates that saturation will occur at a lesser value of base current than before.

With the old, lower-value load resistor in the circuit, a base current of 75 μA would yield a proportional collector current (base current multiplied by β). In the first load line graph, a base current of 75 μA gave a collector current almost twice what was obtained at 40 μA, as the β ratio would predict. Now, however, there is only a marginal increase in collector current between base current values of 75 μA and 40 μA, because the transistor begins to lose sufficient collector-emitter voltage to continue to regulate collector current.

In order to maintain linear (no-distortion) operation, transistor amplifiers shouldn't be operated at points where the transistor will saturate; that is, in any case where the load line will not potentially fall on the horizontal portion of a collector current curve. In this case, we'd have to add a few more curves to the graph before we could tell just how far we could "push" this transistor with increased base currents before it saturates.



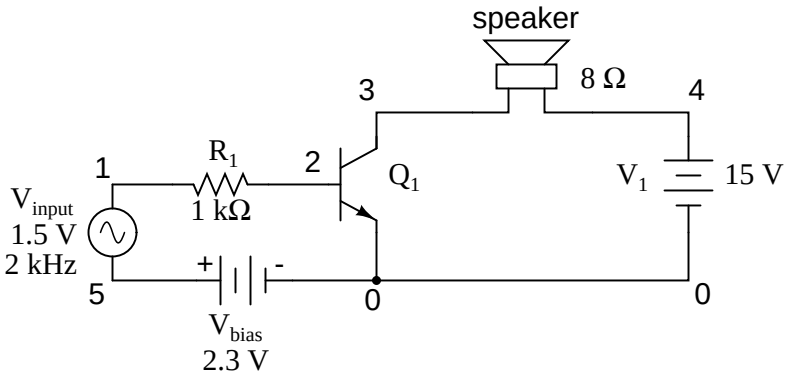
It appears in this graph that the highest-current point on the load line falling on the straight portion of a curve is the point on the 50 μA curve. This new point should be considered the maximum allowable input signal level for class A operation. Also for class A operation, the bias should be set so that the quiescent point is halfway between this new maximum point and cutoff:



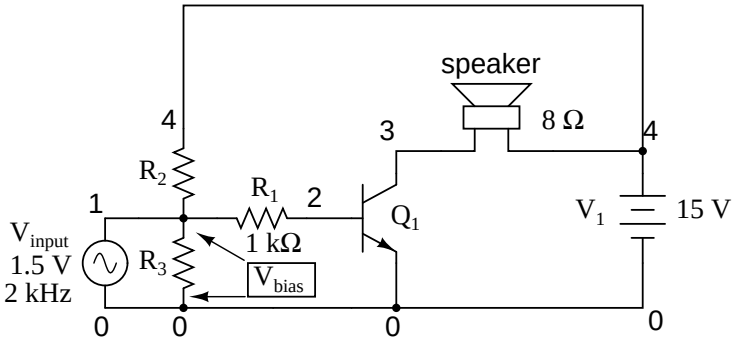
Now that we know a little more about the consequences of different DC bias voltage levels, it is time to investigate practical biasing techniques. So far, I've shown a small DC voltage source (battery) connected in series with the AC input signal to bias the amplifier for whatever desired class of operation. In real life, the connection of a precisely-calibrated battery to the input of an amplifier is simply not practical. Even if it were possible to customize a battery to produce just the right amount of voltage for any given bias requirement, that battery would not remain at its

manufactured voltage indefinitely. Once it started to discharge and its output voltage drooped, the amplifier would begin to drift in the direction of class B operation.

Take this circuit, illustrated in the common-emitter section for a SPICE simulation, for instance:

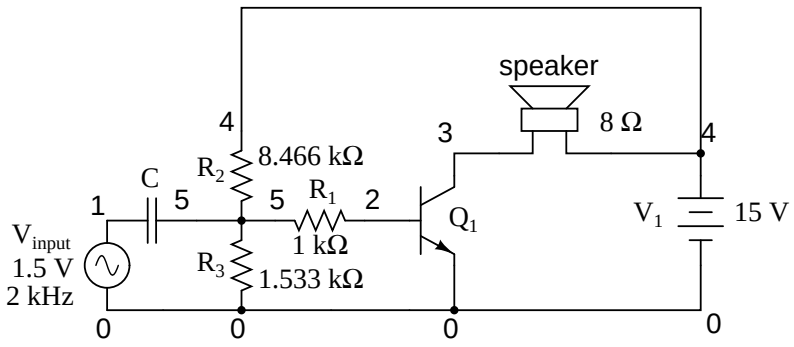


That 2.3 volt "V_{bias}" battery would not be practical to include in a real amplifier circuit. A far more practical method of obtaining bias voltage for this amplifier would be to develop the necessary 2.3 volts using a voltage divider network connected across the 15 volt battery. After all, the 15 volt battery is already there by necessity, and voltage divider circuits are very easy to design and build. Let's see how this might look:



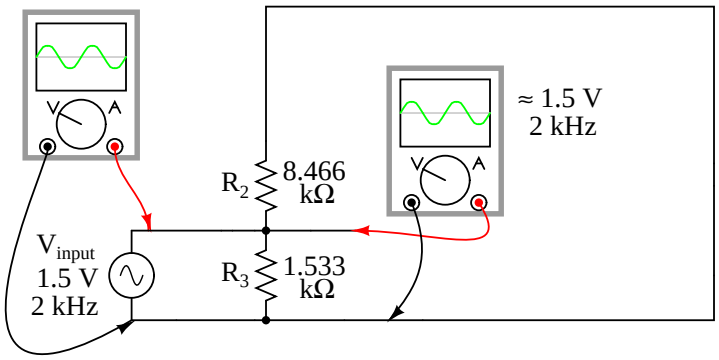
If we choose a pair of resistor values for R_2 and R_3 that will produce 2.3 volts across R_3 from a total of 15 volts (such as 8466 Ω for R_2 and 1533 Ω for R_3), we should have our desired value of 2.3 volts between base and emitter for biasing with no signal input. The only problem is, this circuit configuration places the AC input signal source directly in parallel with R_3 of our voltage divider. This is not acceptable, as the AC source will tend to overpower any DC voltage dropped across R_3. Parallel components *must* have the same voltage, so if an AC voltage source is directly connected across one resistor of a DC voltage divider, the AC source will "win" and there will be no DC bias voltage added to the signal.

One way to make this scheme work, although it may not be obvious *why* it will work, is to place a *coupling capacitor* between the AC voltage source and the voltage divider like this:



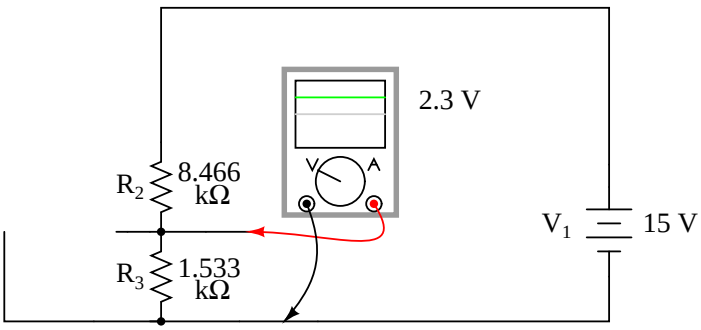
The capacitor forms a high-pass filter between the AC source and the DC voltage divider, passing almost all of the AC signal voltage on to the transistor while blocking all DC voltage from being shorted through the AC signal source. This makes much more sense if you understand the superposition theorem and how it works. According to superposition, any linear, bilateral circuit can be analyzed in a piecemeal fashion by only considering one power source at a time, then algebraically adding the effects of all power sources to find the final result. If we were to separate the capacitor and $R_2 - - R_3$ voltage divider circuit from the rest of the amplifier, it might be easier to understand how this superposition of AC and DC would work.

With only the AC signal source in effect, and a capacitor with an arbitrarily low impedance at signal frequency, almost all the AC voltage appears across R_3:



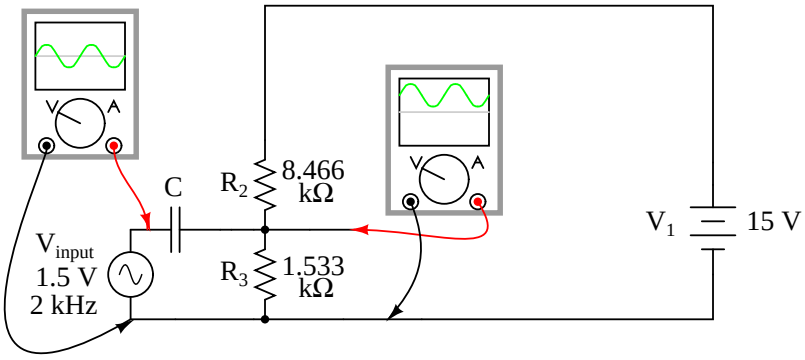
Due to the capacitor's very low impedance at signal frequency, it behaves much like a straight piece of wire and thus can be omitted for the purpose of this step in superposition analysis.

With only the DC source in effect, the capacitor appears to be an open circuit, and thus neither it nor the shorted AC signal source will have any effect on the operation of the $R_2 - - R_3$ voltage divider:



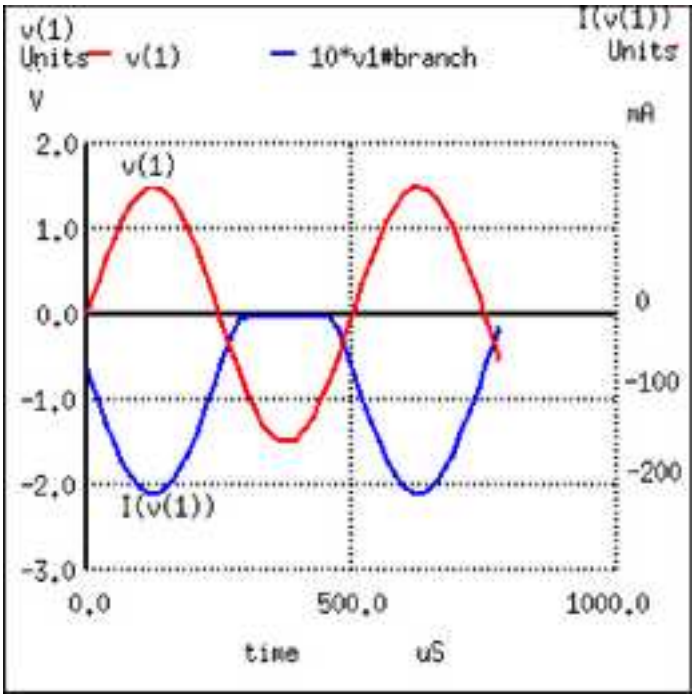
The capacitor appears to be an open circuit as far as DC analysis is concerned

Combining these two separate analyses, we get a superposition of (almost) 1.5 volts AC and 2.3 volts DC, ready to be connected to the base of the transistor:



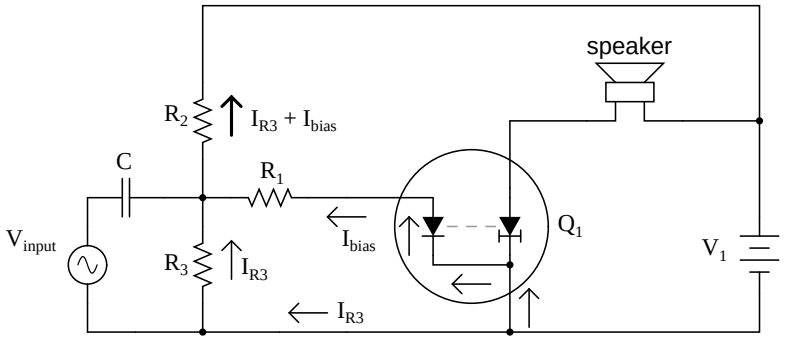
Enough talk – it's about time for a SPICE simulation of the whole amplifier circuit. I'll use a capacitor value of 100 μF to obtain an arbitrarily low (0.796 Ω) impedance at 2000 Hz:

```
voltage divider biasing
vinput 1 0 sin (0 1.5 2000 0 0)
c1 1 5 100u
r1 5 2 1k
r2 4 5 8466
r3 5 0 1533
q1 3 2 0 mod1
rspkr 3 4 8
v1 4 0 dc 15
.model mod1 npn
.tran 0.02m 0.78m
.plot tran v(1,0) i(v1)
.end
```



Notice that there is substantial distortion in the output waveform here: the sine wave is being clipped during most of the input signal's negative half-cycle. This tells us the transistor is entering into cutoff mode when it shouldn't (I'm assuming a goal of class A operation as before). Why is this? This new biasing technique should give us exactly the same amount of DC bias voltage as before, right?

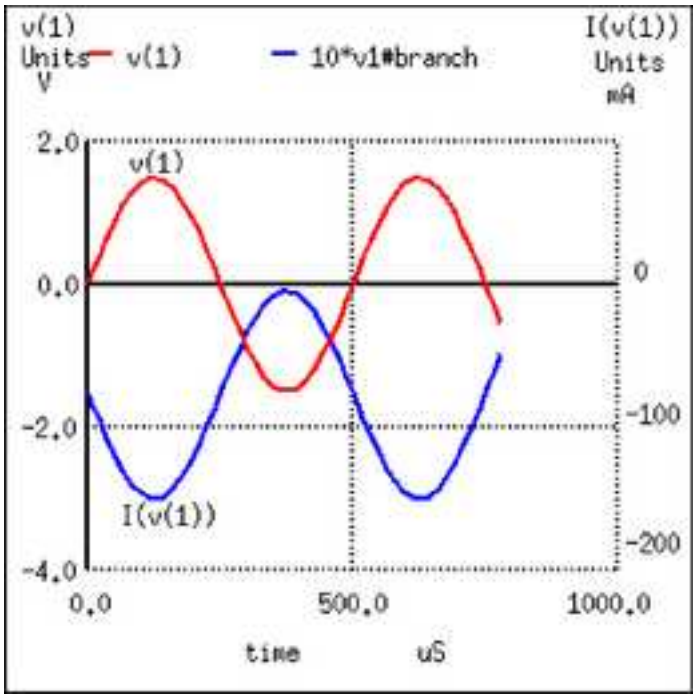
With the capacitor and $R_2 - - R_3$ resistor network unloaded, it will provide exactly 2.3 volts worth of DC bias. However, once we connect this network to the transistor, it is no longer unloaded. Current drawn through the base of the transistor will load the voltage divider, thus reducing the DC bias voltage available for the transistor. Using the diode-regulating diode transistor model to illustrate, the bias problem becomes evident:



A voltage divider's output depends not only on the size of its constituent resistors, but also on

how much current is being divided away from it through a load. In this case, the base-emitter PN junction of the transistor is a load that decreases the DC voltage dropped across R_3, due to the fact that the bias current joins with R_3's current to go through R_2, upsetting the divider ratio formerly set by the resistance values of R_2 and R_3. In order to obtain a DC bias voltage of 2.3 volts, the values of R_2 and/or R_3 must be adjusted to compensate for the effect of base current loading. In this case, we want to *increase* the DC voltage dropped across R_3, so we can lower the value of R_2, raise the value of R_3, or both.

```
voltage divider biasing
vinput 1 0 sin (0 1.5 2000 0 0)
c1 1 5 100u
r1 5 2 1k
r2 4 5 6k       <--- R2 decreased to 6 k ohms
r3 5 0 4k       <--- R3 increased to 4 k ohms
q1 3 2 0 mod1
rspkr 3 4 8
v1 4 0 dc 15
.model mod1 npn
.tran 0.02m 0.78m
.plot tran v(1,0) i(v1)
.end
```



As you can see, the new resistor values of 6 kΩ and 4 kΩ (R_2 and R_3, respectively) results in class A waveform reproduction, just the way we wanted.

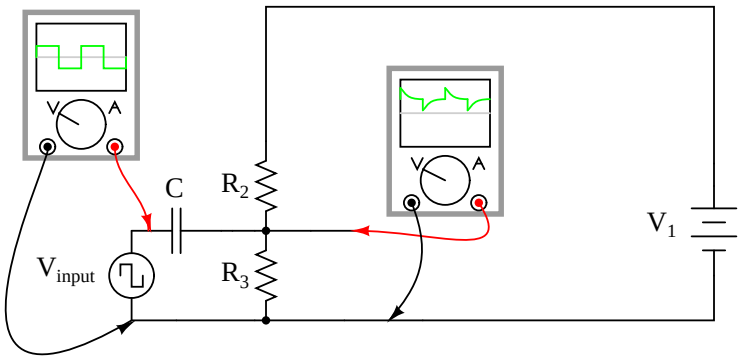
- **REVIEW:**
- *Class A* operation is where an amplifier is biased so as to be in the active mode throughout the entire waveform cycle, thus faithfully reproducing the whole waveform.
- *Class B* operation is where an amplifier is biased so that only half of the input waveform gets reproduced: either the positive half or the negative half. The transistor spends half its time in the active mode and half its time cutoff. Complementary pairs of transistors running in class B operation are often used to deliver high power amplification in audio signal systems, each transistor of the pair handling a separate half of the waveform cycle. Class B operation delivers better power efficiency than a class A amplifier of similar output power.
- *Class AB* operation is where an amplifier is biased at a point somewhere between class A and class B.
- *Class C* operation is where an amplifier's bias forces it to amplify only a small portion of the waveform. A majority of the transistor's time is spent in cutoff mode. In order for there to be a complete waveform at the output, a resonant tank circuit is often used as a "flywheel" to maintain oscillations for a few cycles after each "kick" from the amplifier. Because the transistor is not conducting most of the time, power efficiencies are very high for a class C amplifier.
- *Class D* operation requires an advanced circuit design, and functions on the principle of representing instantaneous input signal amplitude by the duty cycle of a high-frequency squarewave. The output transistor(s) never operate in active mode, only cutoff and saturation. Thus, there is very little heat energy dissipated and energy efficiency is high.
- DC bias voltage on the input signal, necessary for certain classes of operation (especially class A and class C), may be obtained through the use of a voltage divider and *coupling capacitor* rather than a battery connected in series with the AC signal source.

4.9 Input and output coupling

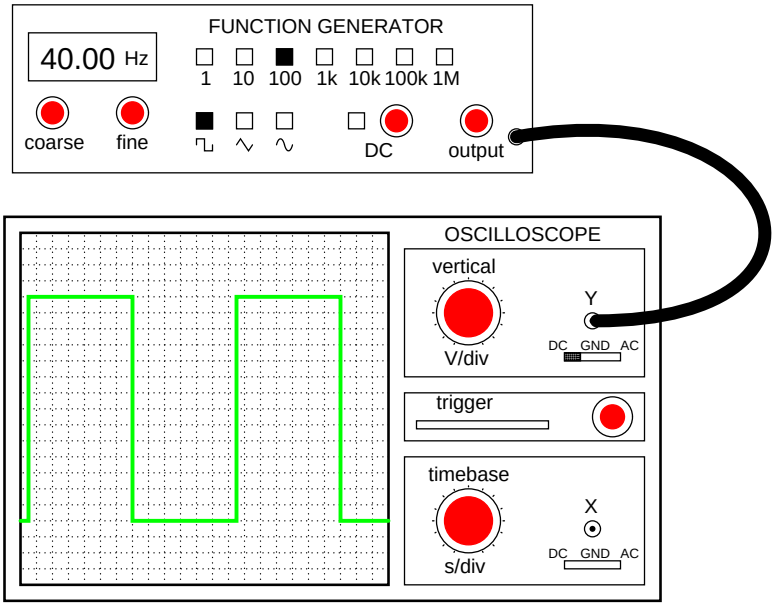
To overcome the challenge of creating necessary DC bias voltage for an amplifier's input signal without resorting to the insertion of a battery in series with the AC signal source, we used a voltage divider connected across the DC power source. To make this work in conjunction with an AC input signal, we "coupled" the signal source to the divider through a capacitor, which acted as a high-pass filter. With that filtering in place, the low impedance of the AC signal source couldn't "short out" the DC voltage dropped across the bottom resistor of the voltage divider. A simple solution, but not without any disadvantages.

Most obvious is the fact that using a high-pass filter capacitor to couple the signal source to the amplifier means that the amplifier can only amplify AC signals. A steady, DC voltage applied to the input would be blocked by the coupling capacitor just as much as the voltage divider bias voltage is blocked from the input source. Furthermore, since capacitive reactance is frequency-dependent, lower-frequency AC signals will not be amplified as much as higher-frequency signals.

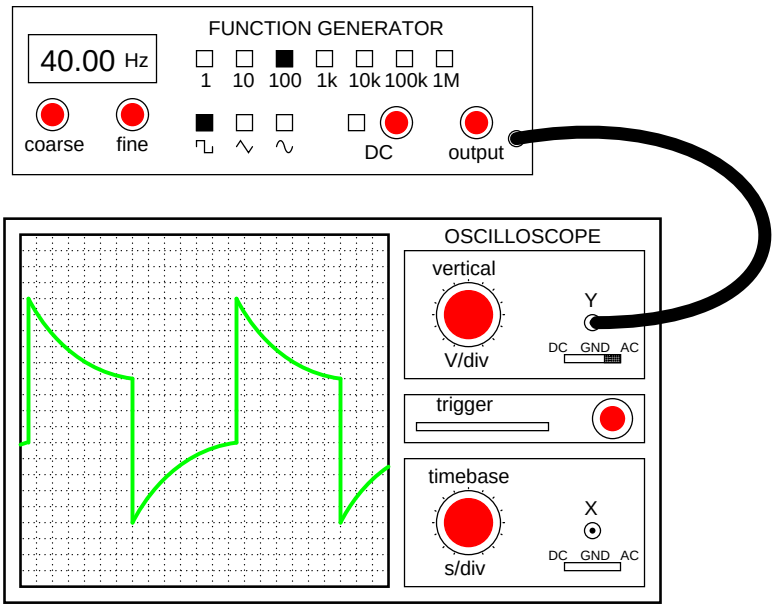
Non-sinusoidal signals will tend to be distorted, as the capacitor responds differently to each of the signal's constituent harmonics. An extreme example of this would be a low-frequency square-wave signal:



Incidentally, this same problem occurs when oscilloscope inputs are set to the "AC coupling" mode. In this mode, a coupling capacitor is inserted in series with the measured voltage signal to eliminate any vertical offset of the displayed waveform due to DC voltage combined with the signal. This works fine when the AC component of the measured signal is of a fairly high frequency, and the capacitor offers little impedance to the signal. However, if the signal is of a low frequency, and/or contains considerable levels of harmonics over a wide frequency range, the oscilloscope's display of the waveform will not be accurate.

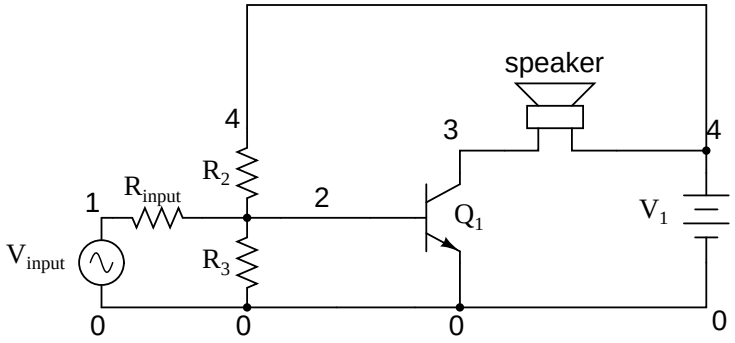


With DC coupling, the oscilloscope properly indicates the shape of the square wave coming from the signal generator.



With AC coupling, the high-pass filtering of the coupling capacitor distorts the square wave's shape so that what is seen is not an accurate representation of the real voltage signal.

In applications where the limitations of capacitive coupling would be intolerable, another solution may be used: *direct coupling*. Direct coupling avoids the use of capacitors or any other frequency-dependent coupling component in favor of resistors. A direct-coupled amplifier circuit might look something like this:



With no capacitor to filter the input signal, this form of coupling exhibits no frequency dependence. DC and AC signals alike will be amplified by the transistor with the same gain (the transistor itself may tend to amplify some frequencies better than others, but that is another subject entirely!).

If direct coupling works for DC as well as for AC signals, then why use capacitive coupling for *any* application? One reason might be to avoid any *unwanted* DC bias voltage naturally present in the signal to be amplified. Some AC signals may be superimposed on an uncontrolled DC voltage right from the source, and an uncontrolled DC voltage would make reliable transistor biasing

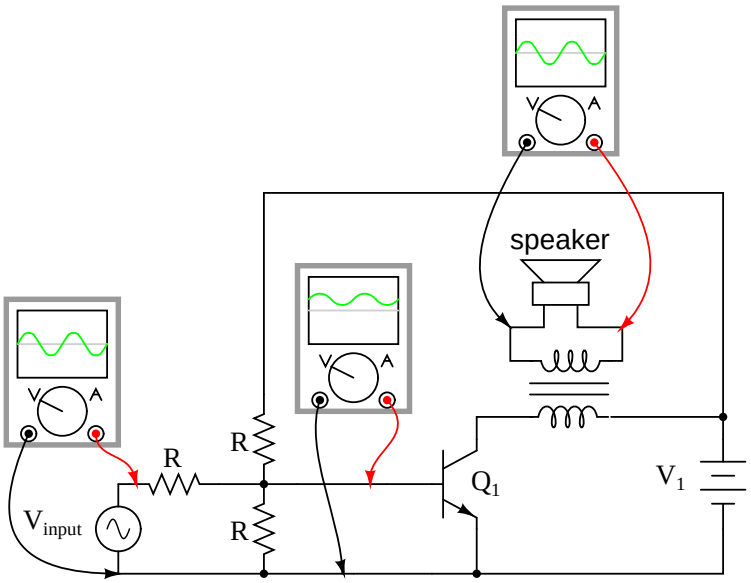
impossible. The high-pass filtering offered by a coupling capacitor would work well here to avoid biasing problems.

Another reason to use capacitive coupling rather than direct is its relative lack of signal attenuation. Direct coupling through a resistor has the disadvantage of diminishing, or attenuating, the input signal so that only a fraction of it reaches the base of the transistor. In many applications, some attenuation is necessary anyway to prevent normal signal levels from "overdriving" the transistor into cutoff and saturation, so any attenuation inherent to the coupling network is useful anyway. However, some applications require that there be *no* signal loss from the input connection to the transistor's base for maximum voltage gain, and a direct coupling scheme with a voltage divider for bias simply won't suffice.

So far, we've discussed a couple of methods for coupling an *input* signal to an amplifier, but haven't addressed the issue of coupling an amplifier's *output* to a load. The example circuit used to illustrate input coupling will serve well to illustrate the issues involved with output coupling.

In our example circuit, the load is a speaker. Most speakers are electromagnetic in design: that is, they use the force generated by an lightweight electromagnet coil suspended within a strong permanent-magnet field to move a thin paper or plastic cone, producing vibrations in the air which our ears interpret as sound. An applied voltage of one polarity moves the cone outward, while a voltage of the opposite polarity will move the cone inward. To exploit cone's full freedom of motion, the speaker must receive true (unbiased) AC voltage. DC bias applied to the speaker coil tends to offset the cone from its natural center position, and this tends to limit the amount of back-and-forth motion it can sustain from the applied AC voltage without overtraveling. However, our example circuit applies a varying voltage of only *one* polarity across the speaker, because the speaker is connected in series with the transistor which can only conduct current one way. This situation would be unacceptable in the case of any high-power audio amplifier.

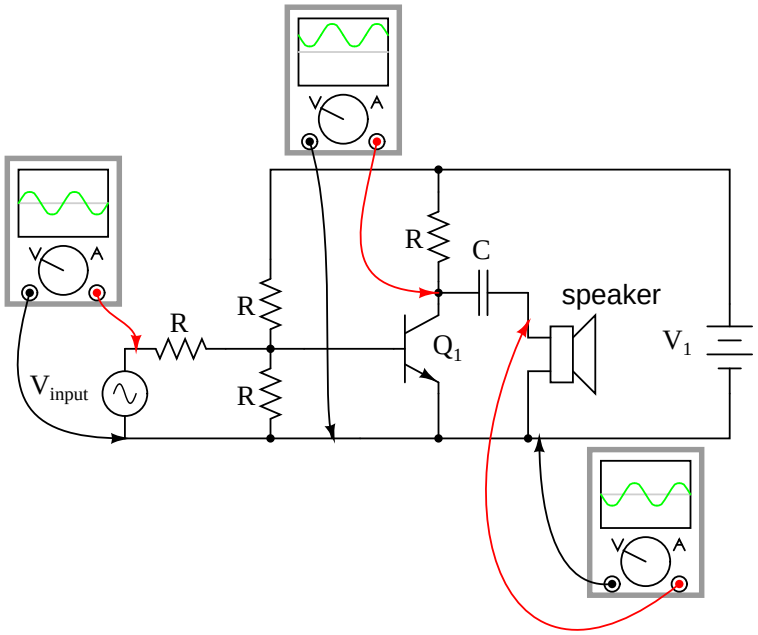
Somehow we need to isolate the speaker from the DC bias of the collector current so that it only receives AC voltage. One way to achieve this goal is to couple the transistor collector circuit to the speaker through a transformer:



Voltage induced in the secondary (speaker-side) of the transformer will be strictly due to *variations* in collector current, because the mutual inductance of a transformer only works on *changes* in winding current. In other words, only the AC portion of the collector current signal will be coupled to the secondary side for powering the speaker. The speaker will "see" true alternating current at its terminals, without any DC bias.

Transformer output coupling works, and has the added benefit of being able to provide impedance matching between the transistor circuit and the speaker coil with custom winding ratios. However, transformers tend to be large and heavy, especially for high-power applications. Also, it is difficult to engineer a transformer to handle signals over a wide range of frequencies, which is almost always required for audio applications. To make matters worse, DC current through the primary winding adds to the magnetization of the core in one polarity only, which tends to make the transformer core saturate more easily in one AC polarity cycle than the other. This problem is reminiscent of having the speaker directly connected in series with the transistor: a DC bias current tends to limit how much output signal amplitude the system can handle without distortion. Generally, though, a transformer can be designed to handle a lot more DC bias current than a speaker without running into trouble, so transformer coupling is still a viable solution in most cases.

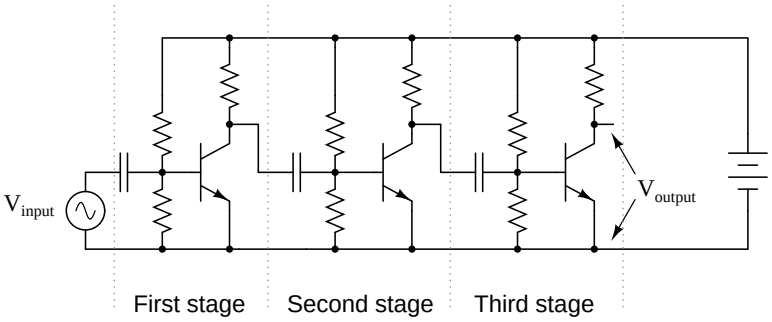
Another method to isolate the speaker from DC bias in the output signal is to alter the circuit a bit and use a coupling capacitor in a manner similar to coupling the input signal to the amplifier:



This circuit resembles the more conventional form of common-emitter amplifier, with the transistor collector connected to the battery through a resistor. The capacitor acts as a high-pass filter, passing most of the AC voltage to the speaker while blocking all DC voltage. Again, the value of this coupling capacitor is chosen so that its impedance at the expected signal frequency will be arbitrarily low.

The blocking of DC voltage from an amplifier's output, be it via a transformer or a capacitor, is useful not only in coupling an amplifier to a load, but also in coupling one amplifier to another amplifier. "Staged" amplifiers are often used to achieve higher power gains than what would be possible using a single transistor:

Three-stage common-emitter amplifier

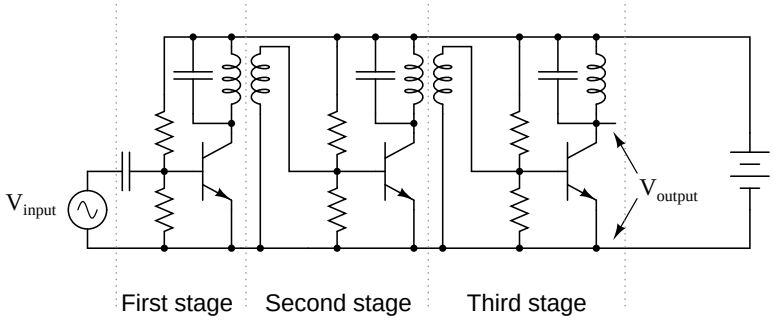


While it is possible to directly couple each stage to the next (via a resistor rather than a capacitor), this makes the whole amplifier *very* sensitive to variations in the DC bias voltage of the first stage, since that DC voltage will be amplified along with the AC signal until the last stage. In other

words, the biasing of the first stage will affect the biasing of the second stage, and so on. However, if the stages are capacitively coupled as shown in the above illustration, the biasing of one stage has no effect on the biasing of the next, because DC voltage is blocked from passing on to the next stage.

Transformer coupling between amplifier stages is also a possibility, but less often seen due to some of the problems inherent to transformers mentioned previously. One notable exception to this rule is in the case of radio-frequency amplifiers where coupling transformers are typically small, have air cores (making them immune to saturation effects), and can be made part of a resonant circuit so as to block unwanted harmonic frequencies from passing on to subsequent stages. The use of resonant circuits assumes that the signal frequency remains constant, of course, but this is typically the case in radio circuitry. Also, the "flywheel" effect of LC tank circuits allows for class C operation for high efficiency:

Three-stage tuned (RF) amplifier



Having said all this, it must be mentioned that it *is* possible to use direct coupling within a multi-stage transistor amplifier circuit. In cases where the amplifier is expected to handle DC signals, this is the only alternative.

- **REVIEW:**

- Capacitive coupling acts like a high-pass filter on the input of an amplifier. This tends to make the amplifier's voltage gain decrease at lower signal frequencies. Capacitive-coupled amplifiers are all but unresponsive to DC input signals.

- Direct coupling with a series resistor instead of a series capacitor avoids the problem of frequency-dependent gain, but has the disadvantage of reducing amplifier gain for all signal frequencies by attenuating the input signal.

- Transformers and capacitors may be used to couple the output of an amplifier to a load, to eliminate DC voltage from getting to the load.

- Multi-stage amplifiers often make use of capacitive coupling between stages to eliminate problems with the bias from one stage affecting the bias of another.

4.10 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.

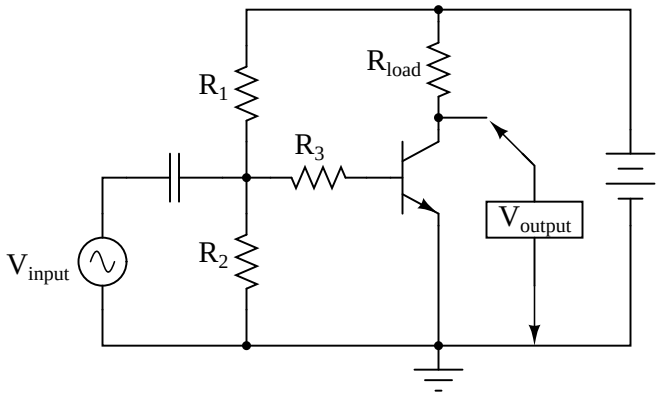
A familiar example of feedback happens in public-address ("PA") systems where someone holds the microphone too close to a speaker: a high-pitched "whine" or "howl" ensues, because the audio amplifier system is detecting and amplifying its own noise. Specifically, this is an example of *positive* or *regenerative* 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.

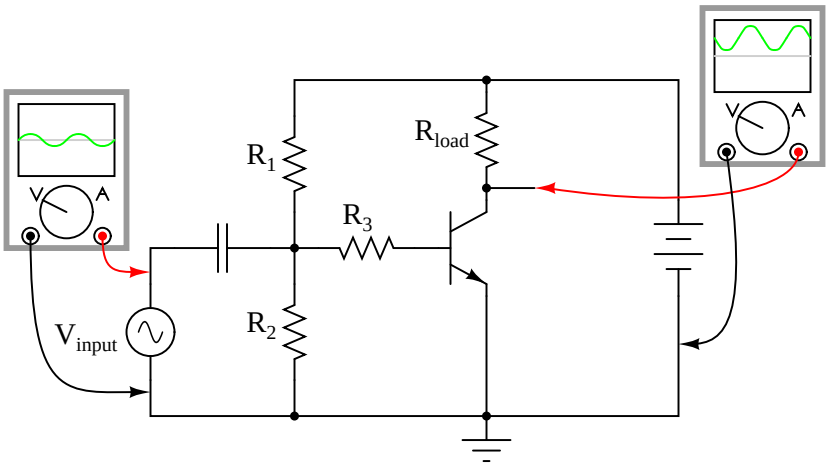
An amplifier circuit equipped with some amount of negative feedback is not only more stable, but it tends to distort the input waveform to a lesser degree and is generally capable of amplifying a wider range of frequencies. The tradeoff for these advantages (there just *has* to be a disadvantage to negative 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:

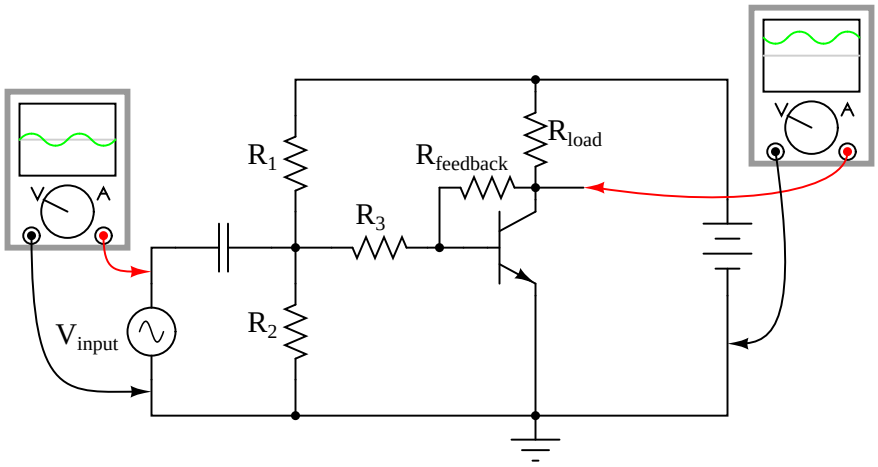


The amplifier configuration shown here is a common-emitter, with a resistor bias network formed by R_1 and R_2. The capacitor couples V_{input} to the amplifier so that the signal source doesn't have a DC voltage imposed on it by the R_1/R_2 divider network. Resistor R_3 serves the purpose of controlling 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:



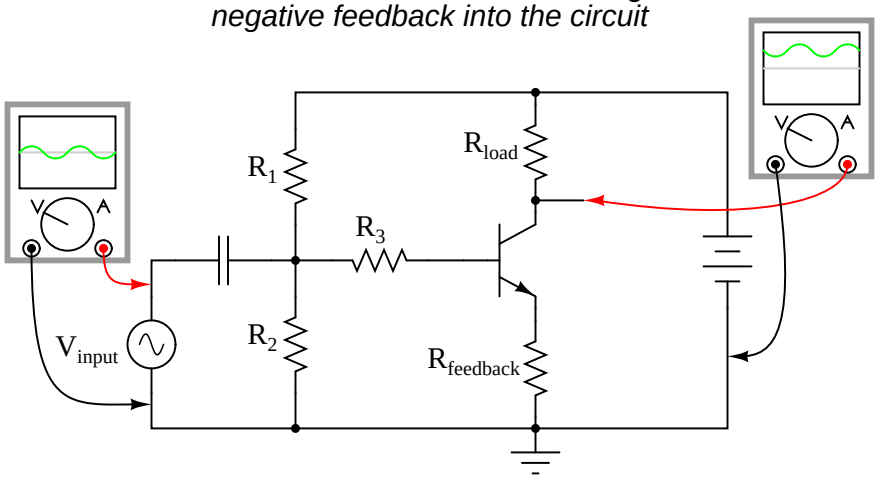
Because the output is an inverted, or mirror-image, reproduction of the input signal, any connection between the output (collector) wire and the input (base) wire of the transistor will result in *negative* feedback:



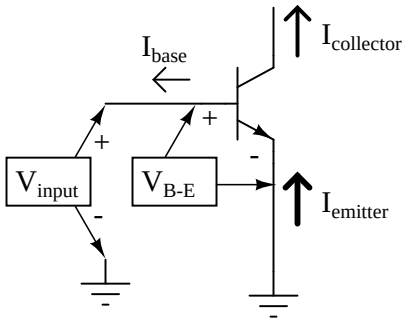
The resistances of R_1, R_2, R_3, and $R_{feedback}$ function together as a signal-mixing network so that the voltage seen at the base of the transistor (in reference to ground) is a weighted average of the input voltage and the feedback voltage, resulting in signal of reduced amplitude going into the 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:

A different method of introducing negative feedback into the circuit

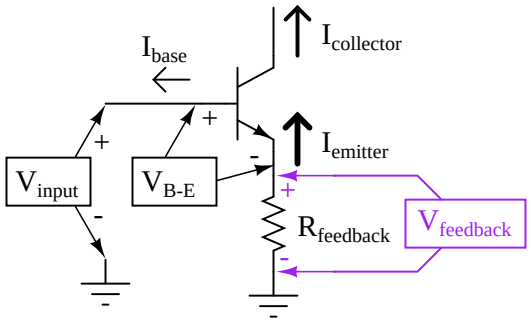


This new feedback resistor drops voltage proportional to the emitter current through the transistor, and it does so in such a way as to oppose the input signal's influence on the base-emitter junction of the transistor. Let's take a closer look at the emitter-base junction and see what difference this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

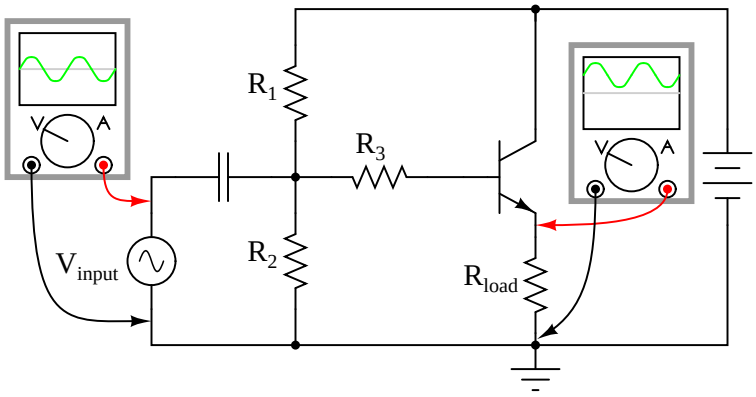
Consequently, the input voltage has less control over the transistor than before, and the voltage gain for the amplifier is reduced: just what we expected from negative feedback.

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.

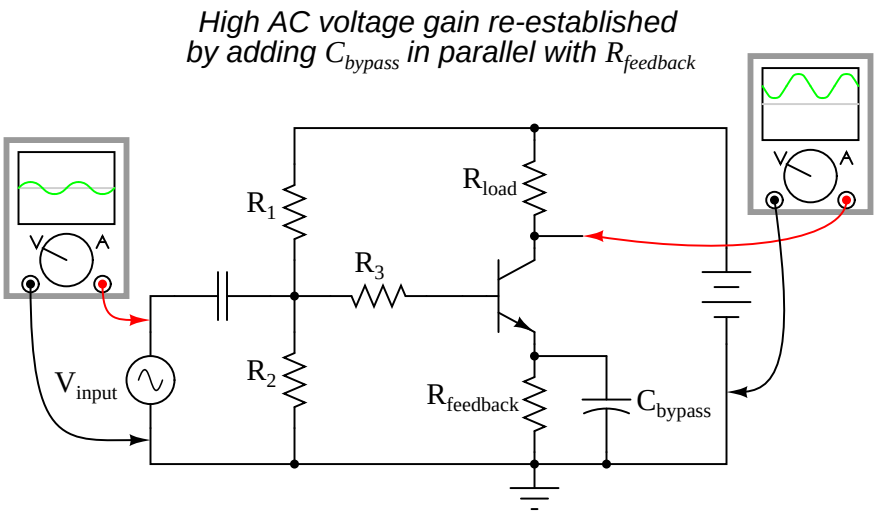
Another benefit of negative feedback, seen clearly in the common-collector circuit, is that it tends to make the voltage gain of the amplifier less dependent on the characteristics of the transistor.

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!

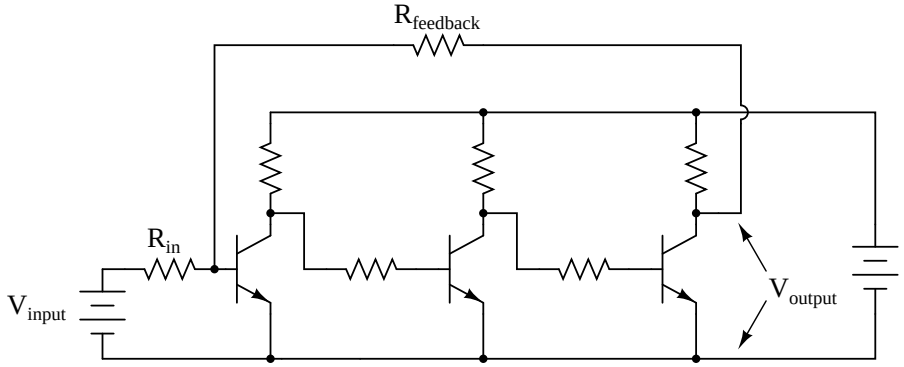
By connecting a capacitor in parallel with the feedback resistor, we create the very situation we need: a path from emitter to ground that is easier for AC than it is for DC:



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.

4.11 Amplifier impedances

***** PENDING *****

- **REVIEW:**
-
-
-

4.12 Current mirrors

An interesting and often-used circuit applying the bipolar junction transistor is the so-called *current mirror*, which serves as a simple current regulator, supplying nearly constant current to a load over a wide range of load resistances.

We know that in a transistor operating in its active mode, collector current is equal to base current multiplied by the ratio β. We also know that the ratio between collector current and emitter current is called α. Because collector current is equal to base current multiplied by β, and emitter current is the sum of the base and collector currents, α should be mathematically derivable from β. If you do the algebra, you'll find that $\alpha = \beta/(\beta+1)$ for any transistor.

We've seen already how maintaining a constant base current through an active transistor results in the regulation of collector current, according to the β ratio. Well, the α ratio works similarly: if emitter current is held constant, collector current will remain at a stable, regulated value so long as the transistor has enough collector-to-emitter voltage drop to maintain it in its active mode. Therefore, if we have a way of holding emitter current constant through a transistor, the transistor will work to regulate collector current at a constant value.

Remember that the base-emitter junction of a BJT is nothing more than a PN junction, just like a diode, and that the "diode equation" specifies how much current will go through a PN junction given forward voltage drop and junction temperature:

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

Where,

I_D = Diode current in amps

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

e = Euler's constant (~ 2.718281828)

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

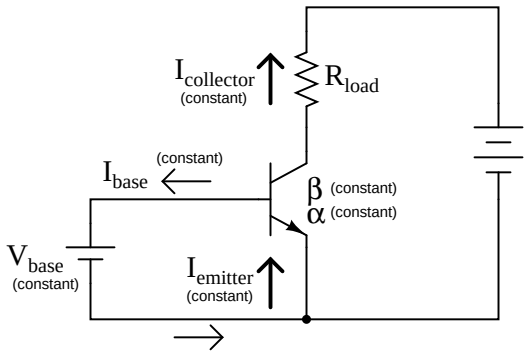
V_D = Voltage applied across diode in volts

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

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

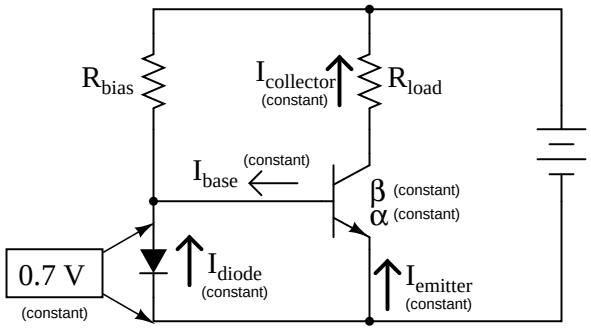
T = Junction temperature in degrees Kelvin

If both junction voltage and temperature are held constant, then the PN junction current will likewise be constant. Following this rationale, if we were to hold the base-emitter voltage of a transistor constant, then its emitter current should likewise be constant, given a constant temperature:



This constant emitter current, multiplied by a constant α ratio, gives a constant collector current through R_{load}, provided that there is enough battery voltage to keep the transistor in its active mode for any change in R_{load}'s resistance.

Maintaining a constant voltage across the transistor's base-emitter junction is easy: use a forward-biased diode to establish a constant voltage of approximately 0.7 volts, and connect it in parallel with the base-emitter junction:



Now, here's where it gets interesting. The voltage dropped across the diode probably won't be 0.7 volts exactly. The exact amount of forward voltage dropped across it depends on the current through the diode, and the diode's temperature, all in accordance with the diode equation. If diode current is increased (say, by reducing the resistance of R_{bias}), its voltage drop will increase slightly, increasing the voltage drop across the transistor's base-emitter junction, which will increase the emitter current by the same proportion, assuming the diode's PN junction and the transistor's base-emitter junction are well-matched to each other. In other words, transistor emitter current will closely equal diode current at any given time. If you change the diode current by changing the resistance value of R_{bias}, then the transistor's emitter current will follow suit, because the emitter current is described by the same equation as the diode's, and both PN junctions experience the same voltage drop.

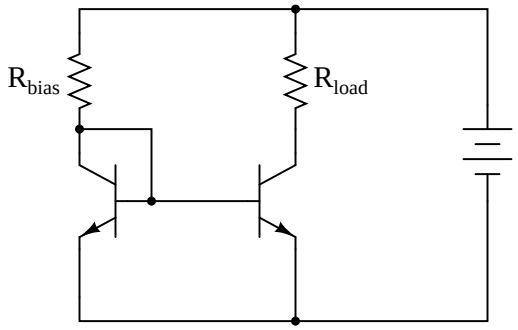
Remember, the transistor's collector current is almost equal to its emitter current, as the α ratio of a typical transistor is almost unity (1). If we have control over the transistor's emitter current by setting diode current with a simple resistor adjustment, then we likewise have control over the transistor's collector current. In other words, collector current mimics, or *mirrors*, diode current.

Current through resistor R_{load} is therefore a function of current set by the bias resistor, the two

being nearly equal. This is the function of the current mirror circuit: to regulate current through the load resistor by conveniently adjusting the value of R_{bias}. It is very easy to create a set amount of diode current, as current through the diode is described by a simple equation: power supply voltage minus diode voltage (almost a constant value), divided by the resistance of R_{bias}.

To better match the characteristics of the two PN junctions (the diode junction and the transistor base-emitter junction), a transistor may be used in place of a regular diode, like this:

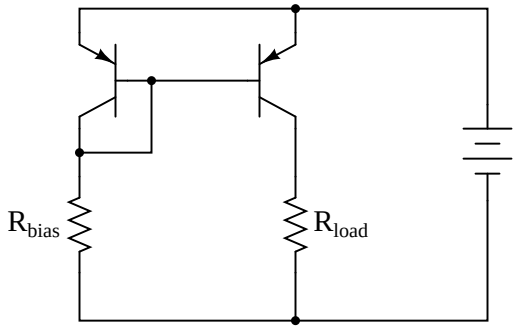
A current mirror circuit using two transistors



Because temperature is a factor in the "diode equation," and we want the two PN junctions to behave identically under all operating conditions, we should maintain the two transistors at exactly the same temperature. This is easily done using discrete components by gluing the two transistor cases back-to-back. If the transistors are manufactured together on a single chip of silicon (as a so-called *integrated circuit*, or *IC*), the designers should locate the two transistors very close to one another to facilitate heat transfer between them.

The current mirror circuit shown with two NPN transistors is sometimes called a *current-sinking* type, because the regulating transistor conducts current to the load *from ground* ("sinking" current), rather than *from the positive side of the battery* ("sourcing" current). If we wish to have a grounded load, and a *current sourcing* mirror circuit, we could use PNP transistors like this:

*A **current-sourcing** mirror circuit*



- **REVIEW:**

- A *current mirror* is a transistor circuit that regulates current through a load resistance, the regulation point being set by a simple resistor adjustment.
- Transistors in a current mirror circuit must be maintained at the same temperature for precise operation. When using discrete transistors, you may glue their cases together to help accomplish this.
- Current mirror circuits may be found in two basic varieties: the current *sinking* configuration, where the regulating transistor connects the load to ground; and the current *sourcing* configuration, where the regulating transistor connects the load to the positive terminal of the DC power supply.

4.13 Transistor ratings and packages

***** INCOMPLETE *****

Like all electrical and electronic components, transistors are limited in the amounts of voltage and current they can handle without sustaining damage. Since transistors are a bit more complex than some of the other components you're used to seeing at this point, they tend to have more kinds of ratings. What follows is an itemized description of some typical transistor ratings.

Power dissipation: When a transistor conducts current between collector and emitter, it also drops voltage between those two points. At any given time, the power dissipated by a transistor is equal to the product (multiplication) of collector current and collector-emitter voltage. Just like resistors, transistors are rated in terms of how many watts they can safely dissipate without sustaining damage. High temperature is the mortal enemy of all semiconductor devices, and bipolar transistors tend to be more susceptible to thermal damage than most. Power ratings are always given in reference to the temperature of ambient (surrounding) air. When transistors are to be used in hotter-than-normal environments, their power ratings must be *derated* to avoid a shortened service life.

Reverse voltages: As with diodes, bipolar transistors are rated for maximum allowable reverse-bias voltage across their PN junctions. This includes voltage ratings for the base-emitter junction, base-collector junction, and also from collector to emitter. The rating for maximum collector-emitter voltage can be thought of in terms of the maximum voltage it can withstand while in full-cutoff mode (no base current). This rating is of particular importance when using a bipolar transistor as a switch.

Collector current: A maximum value for collector current will be given by the manufacturer in amps. Understand that this maximum figure assumes a saturated state (minimum collector-emitter voltage drop). If the transistor is *not* saturated, and in fact is dropping substantial voltage between collector and emitter, the maximum power dissipation rating will probably be exceeded before the maximum collector current rating will. Just something to keep in mind when designing a transistor circuit!

Saturation voltages: Ideally, a saturated transistor acts as a closed switch contact between collector and emitter, dropping zero voltage at full collector current. In reality this is *never* true. Manufacturers will specify the maximum voltage drop of a transistor at saturation, both between the collector and emitter, and also between base and emitter (forward voltage drop of that PN junction). Collector-emitter voltage drop at saturation is generally expected to be 0.3 volts or less, but this figure is of course dependent on the specific type of transistor. Base-emitter forward voltage drop is very similar to that of an equivalent diode, which should come as no surprise.

Beta: The ratio of collector current to base current, β is the fundamental parameter characterizing the amplifying ability of a bipolar transistor. β is usually assumed to be a constant figure in circuit calculations, but unfortunately this is far from true in practice. As such, manufacturers provide a set of β (or "h_{fe}") figures for a given transistor over a wide range of operating conditions, usually in the form of maximum/minimum/typical ratings. It may surprise you to see just how widely β can be expected to vary within normal operating limits. One popular small-signal transistor, the 2N3903, is advertised as having a β ranging from 15 to 150 depending on the amount of collector current. Generally, β is highest for medium collector currents, decreasing for very low and very high collector currents.

Alpha: the ratio of collector current to emitter current, α may be derived from β, being equal to $\beta/(\beta+1)$.

Bipolar transistors come in a wide variety of physical packages. Package type is primarily dependent upon the power dissipation of the transistor, much like resistors: the greater the maximum power dissipation, the larger the device has to be to stay cool. There are several standardized package types for three-terminal semiconductor devices, any of which may be used to house a bipolar transistor. This is an important fact to consider: there are many other semiconductor devices other than bipolar transistors which have three connection points. It is *impossible* to positively identify a three-terminal semiconductor device without referencing the part number printed on it, and/or subjecting it to a set of electrical tests.

- **REVIEW:**
-
-
-

4.14 BJT quirks

***** PENDING *****

Nonlinearity Temperature drift Thermal runaway Junction capacitance Noise Mismatch (problem with paralleling transistors) β cutoff frequency Alpha cutoff frequency

- **REVIEW:**
-
-
-

Chapter 5

JUNCTION FIELD-EFFECT TRANSISTORS

Contents

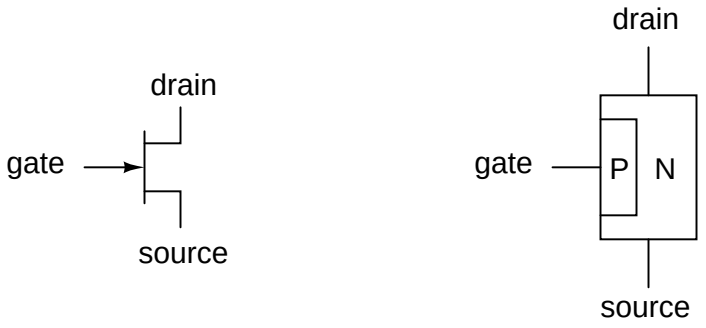
***** INCOMPLETE *****

5.1 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:

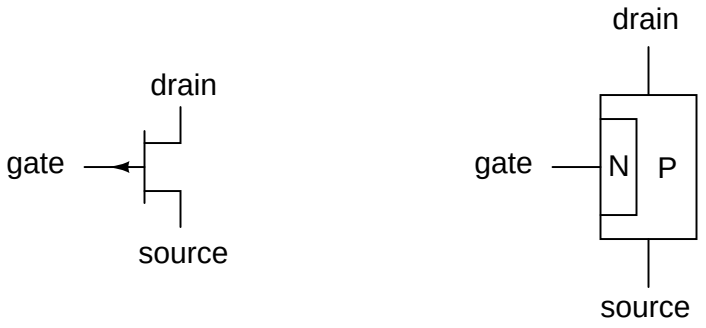
***N-channel* JFET**



schematic symbol *physical diagram*

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:

***P-channel* JFET**

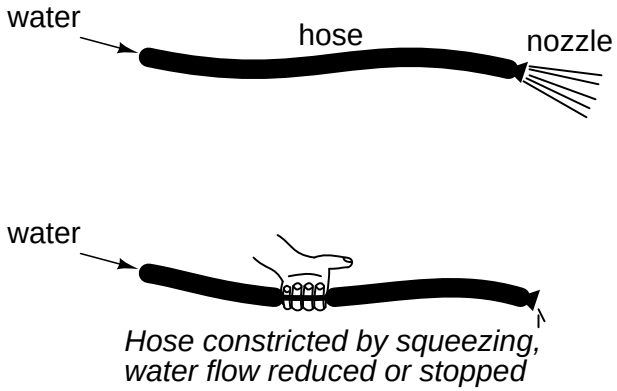


schematic symbol *physical diagram*

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.



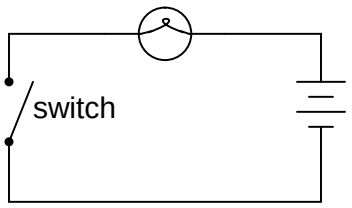
Note how this operational behavior is exactly opposite of the bipolar junction transistor. Bipolar transistors are *normally-off* devices: no current through the base, no current through the collector or the emitter. JFETs, on the other hand, are *normally-on* devices: no voltage applied to the gate 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.

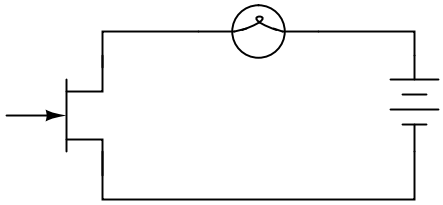
This is a very condensed overview of JFET operation. In the next section, we'll explore the use of the JFET as a switching device.

5.2 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:

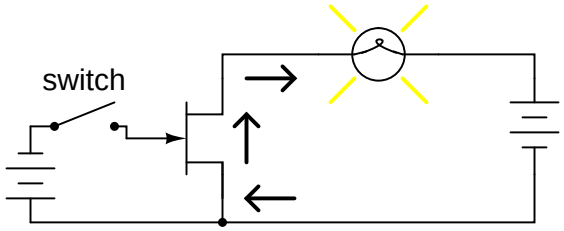


Remembering that the *controlled* current in a JFET flows between source and drain, we substitute the source and drain connections of a JFET for the two ends of the switch in the above circuit:

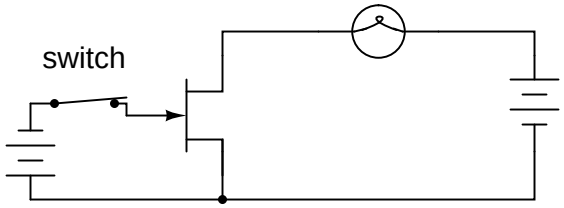


If you haven't noticed by now, the source and drain connections on a JFET look identical on the schematic symbol. Unlike the bipolar junction transistor where the emitter is clearly distinguished from the collector by the arrowhead, a JFET's source and drain lines both run perpendicular into the bar 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:

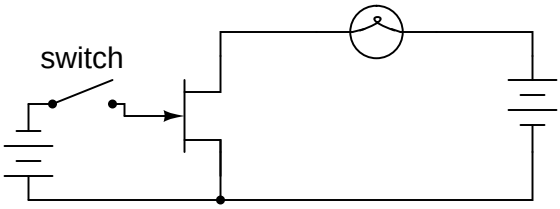


Closing this switch will "pinch off" the JFET's channel, thus forcing it into cutoff and turning the lamp off:



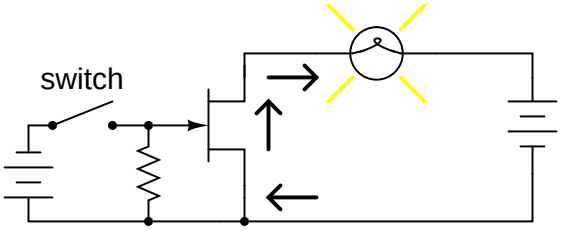
Note that there is no current going through the gate. As a reverse-biased PN junction, it firmly opposes the flow of any electrons through it. As a voltage-controlled device, the JFET requires negligible input current. This is an advantageous trait of the JFET over the bipolar transistor: there is virtually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



No lamp current after the switch opens!

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.

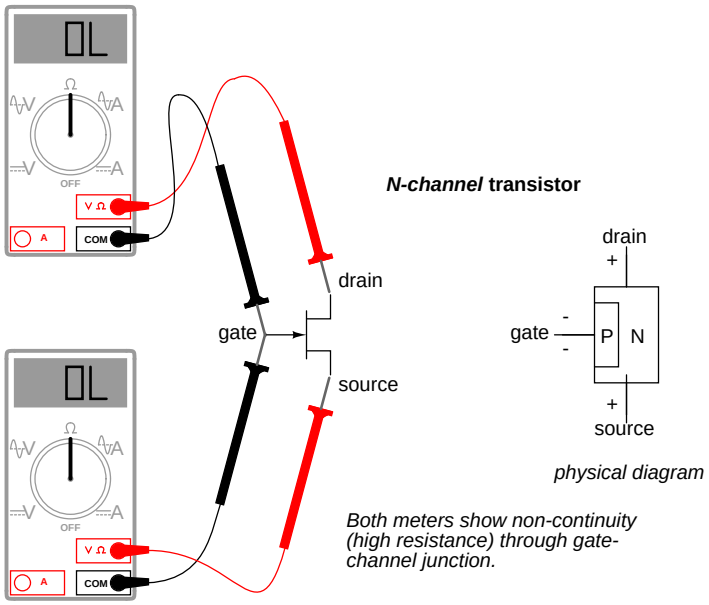
- **REVIEW:**
- Field-effect transistors control the current between source and drain connections by a voltage applied between the gate and source. In a *junction* field-effect transistor (JFET), there is a PN junction between the gate and source which is normally reverse-biased for control of source-drain current.
- JFETs are normally-on (normally-saturated) devices. The application of a reverse-biasing

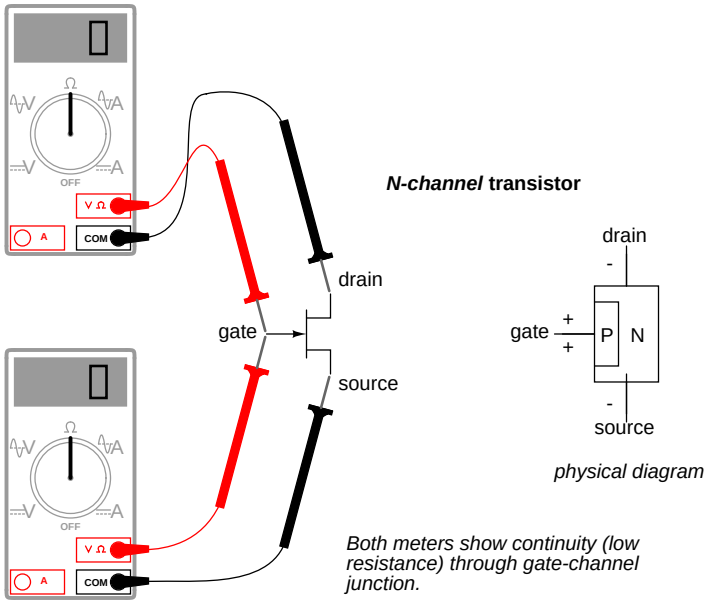
voltage between gate and source causes the depletion region of that junction to expand, thereby "pinching off" the channel between source and drain through which the controlled current travels.

- 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.

5.3 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.





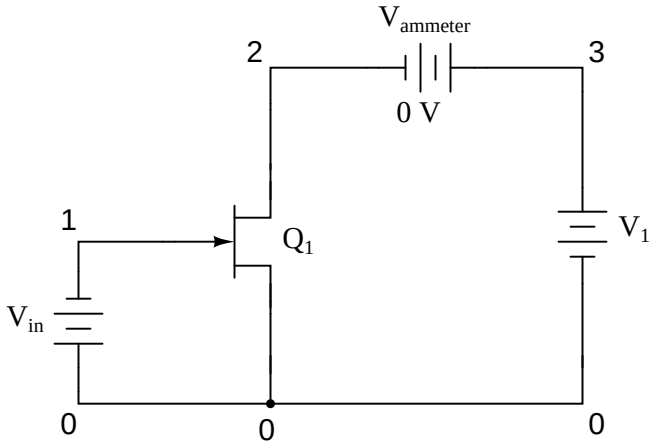
Testing continuity through the drain-source channel is another matter, though. Remember from the last section how a stored charge across the capacitance of the gate-channel PN junction could hold the JFET in a pinched-off state without any external voltage being applied across it? This can occur even 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.

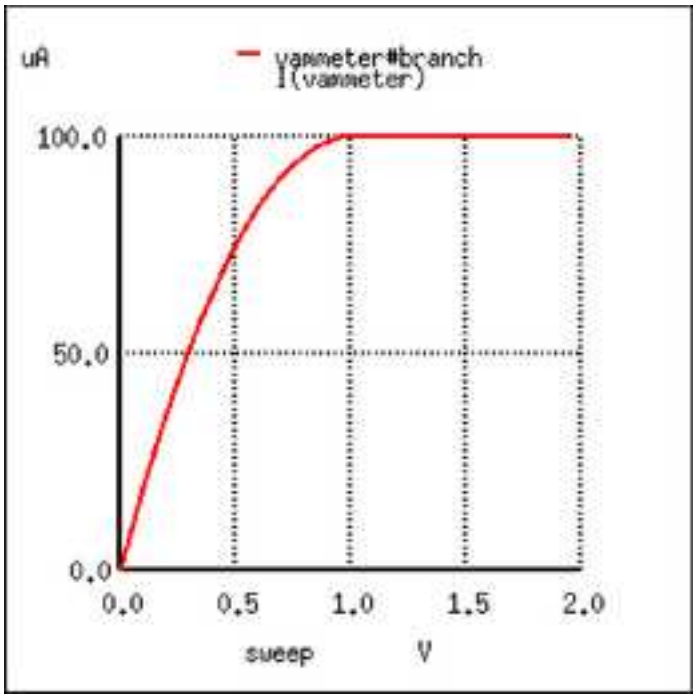
5.4 Active-mode operation

JFETs, like bipolar transistors, are able to "throttle" current in a mode between cutoff and saturation called the *active* mode. To better understand JFET operation, let's set up a SPICE simulation similar to the one used to explore basic bipolar transistor function:




```
jfet simulation
vin 0 1 dc 1
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 njf
.dc v1 0 2 0.05
.plot dc i(vammeter)
.end
```

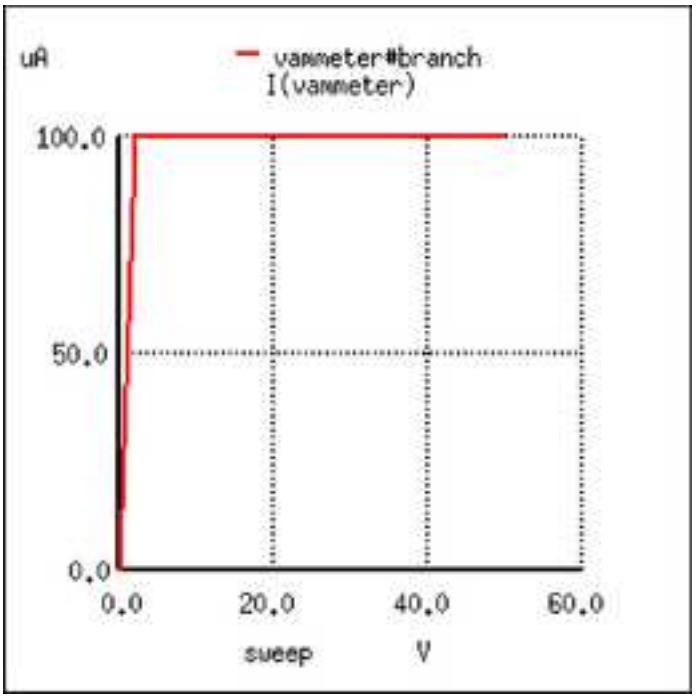
Note that the transistor labeled "Q_1" in the schematic is represented in the SPICE netlist as `j1`. Although all transistor types are commonly referred to as "Q" devices in circuit schematics – just as resistors are referred to by "R" designations, and capacitors by "C" – SPICE needs to be told what type of transistor this is by means of a different letter designation: `q` for bipolar junction transistors, and `j` for junction field-effect transistors.



Here, the controlling signal is a steady voltage of 1 volt, applied with negative towards the JFET gate and positive toward the JFET source, to reverse-bias the PN junction. In the first BJT simulation of chapter 4, a constant-current source of 20 μA was used for the controlling signal, but remember that a JFET is a *voltage-controlled* device, not a current-controlled device like the bipolar junction transistor.

Like the BJT, the JFET tends to regulate the controlled current at a fixed level above a certain power supply voltage, no matter how high that voltage may climb. Of course, this current regulation has limits in real life – no transistor can withstand infinite voltage from a power source – and with enough drain-to-source voltage the transistor will "break down" and drain current will surge. But within normal operating limits the JFET keeps the drain current at a steady level independent of power supply voltage. To verify this, we'll run another computer simulation, this time sweeping the power supply voltage (V_1) all the way to 50 volts:

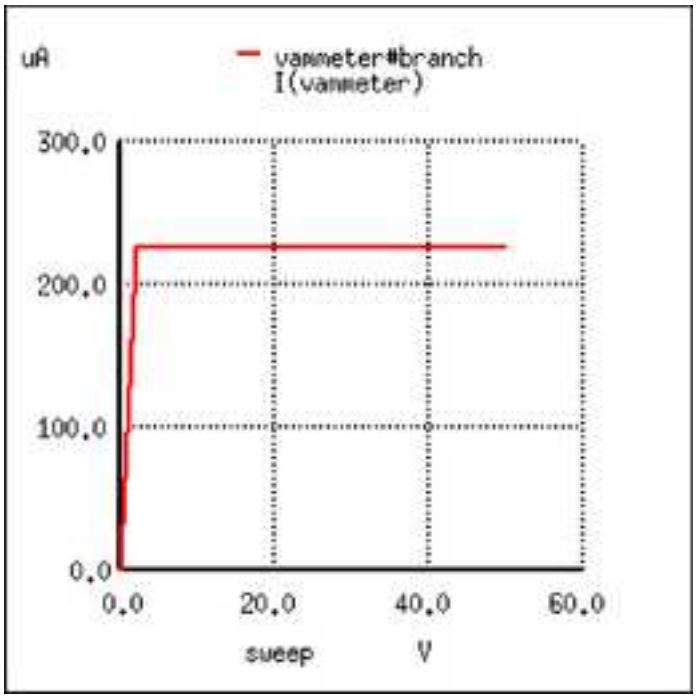
```
jfet simulation
vin 0 1 dc 1
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 njf
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



Sure enough, the drain current remains steady at a value of 100 μA (1.000E-04 amps) no matter how high the power supply voltage is adjusted.

Because the input voltage has control over the constriction of the JFET's channel, it makes sense that changing this voltage should be the only action capable of altering the current regulation point for the JFET, just like changing the base current on a BJT is the only action capable of altering collector current regulation. Let's decrease the input voltage from 1 volt to 0.5 volts and see what happens:

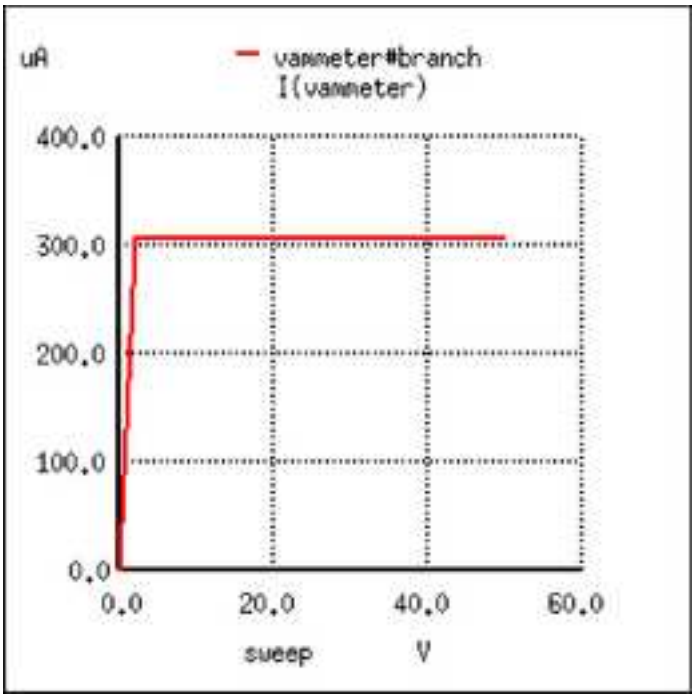
```
jfet simulation
vin 0 1 dc 0.5
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 njf
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



As expected, the drain current is greater now than it was in the previous simulation. With less reverse-bias voltage impressed across the gate-source junction, the depletion region is not as wide as it was before, thus "opening" the channel for charge carriers and increasing the drain current figure.

Please note, however, the actual value of this new current figure: 225 μA (2.250E-04 amps). The last simulation showed a drain current of 100 μA, and that was with a gate-source voltage of 1 volt. Now that we've reduced the controlling voltage by a factor of 2 (from 1 volt down to 0.5 volts), the drain current increased, but not by the same 2:1 proportion! Let's reduce our gate-source voltage once more by another factor of 2 (down to 0.25 volts) and see what happens:

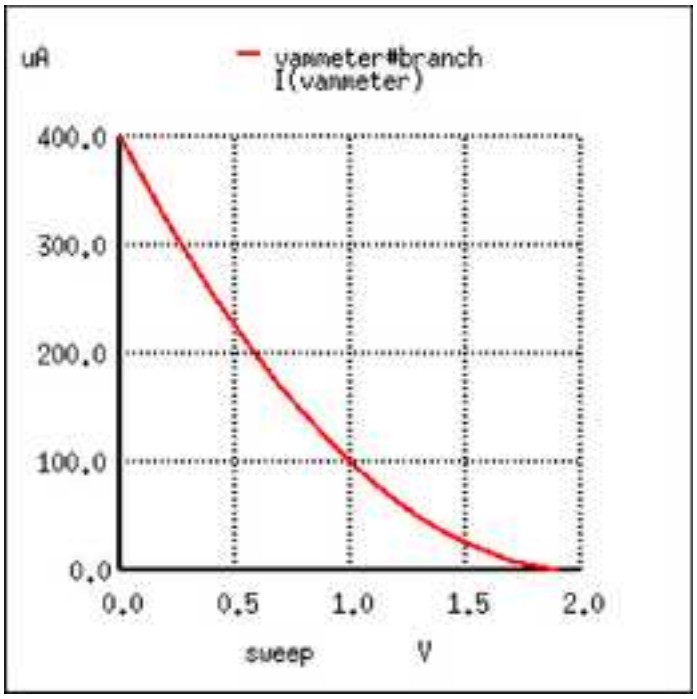
```
jfet simulation
vin 0 1 dc 0.25
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 njf
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



With the gate-source voltage set to 0.25 volts, one-half what it was before, the drain current is 306.3 μA. Although this is still an increase over the 225 μA from the prior simulation, it isn't *proportional* to the change of the controlling voltage.

To obtain a better understanding of what is going on here, we should run a different kind of simulation: one that keeps the power supply voltage constant and instead varies the controlling (voltage) signal. When this kind of simulation was run on a BJT, the result was a straight-line graph, showing how the input current / output current relationship of a BJT is linear. Let's see what kind of relationship a JFET exhibits:

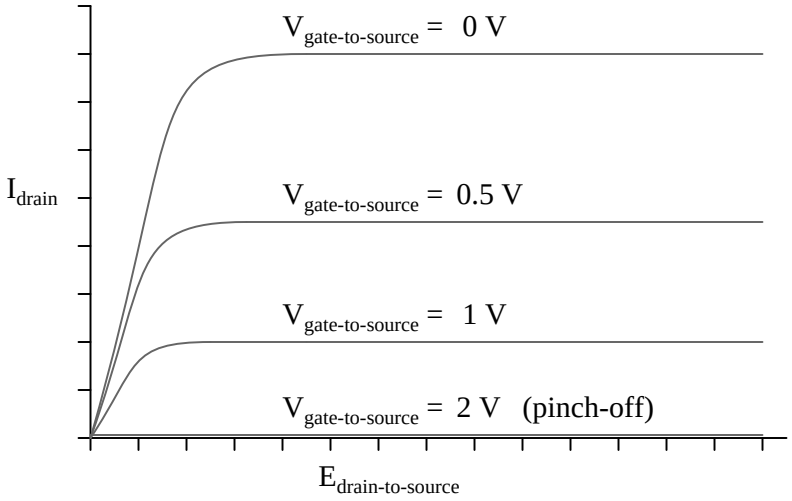
```
jfet simulation
vin 0 1 dc
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc 25
.model mod1 njf
.dc vin 0 2 0.1
.plot dc i(vammeter)
.end
```



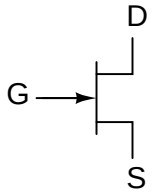
This simulation directly reveals an important characteristic of the junction field-effect transistor: the control effect of gate voltage over drain current is *nonlinear*. Notice how the drain current does not decrease linearly as the gate-source voltage is increased. With the bipolar junction transistor, collector current was directly proportional to base current: output signal proportionately followed input signal. Not so with the JFET! The controlling signal (gate-source voltage) has less and less effect over the drain current as it approaches cutoff. In this simulation, most of the controlling action (75 percent of drain current decrease – from 400 μA to 100 μA) takes place within the first volt of gate-source voltage (from 0 to 1 volt), while the remaining 25 percent of drain current reduction takes another whole volt worth of input signal. Cutoff occurs at 2 volts input.

Linearity is generally important for a transistor because it allows it to faithfully amplify a waveform without distorting it. If a transistor is nonlinear in its input/output amplification, the shape of the input waveform will become corrupted in some way, leading to the production of harmonics in the output signal. The only time linearity is *not* important in a transistor circuit is when it's being operated at the extreme limits of cutoff and saturation (on and off like a switch).

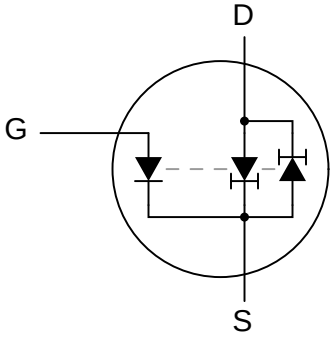
A JFET's characteristic curves display the same current-regulating behavior as for a BJT, and the nonlinearity between gate-to-source voltage and drain current is evident in the disproportionate vertical spacings between the curves:



To better comprehend the current-regulating behavior of the JFET, it might be helpful to draw a model made up of simpler, more common components, just as we did for the BJT:

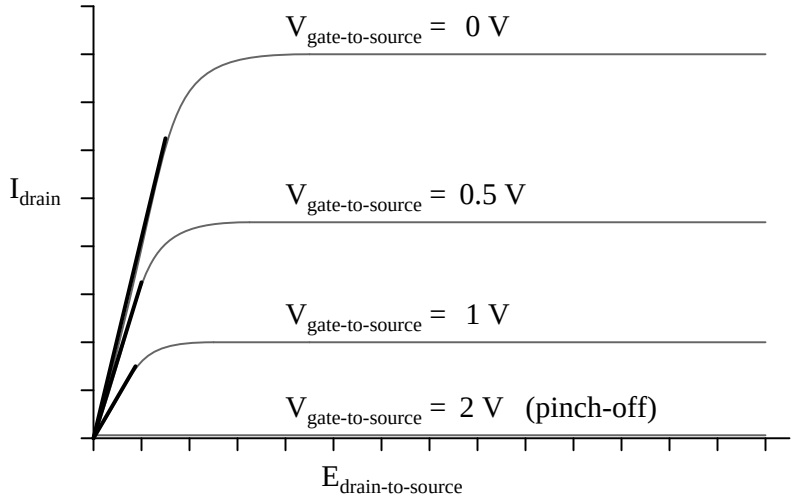


N-channel JFET diode-regulating diode model

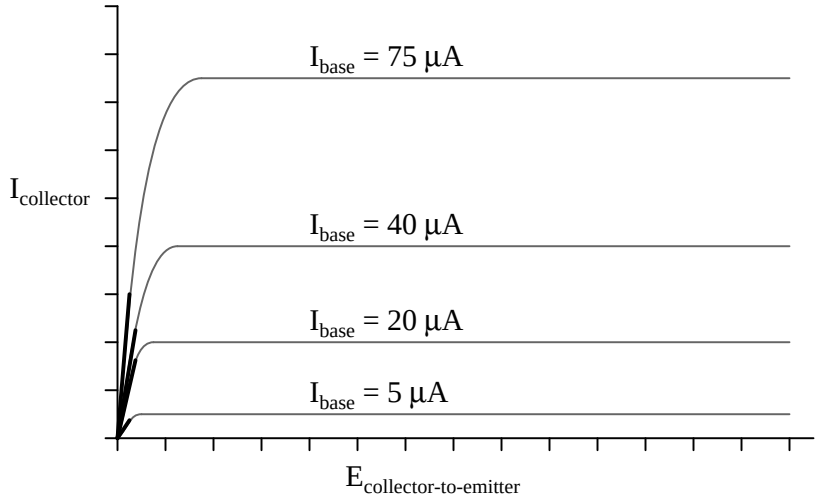


In the case of the JFET, it is the *voltage* across the reverse-biased gate-source diode which sets the current regulation point for the pair of constant-current diodes. A pair of opposing constant-current diodes is included in the model to facilitate current in either direction between source and drain, a trait made possible by the unipolar nature of the channel. With no PN junctions for the source-drain current to traverse, there is no polarity sensitivity in the controlled current. For this reason, JFETs are often referred to as *bilateral* devices.

A contrast of the JFET's characteristic curves against the curves for a bipolar transistor reveals a notable difference: the linear (straight) portion of each curve's saturation region (non-horizontal area) is surprisingly long compared to the respective portions of a BJT's characteristic curves:

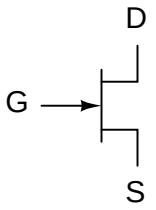


"Ohmic regions"

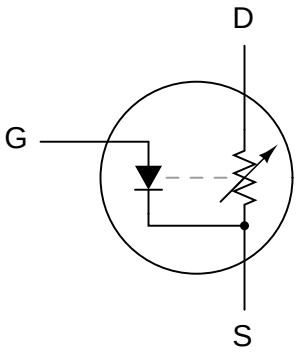


A JFET transistor in a condition of saturation tends to act very much like a plain resistor as measured from drain to source. Like all simple resistances, its current/voltage graph is a straight line. For this reason, the saturation (non-horizontal) portion of a JFET's characteristic curve is sometimes referred to as the *ohmic region*. In this mode of operation where there isn't enough drain-to-source voltage to bring drain current up to the regulated point, the drain current is directly proportional to the drain-to-source voltage. In a carefully designed circuit, this phenomenon can be used to an advantage. Operated in this region of the curve, the JFET acts like a voltage-controlled

resistance rather than a voltage-controlled *current regulator*, and the appropriate model for the transistor is different:



N-channel JFET diode-rheostat model
(for saturation, or "ohmic," mode only!)



Here and here alone the rheostat (variable resistor) model of a transistor is accurate. It must be remembered, however, that this model of the transistor holds true only for a narrow range of its operation: when it is extremely saturated (far less voltage applied between drain and source than what is needed to achieve full regulated current through the drain). The amount of resistance (measured in ohms) between drain and source in this mode is controlled by how much reverse-bias voltage is applied between gate and source. The less gate-to-source voltage, the less resistance (steeper line on graph).

Because JFETs are *voltage*-controlled current regulators (at least when they're allowed to operate in their active mode, not saturated), their inherent amplification factor cannot be expressed as a unitless ratio as with BJTs. In other words, there is no β ratio for a JFET. This is true for all voltage-controlled active devices, including other types of field-effect transistors and even electron tubes. There is, however, an expression of controlled (drain) current to controlling (gate-source) voltage, and it is called *transconductance*. Its unit is Siemens, the same unit for conductance (formerly known as the *mho*).

Why this choice of units? Because the equation takes on the general form of current (output signal) divided by voltage (input signal).

$$g_{fs} = \frac{\Delta I_D}{\Delta V_{GS}}$$

Where,

g_{fs} = Transconductance in Siemens
ΔI_D = Change in drain current
ΔV_{GS} = Change in gate-source voltage

Unfortunately, the transconductance value for any JFET is not a stable quantity: it varies significantly with the amount of gate-to-source control voltage applied to the transistor. As we saw in the SPICE simulations, the drain current does not change proportionally with changes in gate-source voltage. To calculate drain current for any given gate-source voltage, there is another equation that may be used. It is obviously nonlinear upon inspection (note the power of 2), reflecting the nonlinear behavior we've already experienced in simulation:

$$I_D = I_{DSS} \left(1 - \frac{V_{GS}}{V_{GS(cutoff)}} \right)^2$$

Where,

I_D = Drain current
I_{DSS} = Drain current with gate shorted to source
V_{GS} = Gate-to-source voltage
$V_{GS(cutoff)}$ = Pinch-off gate-to-source voltage

- **REVIEW:**
- In their active modes, JFETs regulate drain current according to the amount of reverse-bias voltage applied between gate and source, much like a BJT regulates collector current according to base current. The mathematical ratio between drain current (output) and gate-to-source voltage (input) is called *transconductance*, and it is measured in units of Siemens.
- The relationship between gate-source (control) voltage and drain (controlled) current is nonlinear: as gate-source voltage is decreased, drain current increases exponentially. That is to say, the transconductance of a JFET is not constant over its range of operation.
- In their saturation modes, JFETs regulate drain-to-source *resistance* according to the amount of reverse-bias voltage applied between gate and source. In other words, they act like voltage-controlled resistances.

5.5 The common-source amplifier – PENDING

***** PENDING *****

- **REVIEW:**

-
-
-

5.6 The common-drain amplifier – PENDING

***** PENDING *****

- **REVIEW:**
-
-
-

5.7 The common-gate amplifier – PENDING

***** PENDING *****

- **REVIEW:**
-
-
-

5.8 Biasing techniques – PENDING

***** PENDING *****

- **REVIEW:**
-
-
-

5.9 Transistor ratings and packages – PENDING

***** PENDING *****

- **REVIEW:**
-
-
-

5.10 JFET quirks – PENDING

***** PENDING *****

- **REVIEW:**
-
-
-

Chapter 6

INSULATED-GATE FIELD-EFFECT TRANSISTORS

Contents

***** INCOMPLETE *****

6.1 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*

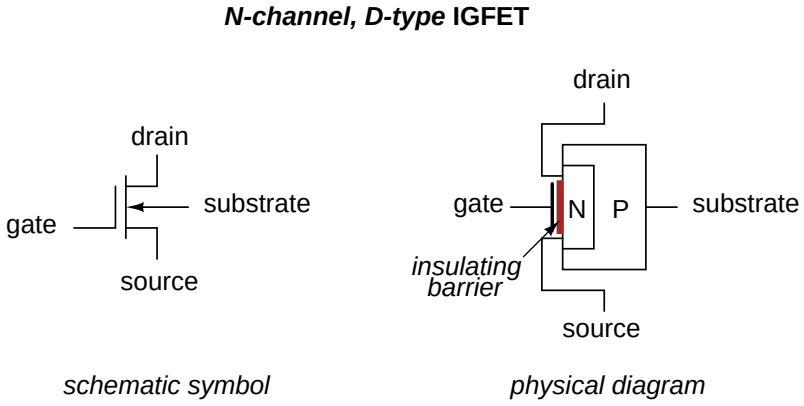
gate. This insulating barrier acts like the dielectric layer of a capacitor, and allows gate-to-source voltage to influence the depletion region electrostatically rather than by direct connection.

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.

6.2 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.

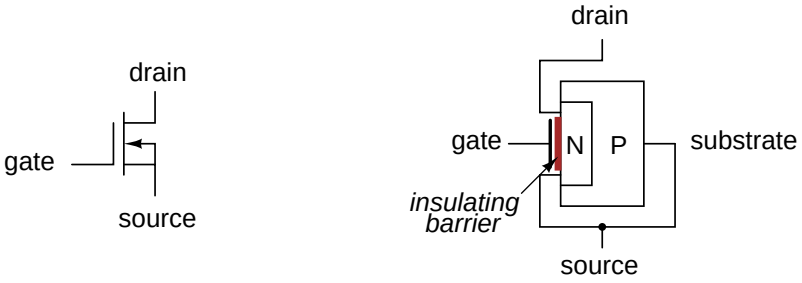
Here is a diagram of an N-channel IGFET of the "depletion" type:



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:

***N-channel, D-type* IGFET**

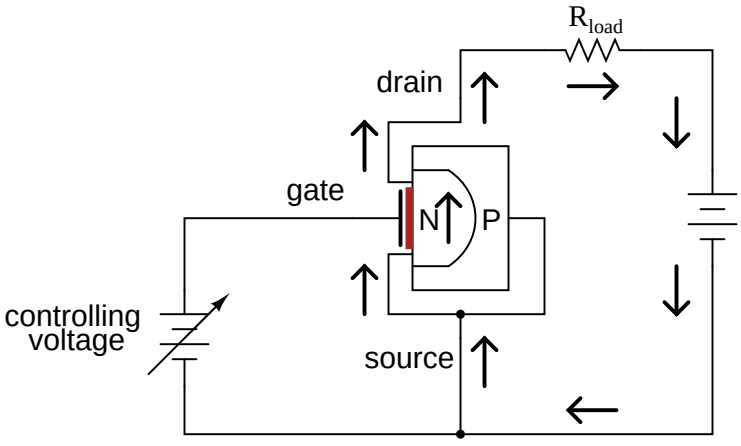


schematic symbol *physical diagram*

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.

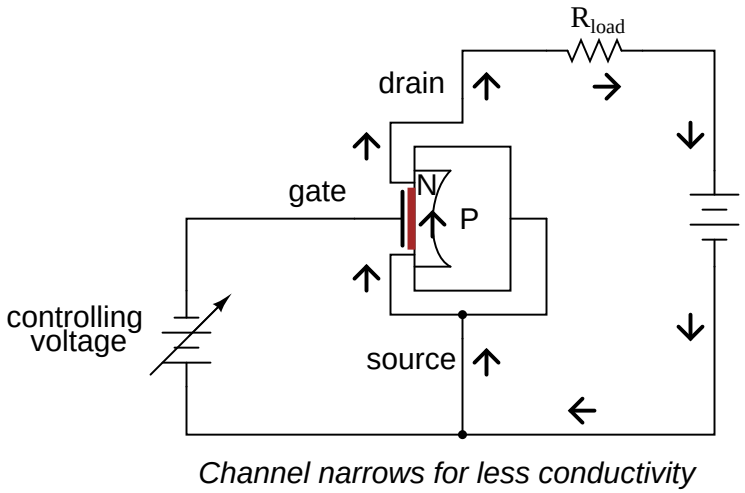
Indeed it is, for when a controlling voltage is applied between gate and source, the conductivity of the channel is changed as a result of the depletion region *moving* closer to or further away from the gate. In other words, the channel's effective width changes just as with the JFET, but this change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

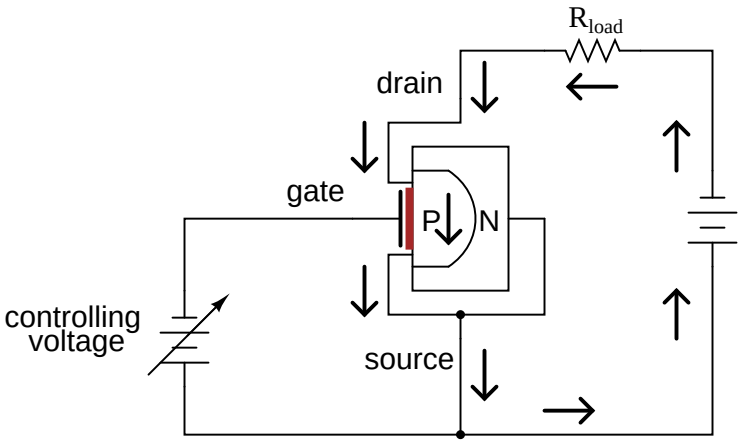
Reversing the controlling voltage's polarity has the opposite effect, attracting the depletion region and narrowing the channel, consequently reducing channel conductivity:



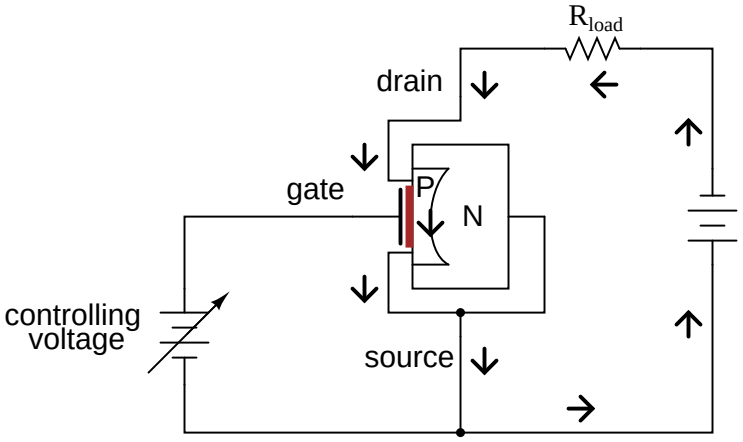
The insulated gate allows for controlling voltages of any polarity without danger of forward-biasing a junction, as was the concern with JFETs. This type of IGFET, although it's called a "depletion-type," actually has the capability of having its channel *either* depleted (channel narrowed) *or* enhanced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

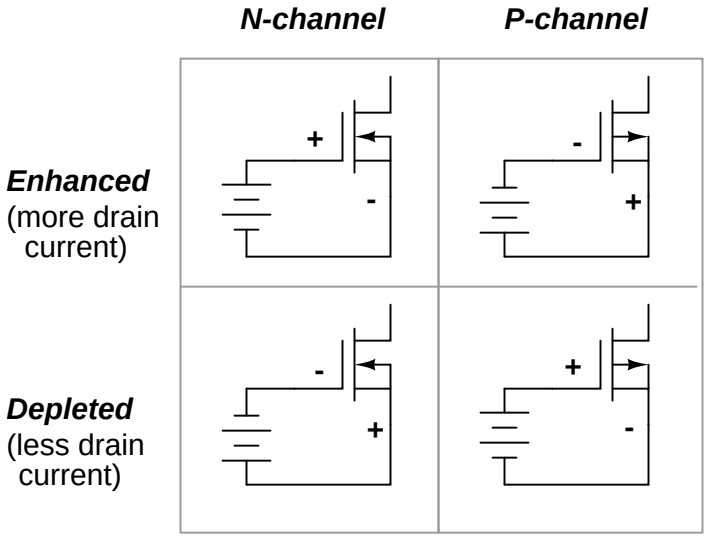


Channel expands for greater conductivity



Channel narrows for less conductivity

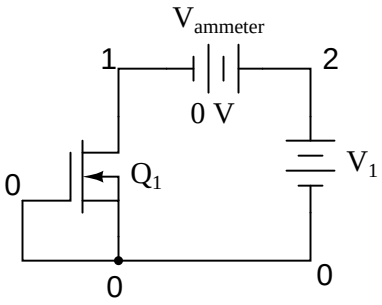
Illustrating the proper biasing polarities with standard IGFET symbols:



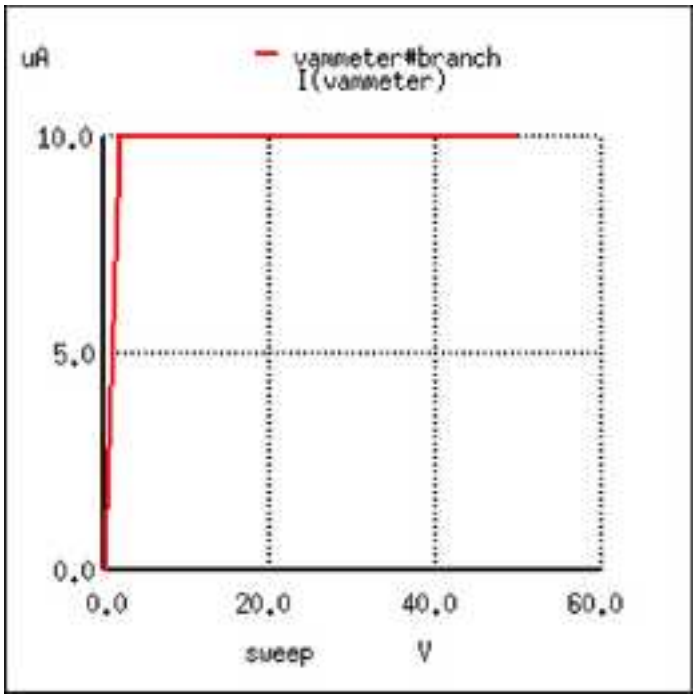
When there is zero voltage applied between gate and source, the IGFET will conduct current between source and drain, but not as much current as it would if it were enhanced by the proper gate voltage. This places the depletion-type, or simply *D-type*, IGFET in a category of its own in the transistor 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.

A few SPICE analyses will demonstrate the current-regulating behavior of D-type IGFETs. First, a test with zero input voltage (gate shorted to source) and the power supply swept from 0 to 50 volts. The graph shows drain current:

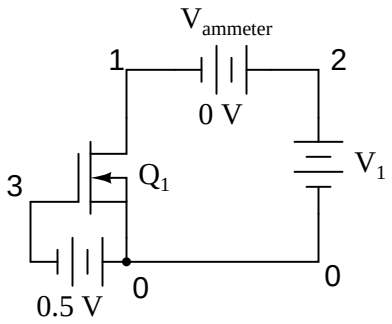



```
n-channel igfet characteristic curve
m1 1 0 0 0 mod1
vammeter 2 1 dc 0
v1 2 0
.model mod1 nmos vto=-1
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```

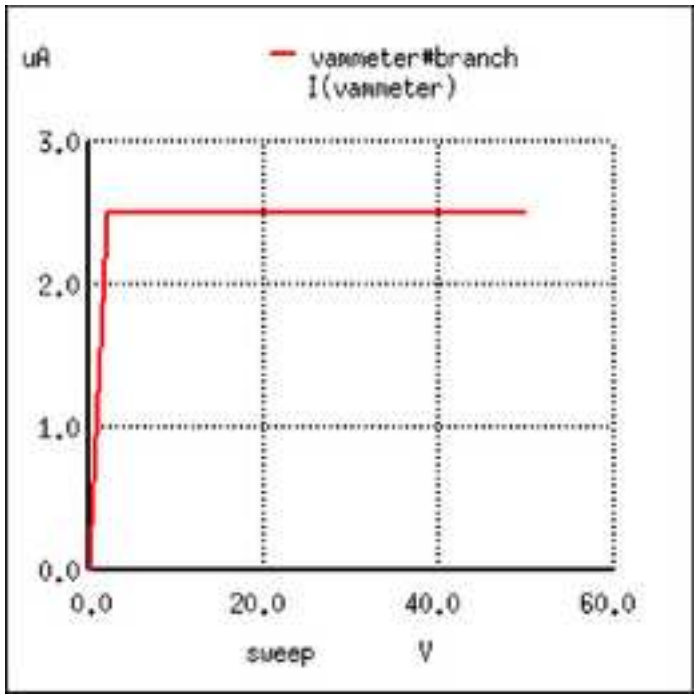


As expected for any transistor, the controlled current holds steady at a regulated value over a wide range of power supply voltages. In this case, that regulated point is 10 μA (1.000E-05). Now

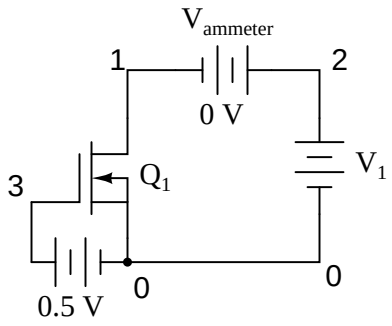
let's see what happens when we apply a negative voltage to the gate (with reference to the source) and sweep the power supply over the same range of 0 to 50 volts:



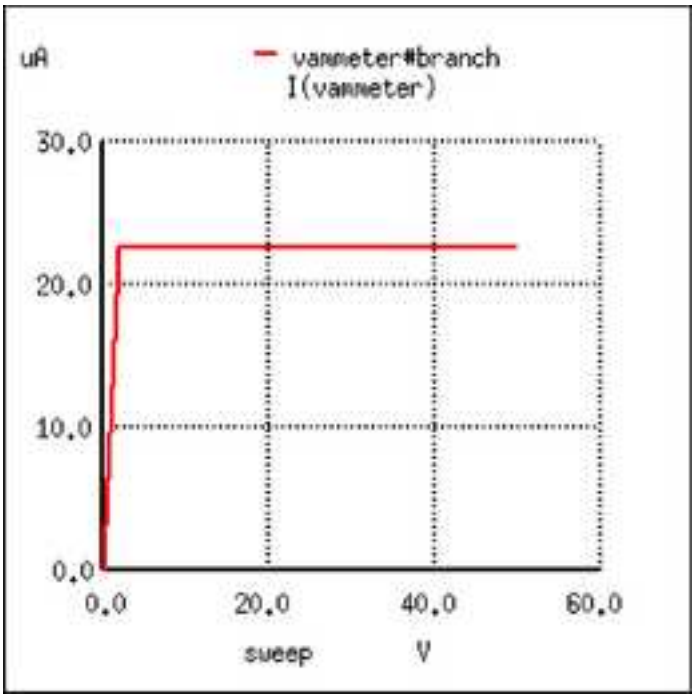

```
n-channel igfet characteristic curve
m1 1 3 0 0 mod1
vin 0 3 dc 0.5
vammeter 2 1 dc 0
v1 2 0
.model mod1 nmos vto=-1
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



Not surprisingly, the drain current is now regulated at a lower value of 2.5 μA (down from 10 μA with zero input voltage). Now let's apply an input voltage of the other polarity, to *enhance* the IGFET:

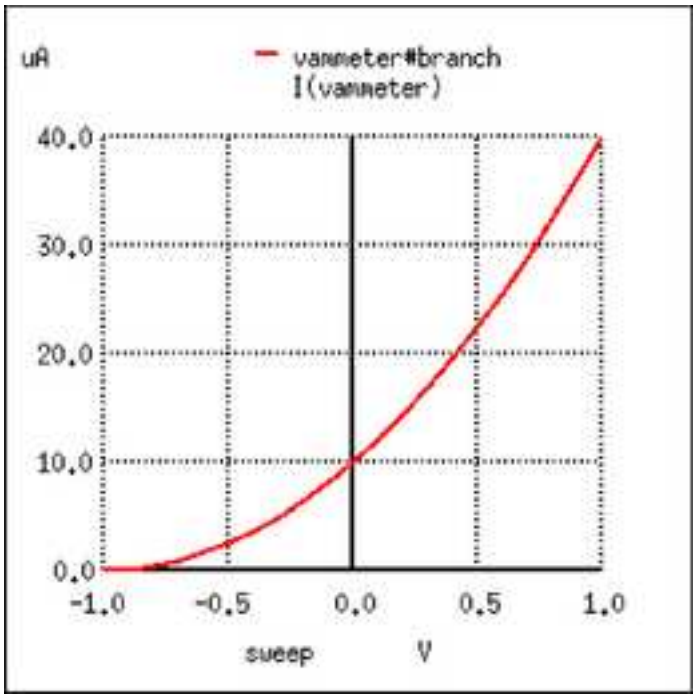



```
n-channel igfet characteristic curve
m1 1 3 0 0 mod1
vin 3 0 dc 0.5
vammeter 2 1 dc 0
v1 2 0
.model mod1 nmos vto=-1
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



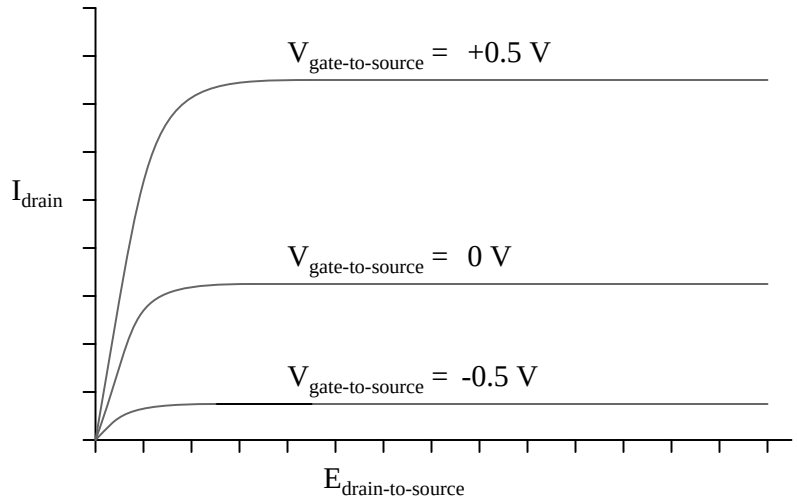
With the transistor enhanced by the small controlling voltage, the drain current is now at an increased value of 22.5 μA (2.250E-05). It should be apparent from these three sets of voltage and current figures that the relationship of drain current to gate-source voltage is nonlinear just as it was with the JFET. With 1/2 volt of depleting voltage, the drain current is 2.5 μA; with 0 volts input the drain current goes up to 10 μA; and with 1/2 volt of enhancing voltage, the current is at 22.5 μA. To obtain a better understanding of this nonlinearity, we can use SPICE to plot the drain current over a range of input voltage values, sweeping from a negative (depleting) figure to a positive (enhancing) figure, maintaining the power supply voltage of V_1 at a constant value:

```
n-channel igfet
m1 1 3 0 0 mod1
vin 3 0
vammeter 2 1 dc 0
v1 2 0 dc 24
.model mod1 nmos vto=-1
.dc vin -1 1 0.1
.plot dc i(vammeter)
.end
```



Just as it was with JFETs, this inherent nonlinearity of the IGFET has the potential to cause distortion in an amplifier circuit, as the input signal will not be reproduced with 100 percent accuracy at the output. Also notice that a gate-source voltage of about 1 volt in the depleting direction is able to pinch off the channel so that there is virtually no drain current. D-type IGFETs, like JFETs, have a certain pinch-off voltage rating. This rating varies with the precise unique of the transistor, and may not be the same as in our simulation here.

Plotting a set of characteristic curves for the IGFET, we see a pattern not unlike that of the JFET:



- **REVIEW:**
-
-
-

6.3 Enhancement-type IGFETs – PENDING

- **REVIEW:**
-
-
-

6.4 Active-mode operation – PENDING

- **REVIEW:**
-
-
-

6.5 The common-source amplifier – PENDING

- **REVIEW:**
-
-
-

6.6 The common-drain amplifier – PENDING

- **REVIEW:**
-
-
-

6.7 The common-gate amplifier – PENDING

- **REVIEW:**
-
-
-

6.8 Biasing techniques – PENDING

- **REVIEW:**
-
-
-

6.9 Transistor ratings and packages – PENDING

- **REVIEW:**
-
-
-

6.10 IGFET quirks – PENDING

- **REVIEW:**
-
-
-

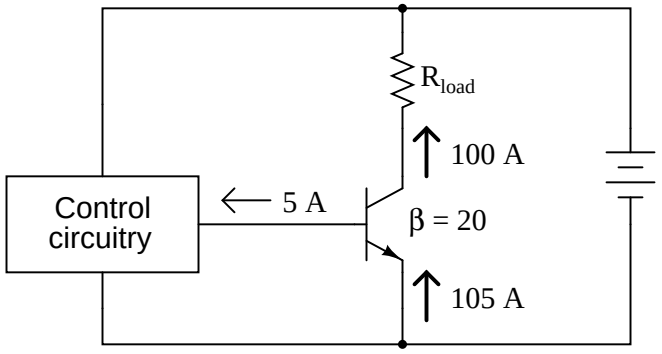
6.11 MESFETs – PENDING

- **REVIEW:**
-
-
-

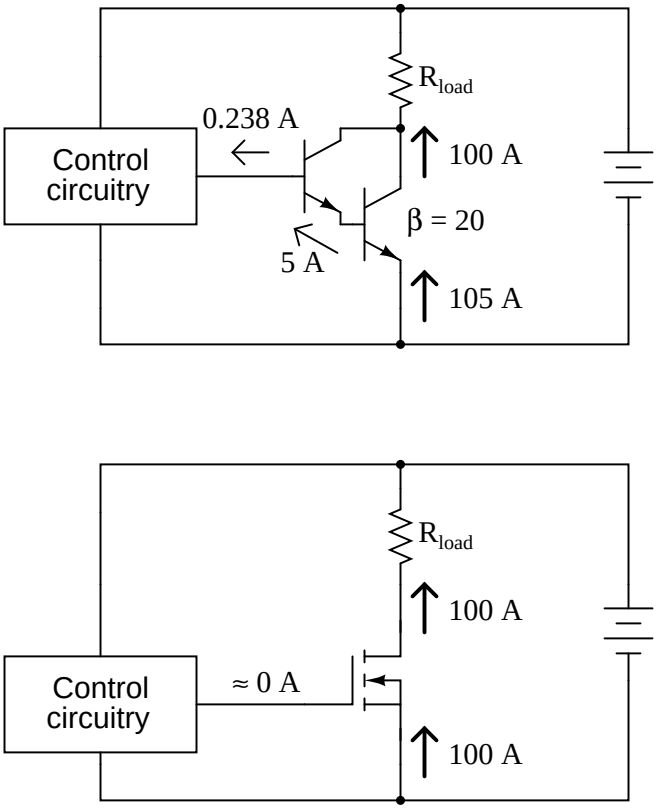
6.12 IGBTs

Because of their insulated gates, IGFETs of all types have extremely high current gain: there can be no sustained gate current if there is no continuous gate *circuit* in which electrons may continually flow. The only current we see through the gate terminal of an IGFET, then, is whatever transient (brief surge) may be required to charge the gate-channel capacitance and displace the depletion region as the transistor switches from an "on" state to an "off" state, or vice versa.

This high current gain would at first seem to place IGFET technology at a decided advantage over bipolar transistors for the control of very large currents. If a bipolar junction transistor is used to control a large collector current, there must be a substantial base current sourced or sunk by some control circuitry, in accordance with the β ratio. To give an example, in order for a power BJT with a β of 20 to conduct a collector current of 100 amps, there must be at least 5 amps of base current, a substantial amount of current in itself for miniature discrete or integrated control circuitry to handle:



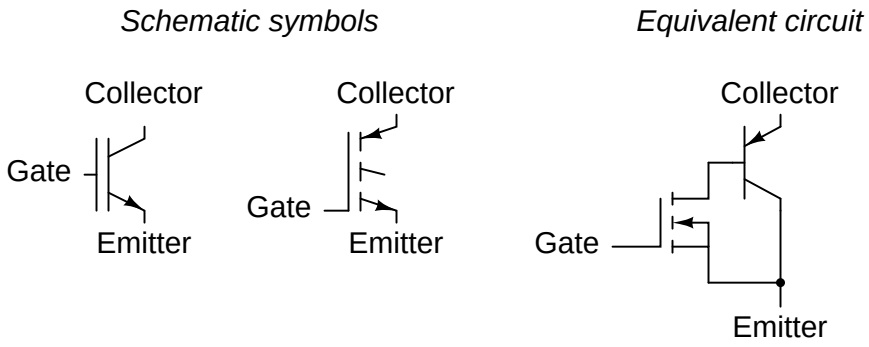
It would be nice from the standpoint of control circuitry to have power transistors with high current gain, so that far less current is needed for control of load current. Of course, we can use Darlington pair transistors to increase the current gain, but this kind of arrangement still requires *far* more controlling current than an equivalent power IGFET:



Unfortunately, though, IGFETs have problems of their own controlling high current: they typically exhibit greater drain-to-source voltage drop while saturated than the collector-to-emitter voltage drop of a saturated BJT. This greater voltage drop equates to higher power dissipation for the same amount of load current, limiting the usefulness of IGFETs as high-power devices. Although some specialized designs such as the so-called VMOS transistor have been designed to minimize this inherent disadvantage, the bipolar junction transistor is still superior in its ability to switch high currents.

An interesting solution to this dilemma leverages the best features of IGFETs with the best of features of BJTs, in one device called an *Insulated-Gate Bipolar Transistor*, or *IGBT*. Also known as an *Bipolar-mode MOSFET*, a *Conductivity-Modulated Field-Effect Transistor* (*COMFET*), or simply as an *Insulated-Gate Transistor* (*IGT*), it is equivalent to a Darlington pair of IGFET and BJT:

Insulated-Gate Bipolar Transistor (IGBT)
(N-channel)



In essence, the IGFET controls the base current of a BJT, which handles the main load current between collector and emitter. This way, there is extremely high current gain (since the insulated gate of the IGFET draws practically no current from the control circuitry), but the collector-to-emitter voltage drop during full conduction is as low as that of an ordinary BJT.

One disadvantage of the IGBT over a standard BJT is its slower turn-off time. For *fast* switching and high current-handling capacity, it's difficult to beat the bipolar junction transistor. Faster turn-off times for the IGBT may be achieved by certain changes in design, but only at the expense of a higher saturated voltage drop between collector and emitter. However, the IGBT provides a good alternative to IGFETs and BJTs for high-power control applications.

- **REVIEW:**
-
-
-

Chapter 7

THYRISTORS

Contents

***** INCOMPLETE *****

7.1 Hysteresis

Thyristors are a class of semiconductor components exhibiting *hysteresis*, that property whereby a system fails to return to its original state after some cause of state change has been removed. A very simple example of hysteresis is the mechanical action of a toggle switch: when the lever is pushed, it flips to one of two extreme states (positions) and will remain there even after the source of motion is removed (after you remove your hand from the switch lever). To illustrate the absence of hysteresis, consider the action of a "momentary" pushbutton switch, which returns to its original state after the button is no longer pressed: when the stimulus is removed (your hand), the system (switch) immediately and fully returns to its prior state with no "latching" behavior.

Bipolar, junction field-effect, and insulated gate field-effect transistors are all non-hysteretic devices. That is, they do not inherently "latch" into a state after being stimulated by a voltage or current signal. For any given input signal at any given time, a transistor will exhibit a predictable output response as defined by its characteristic curve. Thyristors, on the other hand, are semiconductor devices that tend to stay "on" once turned on, and tend to stay "off" once turned off. A

momentary event is able to flip these devices into either their on or off states where they will remain that way on their own, even after the cause of the state change is taken away. As such, they are useful only as on/off switching devices – much like a toggle switch – and cannot be used as analog signal amplifiers.

Thyristors are constructed using the same technology as bipolar junction transistors, and in fact may be analyzed as circuits comprised of transistor pairs. How then, can a hysteretic device (a thyristor) be made from non-hysteretic devices (transistors)? The answer to this question is *positive feedback*, also known as *regenerative feedback*. As you should recall, feedback is the condition where a percentage of the output signal is "fed back" to the input of an amplifying device. Negative, or degenerative, feedback results in a diminishing of voltage gain with increases in stability, linearity, and bandwidth. Positive feedback, on the other hand, results in a kind of instability where the amplifier's output tends to "saturate." In the case of thyristors, this saturating tendency equates to the device "wanting" to stay on once turned on, and off once turned off.

In this chapter we will explore several different kinds of thyristors, most of which stem from a single, basic two-transistor core circuit. Before we do that, though, it would be beneficial to study the technological predecessor to thyristors: gas discharge tubes.

7.2 Gas discharge tubes

If you've ever witnessed a lightning storm, you've seen electrical hysteresis in action (and probably didn't realize what you were seeing). The action of strong wind and rain accumulates tremendous static electric charges between cloud and earth, and between clouds as well. Electric charge imbalances manifest themselves as high voltages, and when the electrical resistance of air can no longer hold these high voltages at bay, huge surges of current travel between opposing poles of electrical charge which we call "lightning."

The buildup of high voltages by wind and rain is a fairly continuous process, the rate of charge accumulation increasing under the proper atmospheric conditions. However, lightning bolts are anything but continuous: they exist as relatively brief surges rather than continuous discharges. Why is this? Why don't we see soft, glowing lightning *arcs* instead of violently brief lightning *bolts*? The answer lies in the nonlinear (and hysteretic) resistance of air.

Under ordinary conditions, air has an extremely high amount of resistance. It is so high, in fact, that we typically treat its resistance as infinite and electrical conduction through the air as negligible. The presence of water and/or dust in air lowers its resistance some, but it is still an insulator for most practical purposes. When a sufficient amount of high voltage is applied across a distance of air, though, its electrical properties change: electrons become "stripped" from their normal positions around their respective atoms and are liberated to constitute a current. In this state, air is considered to be *ionized* and is referred to as a *plasma* rather than a normal *gas*. This usage of the word "plasma" is not to be confused with the medical term (meaning the fluid portion of blood), but is a fourth state of matter, the other three being solid, liquid, and vapor (gas). Plasma is a relatively good conductor of electricity, its specific resistance being much lower than that of the same substance in its gaseous state.

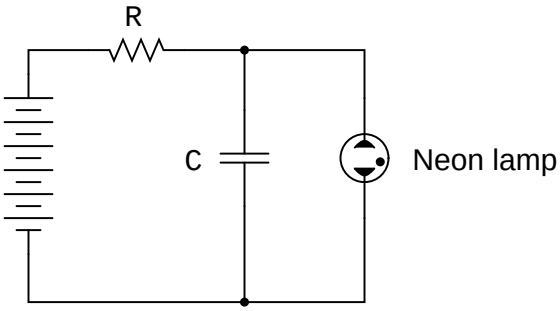
As an electric current moves through the plasma, there is energy dissipated in the plasma in the form of heat, just as current through a solid resistor dissipates energy in the form of heat. In the case of lightning, the temperatures involved are extremely high. High temperatures are also sufficient to convert gaseous air into a plasma or maintain plasma in that state without the presence

of high voltage. As the voltage between cloud and earth, or between cloud and cloud, decreases as the charge imbalance is neutralized by the current of the lightning bolt, the heat dissipated by the bolt maintains the air path in a plasma state, keeping its resistance low. The lightning bolt remains a plasma until the voltage decreases to too low a level to sustain enough current to dissipate enough heat. Finally, the air returns to a normal, gaseous state and stops conducting current, thus allowing voltage to build up once more.

Note how throughout this cycle, the air exhibits hysteresis. When not conducting electricity, it tends to *remain an insulator* until voltage builds up past a critical threshold point. Then, once it changes state and becomes a plasma, it tends to *remain a conductor* until voltage falls below a lower critical threshold point. Once "turned on" it tends to stay "on," and once "turned off" it tends to stay "off." This hysteresis, combined with a steady buildup of voltage due to the electrostatic effects of wind and rain, explains the action of lightning as brief bursts.

In electronic terms, what we have here in the action of lightning is a simple *relaxation oscillator*. Oscillators are electronic circuits that produce an oscillating (AC) voltage from a steady supply of DC power. A relaxation oscillator is one that works on the principle of a charging capacitor that is suddenly discharged every time its voltage reaches a critical threshold value. One of the simplest relaxation oscillators in existence is comprised of three components (not counting the DC power supply): a resistor, capacitor, and neon lamp:

Simple relaxation oscillator



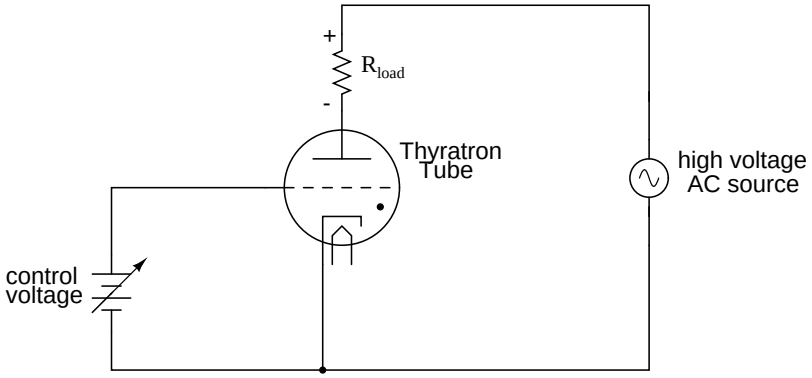
Neon lamps are nothing more than two metal electrodes inside a sealed glass bulb, separated by the neon gas inside. At room temperatures and with no applied voltage, the lamp has nearly infinite resistance. However, once a certain threshold voltage is exceeded (this voltage depends on the gas pressure and geometry of the lamp), the neon gas will become ionized (turned into a plasma) and its resistance dramatically reduced. In effect, the neon lamp exhibits the same characteristics as air in a lightning storm, complete with the emission of light as a result of the discharge, albeit on a much smaller scale.

The capacitor in the relaxation oscillator circuit shown above charges at an inverse exponential rate determined by the size of the resistor. When its voltage reaches the threshold voltage of the lamp, the lamp suddenly "turns on" and quickly discharges the capacitor to a low voltage value. Once discharged, the lamp "turns off" and allows the capacitor to build up a charge once more. The result is a series of brief flashes of light from the lamp, the rate of which dictated by battery voltage, resistor resistance, capacitor capacitance, and lamp threshold voltage.

While gas-discharge lamps are more commonly used as sources of illumination, their hysteretic

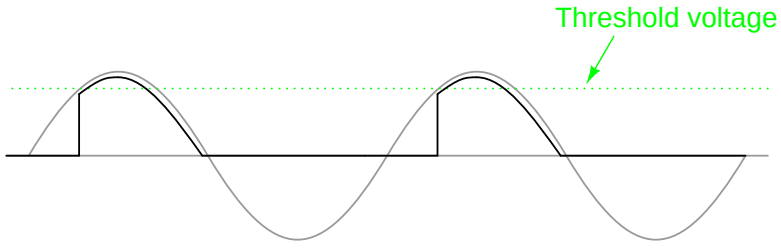
properties were leveraged in slightly more sophisticated variants known as *thyratron tubes*. Essentially a gas-filled triode tube (a triode being a three-element vacuum electron tube performing much a similar function to the N-channel, D-type IGFET), the thyratron tube could be turned on with a small control voltage applied between grid and cathode, and turned off by reducing the plate-to-cathode voltage.

(Simple) Thyratron control circuit



In essence, thyratron tubes were *controlled* versions of neon lamps built specifically for switching current to a load. The dot inside the circle of the schematic symbol indicates a gas fill, as opposed to the hard vacuum normally seen in other electron tube designs. In the circuit shown above, the thyratron tube allows current through the load in one direction (note the polarity across the load resistor) when triggered by the small DC control voltage connected between grid and cathode. Note that the load's power source is AC, which provides a clue as to how the thyratron turns off after it's been triggered on: since AC voltage periodically passes through a condition of 0 volts between half-cycles, the current through an AC-powered load must also periodically halt. This brief pause of current between half-cycles gives the tube's gas time to cool, letting it return to its normal "off" state. Conduction may resume only if there is enough voltage applied by the AC power source (some other time in the wave's cycle) *and* if the DC control voltage allows it.

An oscilloscope display of load voltage in such a circuit would look something like this:



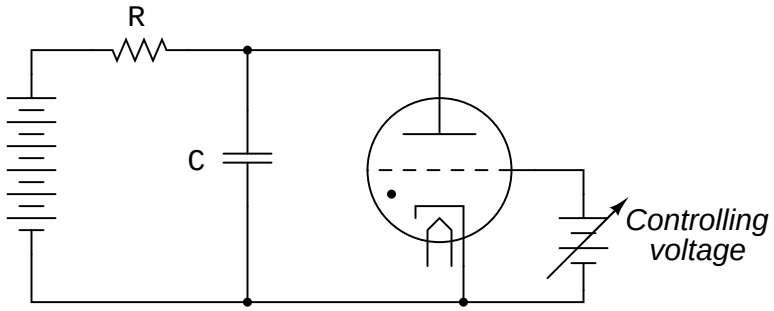
AC supply voltage

Load voltage

As the AC supply voltage climbs from zero volts to its first peak, the load voltage remains at zero (no load current) until the threshold voltage is reached. At that point, the tube switches "on"

and begins to conduct, the load voltage now following the AC voltage through the rest of the half cycle. Notice how there is load voltage (and thus load current) even when the AC voltage waveform has dropped below the threshold value of the tube. This is hysteresis at work: the tube stays in its conductive mode past the point where it first turned on, continuing to conduct until there the supply voltage drops off to almost zero volts. Because thyratron tubes are one-way (diode) devices, there is no voltage across the load through the negative half-cycle of AC. In practical thyratron circuits, multiple tubes arranged in some form of full-wave rectifier circuit to facilitate full-wave DC power to the load.

Although I'm not sure if this was ever done, someone could have applied the thyratron tube to a relaxation oscillator circuit and control the frequency with a small DC voltage between grid and cathode, making a crude voltage-controlled oscillator, otherwise known as a *VCO*. Relaxation oscillators tend to have poor frequency control, not to mention a very non-sinusoidal output, and so they exist mostly as demonstration circuits (as is the case here) or in applications where precise frequency control isn't important. Consequently, this use of a thyratron tube would not have been a very practical one.

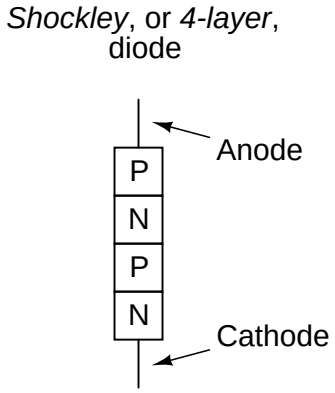


I speak of thyratron tubes in the past tense for good reason: modern semiconductor components have obsoleted thyratron tube technology for all but a few very special applications. It is no coincidence that the word *thyristor* bears so much similarity to the word *thyratron*, for this class of semiconductor components does much the same thing: use *hys*te*r*etically switch current on and off. It is these modern devices that we now turn our attention to.

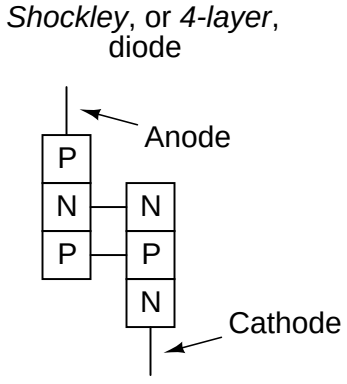
- **REVIEW:**
- Electrical *hysteresis*, the tendency for a component to remain "on" (conducting) after it begins to conduct and to remain "off" (nonconducting) after it ceases to conduct, helps to explain why lightning bolts exist as momentary surges of current rather than continuous discharges through the air.
- Simple gas-discharge tubes such as neon lamps exhibit electrical hysteresis.
- More advanced gas-discharge tubes have been made with control elements so that their "turn-on" voltage could be adjusted by an external signal. The most common of these tubes was called the *thyratron*.
- Simple oscillator circuits called *relaxation oscillators* may be created with nothing more than a resistor-capacitor charging network and a hysteretic device connected across the capacitor.

7.3 The Shockley Diode

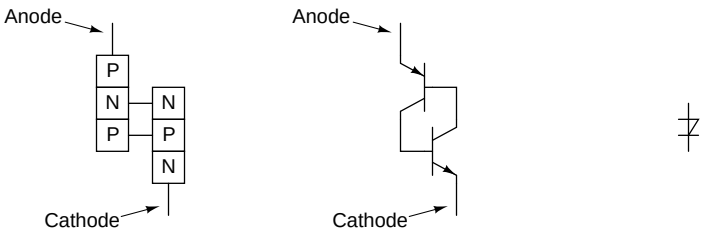
Our exploration of thyristors begins with a device called the *four-layer diode*, also known as a *PNPN diode*, or a *Shockley diode* after its inventor, William Shockley. This is not to be confused with a *Schottky diode*, that two-layer metal-semiconductor device known for its high switching speed. A crude illustration of the Shockley diode, often seen in textbooks, is a four-layer sandwich of P-N-P-N semiconductor material:



Unfortunately, this simple illustration does nothing to enlighten the viewer on how it works or why. Consider an alternative rendering of the device's construction:

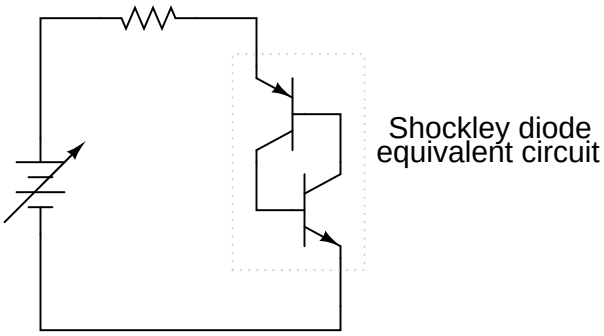


Shown like this, it appears to be a set of interconnected bipolar transistors, one PNP and the other NPN. Drawn using standard schematic symbols, and respecting the layer doping concentrations not shown in the last image, the Shockley diode looks like this:



Physical diagram *Equivalent schematic* *Schematic symbol*

Let's connect one of these devices to a source of variable voltage and see what happens:



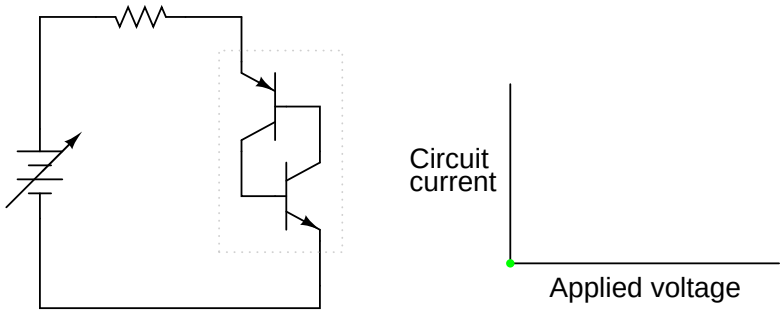
With no voltage applied, of course there will be no current. As voltage is initially increased, there will still be no current because neither transistor is able to turn on: both will be in cutoff mode. To understand why this is, consider what it takes to turn a bipolar junction transistor on: current through the base-emitter junction. As you can see in the diagram, base current through the lower transistor is controlled by the upper transistor, and the base current through the upper transistor is controlled by the lower transistor. In other words, neither transistor can turn on until the *other* transistor turns on. What we have here, in vernacular terms, is known as a Catch-22.

So how can a Shockley diode ever conduct current, if its constituent transistors stubbornly maintain themselves in a state of cutoff? The answer lies in the behavior of *real* transistors as opposed to *ideal* transistors. An ideal bipolar transistor will never conduct collector current if there is no base current, no matter how much or little voltage we apply between collector and emitter. Real transistors, on the other hand, have definite limits to how much collector-emitter voltage they can withstand before they break down and conduct. If two real transistors are connected together in this fashion to form a Shockley diode, they *will* be able to conduct if there is sufficient voltage applied by the battery between anode and cathode to cause one of them to break down. Once one transistor breaks down and begins to conduct, it will allow base current through the other transistor, causing it to turn on in a normal fashion, which then allows base current through the first transistor. The end result is that both transistors will be saturated, now keeping each other turned on instead of off.

So, we can force a Shockley diode to turn on by applying sufficient voltage between anode and cathode. As we have seen, this will inevitably cause one of the transistors to turn on, which then turns the other transistor on, ultimately "latching" both transistors on where they will tend to remain. But how do we now get the two transistors to turn off again? Even if the applied voltage is reduced

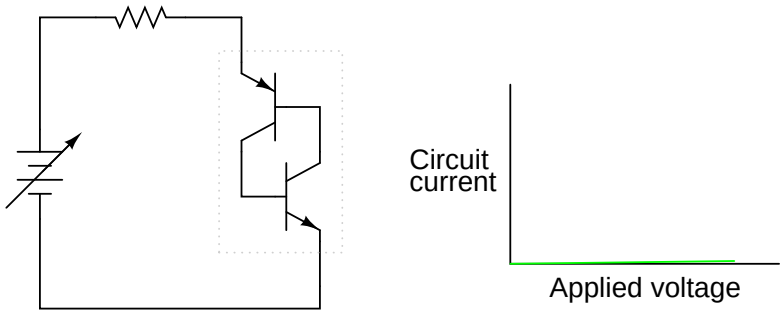
to a point well below what it took to get the Shockley diode conducting, it will remain conducting because both transistors now have base current to maintain regular, controlled conduction. The answer to this is to reduce the applied voltage to a much lower point where there is too little current to maintain transistor bias, at which point one of the transistors will cutoff, which then halts base current through the other transistor, sealing both transistors in the "off" state as they were before any voltage was applied at all.

If we graph this sequence of events and plot the results on an I/V graph, the hysteresis is very evident. First, we will observe the circuit as the DC voltage source (battery) is set to zero voltage:



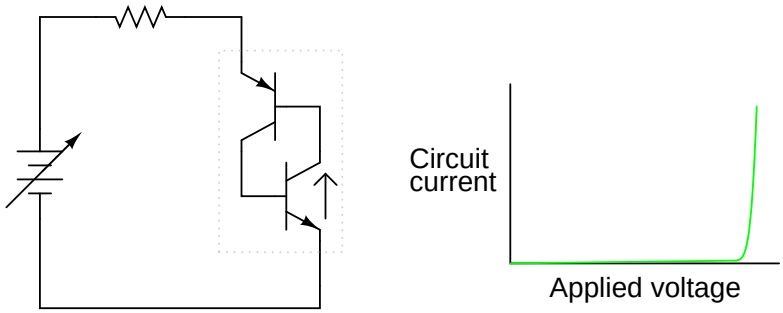
Zero applied voltage; zero current

Next, we will steadily increase the DC voltage. Current through the circuit is at or nearly at zero, as the breakdown limit has not been reached for either transistor:



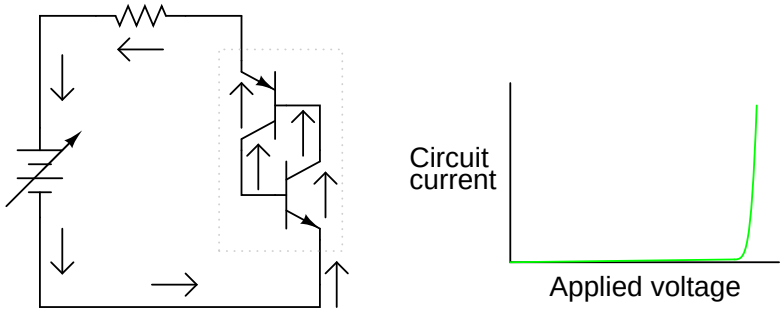
Some voltage applied, still no appreciable current

When the voltage breakdown limit of one transistor is reached, it will begin to conduct collector current even though no base current has gone through it yet. Normally, this sort of treatment would destroy a bipolar junction transistor, but the PNP junctions comprising a Shockley diode are engineered to take this kind of abuse, similar to the way a Zener diode is built to handle reverse breakdown without sustaining damage. For the sake of illustration I'll assume the lower transistor breaks down first, sending current through the base of the upper transistor:



More voltage applied; lower transistor breaks down

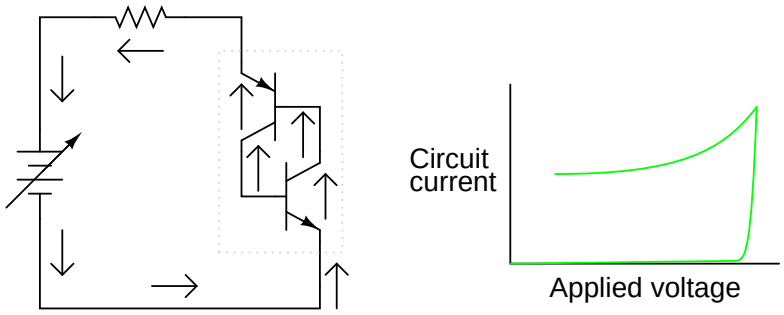
As the upper transistor receives base current, it turns on as expected. This action allows the lower transistor to conduct normally, the two transistors "sealing" themselves in the "on" state. Full current is very quickly seen in the circuit:



Transistors now fully conducting

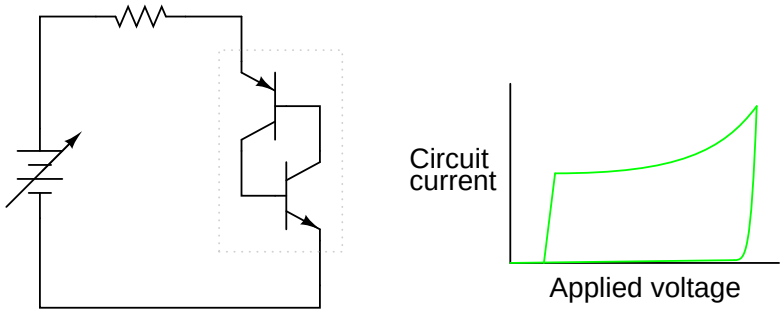
The positive feedback mentioned earlier in this chapter is clearly evident here. When one transistor breaks down, it allows current through the device structure. This current may be viewed as the "output" signal of the device. Once an output current is established, it works to hold both transistors in saturation, thus ensuring the continuation of a substantial output current. In other words, an output current "feeds back" positively to the input (transistor base current) to keep both transistors in the "on" state, thus reinforcing (or *regenerating*) itself.

With both transistors maintained in a state of saturation with the presence of ample base current, they will continue to conduct even if the applied voltage is greatly reduced from the breakdown level. The effect of positive feedback is to keep both transistors in a state of saturation despite the loss of input stimulus (the original, high voltage needed to break down one transistor and cause a base current through the other transistor):



Current maintained even when voltage is reduced

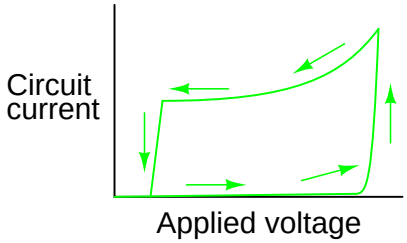
If the DC voltage source is turned down too far, though, the circuit will eventually reach a point where there isn't enough current to sustain both transistors in saturation. As one transistor passes less and less collector current, it reduces the base current for the other transistor, thus reducing base current for the first transistor. The vicious cycle continues rapidly until both transistors fall into cutoff:



If the voltage drops too low, both transistors shut off

Here, positive feedback is again at work: the fact that the cause/effect cycle between both transistors is "vicious" (a decrease in current through one works to decrease current through the other, further decreasing current through the first transistor) indicates a positive relationship between output (controlled current) and input (controlling current through the transistors' bases).

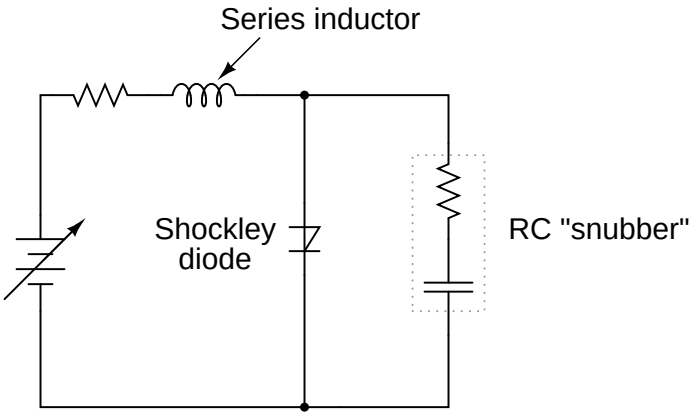
The resulting curve on the graph is classicly hysteretic: as the input signal (voltage) is increased and decreased, the output (current) does not follow the same path going down as it did going up:



Put in simple terms, the Shockley diode tends to stay on once it's turned on, and stay off once it's turned off. There is no "in-between" or "active" mode in its operation: it is a purely on or off device, as are all thyristors.

There are a few special terms applied to Shockley diodes and all other thyristor devices built upon the Shockley diode foundation. First is the term used to describe its "on" state: *latched*. The word "latch" is reminiscent of a door lock mechanism, which tends to keep the door closed once it has been pushed shut. The term *firing* refers to the initiation of a latched state. In order to get a Shockley diode to latch, the applied voltage must be increased until *breakover* is attained. Despite the fact that this action is best described in terms of transistor break*down*, the term break*over* is used instead because the end result is a pair of transistors in mutual saturation rather than destruction as would be the case with a normal transistor. A latched Shockley diode is re-set back into its nonconducting state by reducing current through it until *low-current dropout* occurs.

It should be noted that Shockley diodes may be fired in a way other than breakover: excessive *voltage rise*, or *dv/dt*. This is when the applied voltage across the diode increases at a high rate of change. This is able to cause latching (turning on) of the diode due to inherent junction capacitances within the transistors. Capacitors, as you may recall, oppose *changes* in voltage by drawing or supplying current. If the applied voltage across a Shockley diode rises at too fast a rate, those tiny capacitances will draw enough current during that time to activate the transistor pair, turning them both on. Usually, this form of latching is undesirable, and can be minimized by filtering high-frequency (fast voltage rises) from the diode with series inductors and/or parallel resistor-capacitor networks called *snubbers*:



Both the series inductor and the parallel resistor-capacitor "snubber" circuit help minimize the Shockley diode's exposure to excessively rising voltages.

The voltage rise limit of a Shockley diode is referred to as the *critical rate of voltage rise*. Manufacturers usually provide this specification for the devices they sell.

- **REVIEW:**
- Shockley diodes are four-layer PNPN semiconductor devices. They behave as a pair of interconnected PNP and NPN transistors.

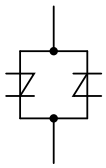
- Like all thyristors, Shockley diodes tend to stay on once they've been turned on (*latched*), and stay off once they've been turned off.

- There are two ways to latch a Shockley diode: exceed the anode-to-cathode *breakover* voltage, or exceed the anode-to-cathode *critical rate of voltage rise*.

- There is only one way to cause a Shockley diode to stop conducting, and that is to reduce the current going through it to a level below its *low-current dropout* threshold.

7.4 The DIAC

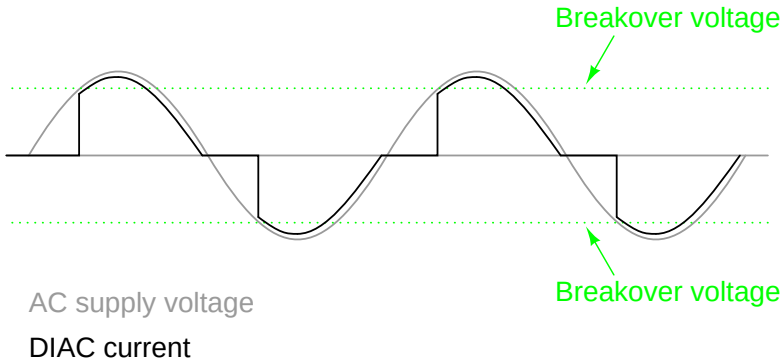
Like all diodes, Shockley diodes are unidirectional devices; that is, they only conduct current in one direction. If bidirectional (AC) operation is desired, two Shockley diodes may be joined in parallel facing different directions to form a new kind of thyristor, the *DIAC*:



DIAC equivalent circuit

DIAC schematic symbol

A DIAC operated with a DC voltage across it behaves exactly the same as a Shockley diode. With AC, however, the behavior is different from what one might expect. Because alternating current repeatedly reverses direction, DIACs will not stay latched longer than one-half cycle. If a DIAC becomes latched, it will continue to conduct current only as long as there is voltage available to push enough current in that direction. When the AC polarity reverses, as it must twice per cycle, the DIAC will drop out due to insufficient current, necessitating another breakover before it conducts again. The result is a current waveform that looks like this:



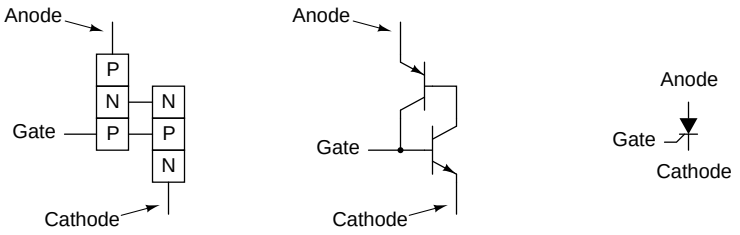
DIACs are almost never used alone, but in conjunction with other thyristor devices.

7.5 The Silicon-Controlled Rectifier (SCR)

Shockley diodes are curious devices, but rather limited in application. Their usefulness may be expanded, however, by equipping them with another means of latching. In doing so, they become true amplifying devices (if only in an on/off mode), and we refer to them as *silicon-controlled rectifiers*, or *SCR*s.

The progression from Shockley diode to SCR is achieved with one small addition, actually nothing more than a third wire connection to the existing PNPN structure:

The Silicon-Controlled Rectifier (SCR)



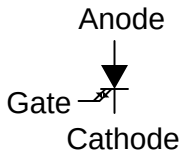
Physical diagram *Equivalent schematic* *Schematic symbol*

If an SCR's gate is left *floating* (disconnected), it behaves exactly as a Shockley diode. It may be latched by breakover voltage or by exceeding the critical rate of voltage rise between anode and cathode, just as with the Shockley diode. Dropout is accomplished by reducing current until one or both internal transistors fall into cutoff mode, also like the Shockley diode. However, because the gate terminal connects directly to the base of the lower transistor, it may be used as an alternative means to latch the SCR. By applying a small voltage between gate and cathode, the lower transistor will be forced *on* by the resulting base current, which will cause the upper transistor to conduct, which then supplies the lower transistor's base with current so that it no longer needs to be activated by a gate voltage. The necessary gate current to initiate latch-up, of course, will be much lower than the current through the SCR from cathode to anode, so the SCR does achieve a measure of amplification.

This method of securing SCR conduction is called *triggering*, and it is by far the most common way that SCRs are latched in actual practice. In fact, SCRs are usually chosen so that their breakover voltage is far beyond the greatest voltage expected to be experienced from the power source, so that it can be turned on *only* by an intentional voltage pulse applied to the gate.

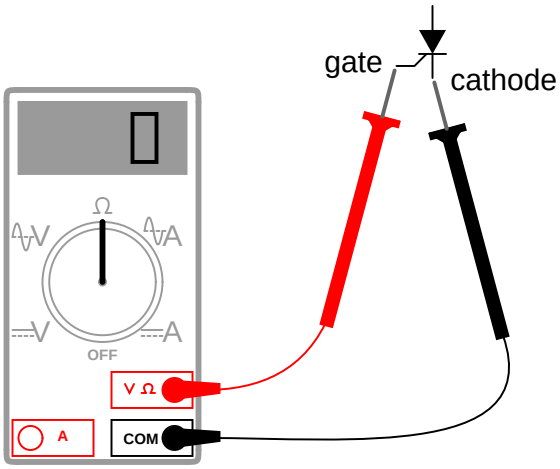
It should be mentioned that SCRs may *sometimes* be turned off by directly shorting their gate and cathode terminals together, or by "reverse-triggering" the gate with a negative voltage (in reference to the cathode), so that the lower transistor is forced into cutoff. I say this is "sometimes" possible because it involves shunting all of the upper transistor's collector current past the lower transistor's base. This current may be substantial, making triggered shut-off of an SCR difficult at best. A variation of the SCR, called a *Gate-Turn-Off* thyristor, or *GTO*, makes this task easier. But even with a GTO, the gate current required to turn it off may be as much as 20% of the anode (load) current! The schematic symbol for a GTO is shown in the following illustration:

Gate Turn-Off thyristor (GTO)



SCRs and GTOs share the same equivalent schematics (two transistors connected in a positive-feedback fashion), the only differences being details of construction designed to grant the NPN transistor a greater β than the PNP. This allows a smaller gate current (forward or reverse) to exert a greater degree of control over conduction from cathode to anode, with the PNP transistor's latched state being more dependent upon the NPN's than vice versa. The Gate-Turn-Off thyristor is also known by the name of *Gate-Controlled Switch*, or *GCS*.

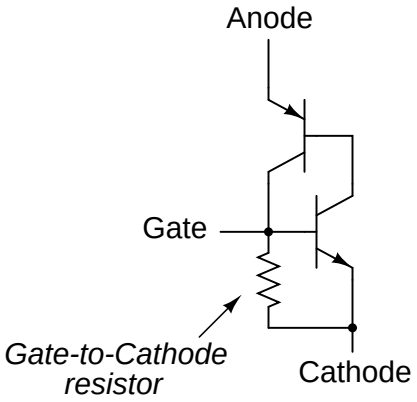
A rudimentary test of SCR function, or at least terminal identification, may be performed with an ohmmeter. Because the internal connection between gate and cathode is a single PN junction, a meter should indicate continuity between these terminals with the red test lead on the gate and the black test lead on the cathode like this:



All other continuity measurements performed on an SCR will show "open" ("OL" on some digital multimeter displays). It must be understood that this test is very crude and does *not* constitute a comprehensive assessment of the SCR. It is possible for an SCR to give good ohmmeter indications and still be defective. Ultimately, the only way to test an SCR is to subject it to a load current.

If you are using a multimeter with a "diode check" function, the gate-to-cathode junction voltage indication you get may or may not correspond to what's expected of a silicon PN junction (approximately 0.7 volts). In some cases, you will read a much lower junction voltage: mere hundredths of a volt. This is due to an internal resistor connected between the gate and cathode incorporated within some SCRs. This resistor is added to make the SCR less susceptible to false triggering by spurious voltage spikes, from circuit "noise" or from static electric discharge. In other words, having a resistor connected across the gate-cathode junction requires that a *strong* triggering signal (substantial current) be applied to latch the SCR. This feature is often found in larger SCRs, not

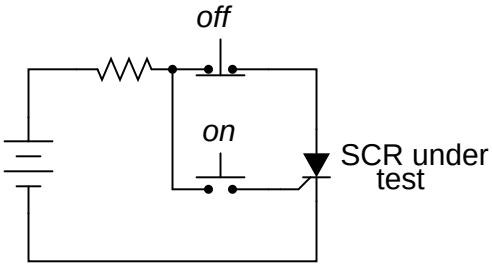
on small SCRs. Bear in mind that an SCR with an internal resistor connected between gate and cathode will indicate continuity *in both directions* between those two terminals:



"Normal" SCRs, lacking this internal resistor, are sometimes referred to as *sensitive gate* SCRs due to their ability to be triggered by the slightest positive gate signal.

The test circuit for an SCR is both practical as a diagnostic tool for checking suspected SCRs and also an excellent aid to understanding basic SCR operation. A DC voltage source is used for powering the circuit, and two pushbutton switches are used to latch and unlatch the SCR, respectively:

SCR testing circuit



Actuating the normally-open "on" pushbutton switch connects the gate to the anode, allowing current from the negative terminal of the battery, through the cathode-gate PN junction, through the switch, through the load resistor, and back to the battery. This gate current should force the SCR to latch on, allowing current to go directly from cathode to anode without further triggering through the gate. When the "on" pushbutton is released, the load should remain energized.

Pushing the normally-closed "off" pushbutton switch breaks the circuit, forcing current through the SCR to halt, thus forcing it to turn off (low-current dropout).

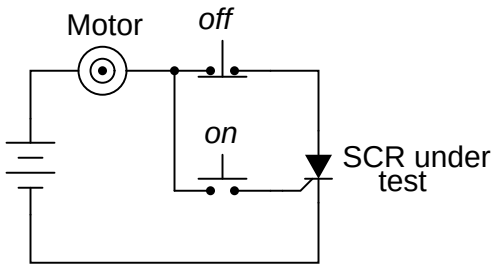
If the SCR fails to latch, the problem may be with the load and not the SCR. There is a certain minimum amount of load current required to hold the SCR latched in the "on" state. This minimum current level is called the *holding current*. A load with too great a resistance value may not draw enough current to keep an SCR latched when gate current ceases, thus giving the false impression of a bad (unlatchable) SCR in the test circuit. Holding current values for different SCRs should

be available from the manufacturers. Typical holding current values range from 1 milliamp to 50 milliamps or more for larger units.

For the test to be fully comprehensive, more than the triggering action needs to be tested. The forward breakover voltage limit of the SCR could be tested by increasing the DC voltage supply (with no pushbuttons actuated) until the SCR latches all on its own. Beware that a breakover test may require very high voltage: many power SCRs have breakover voltage ratings of 600 volts or more! Also, if a pulse voltage generator is available, the critical rate of voltage rise for the SCR could be tested in the same way: subject it to pulsing supply voltages of different V/time rates with no pushbutton switches actuated and see when it latches.

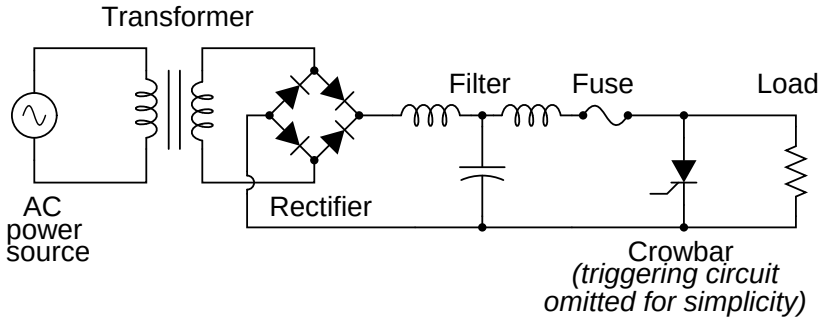
In this simple form, the SCR test circuit could suffice as a start/stop control circuit for a DC motor, lamp, or other practical load:

DC motor start/stop control circuit



Another practical use for the SCR in a DC circuit is as a *crowbar* device for overvoltage protection. A "crowbar" circuit consists of an SCR placed in parallel with the output of a DC power supply, for the purpose of placing a direct short-circuit on the output of that supply to prevent excessive voltage from reaching the load. Damage to the SCR and power supply is prevented by the judicious placement of a fuse or substantial series resistance ahead of the SCR to limit short-circuit current:

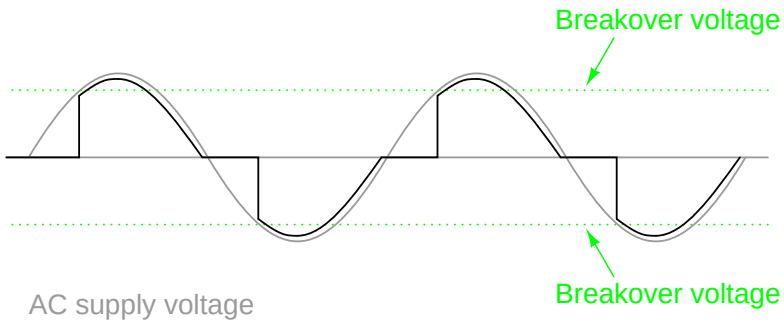
Crowbar as used in an AC-DC power supply



Some device or circuit sensing the output voltage will be connected to the gate of the SCR, so that when an overvoltage condition occurs, voltage will be applied between the gate and cathode, triggering the SCR and forcing the fuse to blow. The effect will be approximately the same as

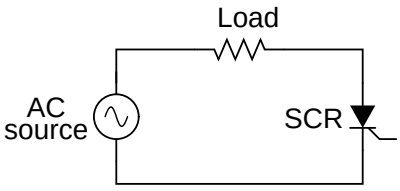
dropping a solid steel crowbar directly across the output terminals of the power supply, hence the name of the circuit.

Most applications of the SCR are for AC power control, despite the fact that SCRs are inherently DC (unidirectional) devices. If bidirectional circuit current is required, multiple SCRs may be used, with one or more facing each direction to handle current through both half-cycles of the AC wave. The primary reason SCRs are used at all for AC power control applications is the unique response of a thyristor to an alternating current. As we saw in the case of the thyratron tube (the electron tube version of the SCR) and the DIAC, a hysteretic device triggered on during a portion of an AC half-cycle will latch and remain on throughout the remainder of the half-cycle until the AC current decreases to zero, as it must to begin the next half-cycle. Just prior to the zero-crossover point of the current waveform, the thyristor will turn off due to insufficient current (this behavior is also known as *natural commutation*) and must be fired again during the next cycle. The result is a circuit current equivalent to a "chopped up" sine wave. For review, here is the graph of a DIAC's response to an AC voltage whose peak exceeds the breakover voltage of the DIAC:



With the DIAC, that breakover voltage limit was a fixed quantity. With the SCR, we have control over exactly when the device becomes latched by triggering the gate at any point in time along the waveform. By connecting a suitable control circuit to the gate of an SCR, we can "chop" the sine wave at any point to allow for time-proportioned power control to a load.

Take the following circuit as an example. Here, an SCR is positioned in a circuit to control power to a load from an AC source:

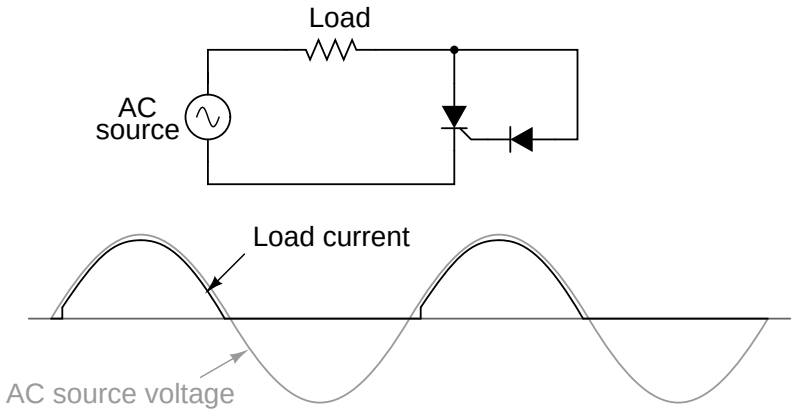


Being a unidirectional (one-way) device, at most we can only deliver half-wave power to the load, in the half-cycle of AC where the supply voltage polarity is positive on the top and negative on the bottom. However, for demonstrating the basic concept of time-proportional control, this simple circuit is better than one controlling full-wave power (which would require two SCRs).

With no triggering to the gate, and the AC source voltage well below the SCR's breakover voltage rating, the SCR will never turn on. Connecting the SCR gate to the anode through a normal rectifying diode (to prevent reverse current through the gate in the event of the SCR containing

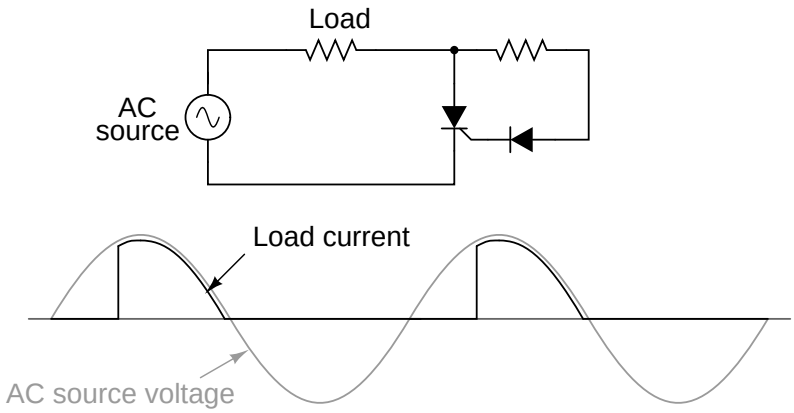
a built-in gate-cathode resistor), will allow the SCR to be triggered almost immediately at the beginning of every positive half-cycle:

Gate connected directly to anode through a diode;
nearly complete half-wave current through load

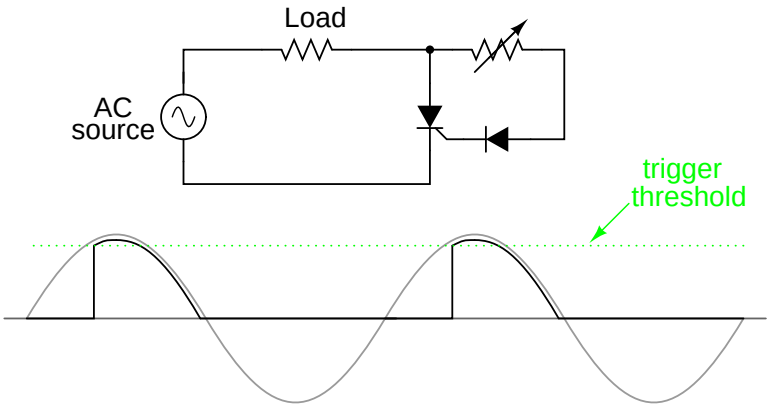


We can delay the triggering of the SCR, however, by inserting some resistance into the gate circuit, thus increasing the amount of voltage drop required before there is enough gate current to trigger the SCR. In other words, if we make it harder for electrons to flow through the gate by adding a resistance, the AC voltage will have to reach a higher point in its cycle before there will be enough gate current to turn the SCR on. The result looks like this:

Resistance inserted in gate circuit;
less than half-wave current through load



With the half-sine wave chopped up to a greater degree by delayed triggering of the SCR, the load receives less average power (power is delivered for less time throughout a cycle). By making the series gate resistor variable, we can make adjustments to the time-proportioned power:

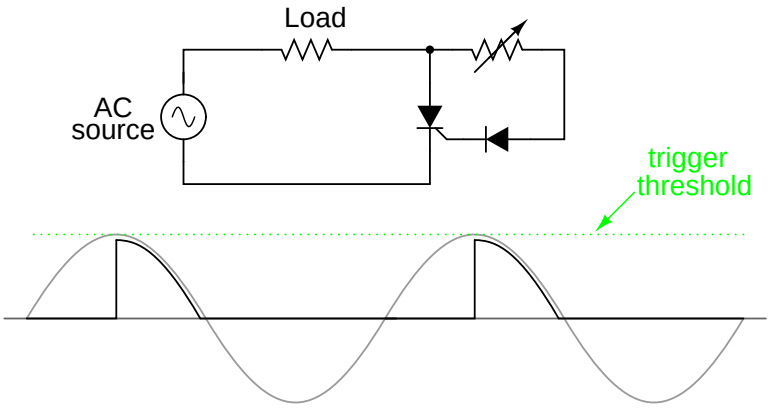


Increasing the resistance raises the threshold level, causing less power to be delivered to the load.

Decreasing the resistance lowers the threshold level, causing more power to be delivered to the load.

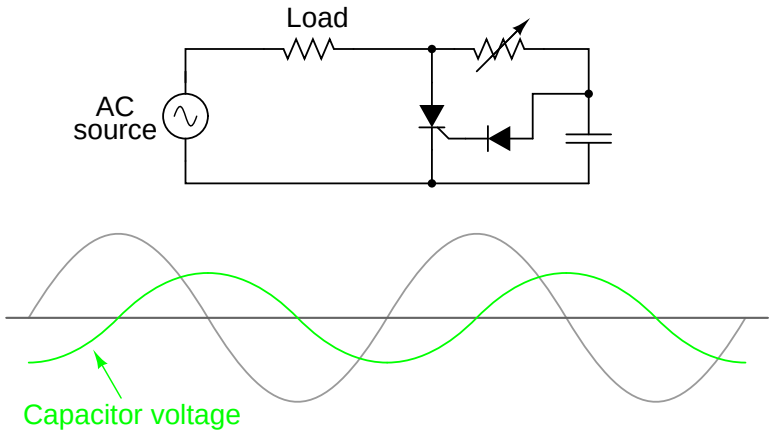
Unfortunately, this control scheme has a significant limitation. In using the AC source waveform for our SCR triggering signal, we limit control to the first half of the waveform's half-cycle. In other words, there is no way for us to wait until *after* the wave's peak to trigger the SCR. This means we can turn down the power only to the point where the SCR turns on at the very peak of the wave:

Circuit at minimum power setting



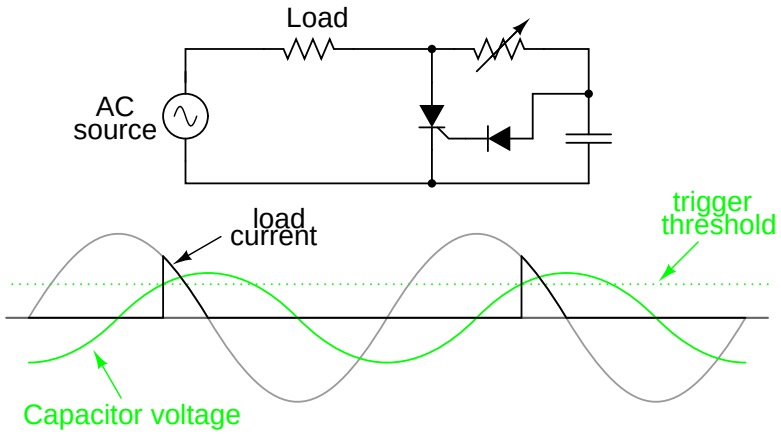
Raising the trigger threshold any more will cause the circuit to not trigger at all, since not even the peak of the AC power voltage will be enough to trigger the SCR. The result will be no power to the load.

An ingenious solution to this control dilemma is found in the addition of a phase-shifting capacitor to the circuit:



The smaller waveform shown on the graph is voltage across the capacitor. For the sake of illustrating the phase shift, I'm assuming a condition of maximum control resistance where the SCR is not triggering at all and there is no load current, save for what little current goes through the control resistor and capacitor. This capacitor voltage will be phase-shifted anywhere from 0^o to 90^o lagging behind the power source AC waveform. When this phase-shifted voltage reaches a high enough level, the SCR will trigger.

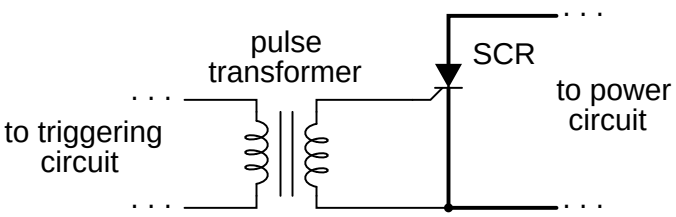
Assuming there is periodically enough voltage across the capacitor to trigger the SCR, the resulting load current waveform will look something like this:



Because the capacitor waveform is still *rising* after the main AC power waveform has reached its peak, it becomes possible to trigger the SCR at a threshold level beyond that peak, thus chopping the load current wave further than it was possible with the simpler circuit. In reality, the capacitor voltage waveform is a bit more complex that what is shown here, its sinusoidal shape distorted every time the SCR latches on. However, what I'm trying to illustrate here is the delayed triggering action gained with the phase-shifting RC network, and so a simplified, undistorted waveform serves the purpose well.

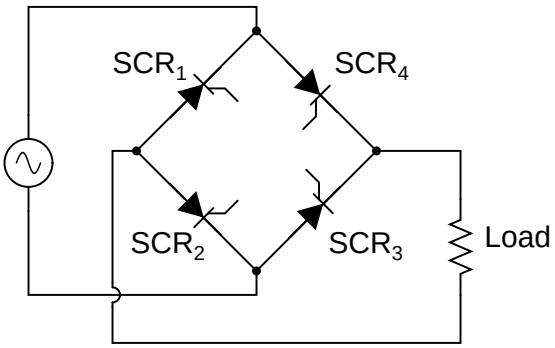
SCRs may also be triggered, or "fired," by more complex circuits. While the circuit previously shown is sufficient for a simple application like a lamp control, large industrial motor controls often

rely on more sophisticated triggering methods. Sometimes, pulse transformers are used to couple a triggering circuit to the gate and cathode of an SCR to provide electrical isolation between the triggering and power circuits:

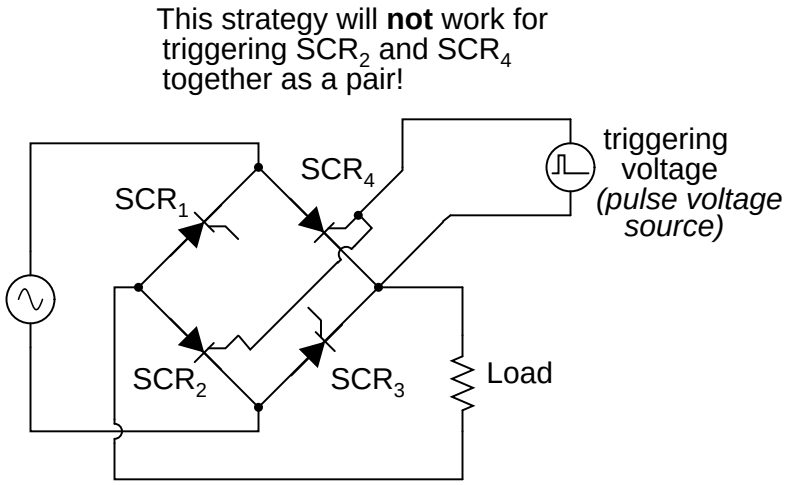


When multiple SCRs are used to control power, their cathodes are often *not* electrically common, making it difficult to connect a single triggering circuit to all SCRs equally. An example of this is the *controlled bridge rectifier* shown here:

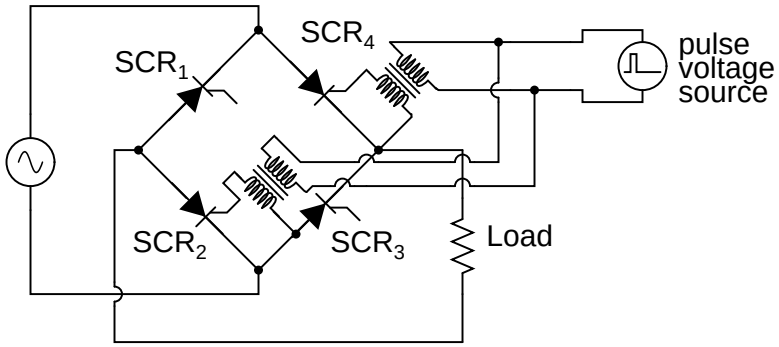
Controlled bridge rectifier



In any bridge rectifier circuit, the rectifying diodes (or in this case, the rectifying SCRs) must conduct in opposite pairs. SCR_1 and SCR_3 must be fired simultaneously, and likewise SCR_2 and SCR_4 must be fired together as a pair. As you will notice, though, these pairs of SCRs do not share the same cathode connections, meaning that it would not work to simply parallel their respective gate connections and connect a single voltage source to trigger both:

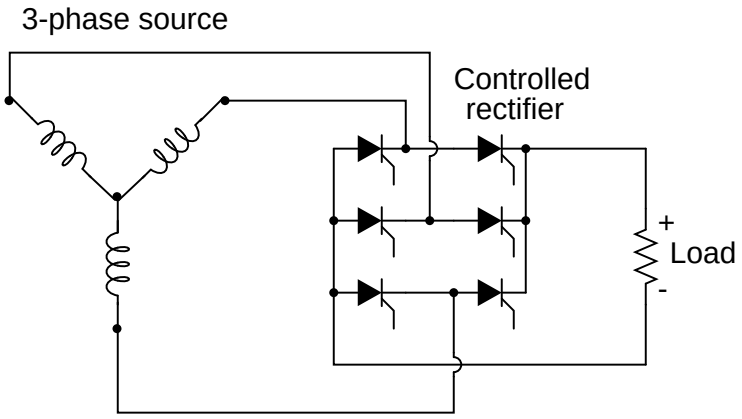


Although the triggering voltage source shown will trigger SCR_4, it will not trigger SCR_2 properly because the two thyristors do not share a common cathode connection to reference that triggering voltage. Pulse transformers connecting the two thyristor gates to a common triggering voltage source *will* work, however:



Bear in mind that this circuit only shows the gate connections for two out of the four SCRs. Pulse transformers and triggering sources for SCR_1 and SCR_3, as well as the details of the pulse sources themselves, have been omitted for the sake of simplicity.

Controlled bridge rectifiers are not limited to single-phase designs. In most industrial control systems, AC power is available in three-phase form for maximum efficiency, and solid-state control circuits are built to take advantage of that. A three-phase controlled rectifier circuit built with SCRs, without pulse transformers or triggering circuitry shown, would look like this:



- **REVIEW:**

- A *Silicon-Controlled Rectifier*, or *SCR*, is essentially a Shockley diode with an extra terminal added. This extra terminal is called the *gate*, and it is used to *trigger* the device into conduction (latch it) by the application of a small voltage.

- To trigger, or *fire*, an SCR, voltage must be applied between the gate and cathode, positive to the gate and negative to the cathode. When testing an SCR, a momentary connection between the gate and anode is sufficient in polarity, intensity, and duration to trigger it.

- SCRs may be fired by intentional triggering of the gate terminal, excessive voltage (breakdown) between anode and cathode, or excessive rate of voltage rise between anode and cathode. SCRs may be turned off by anode current falling below the *holding current value* (low-current dropout), or by "reverse-firing" the gate (applying a negative voltage to the gate). Reverse-firing is only sometimes effective, and always involves high gate current.

- A variant of the SCR, called a Gate-Turn-Off thyristor (GTO), is specifically designed to be turned off by means of reverse triggering. Even then, reverse triggering requires fairly high current: typically 20% of the anode current.

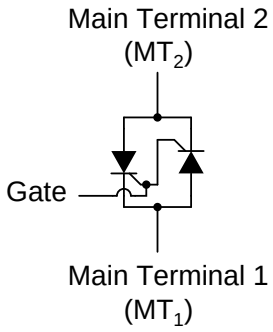
- SCR terminals may be identified by a continuity meter: the only two terminals showing any continuity between them at all should be the gate and cathode. Gate and cathode terminals connect to a PN junction inside the SCR, so a continuity meter should obtain a diode-like reading between these two terminals with the red (+) lead on the gate and the black (-) lead on the cathode. Beware, though, that some large SCRs have an internal resistor connected between gate and cathode, which will affect any continuity readings taken by a meter.

- SCRs are true *rectifiers*: they only allow current through them in one direction. This means they cannot be used alone for full-wave AC power control.

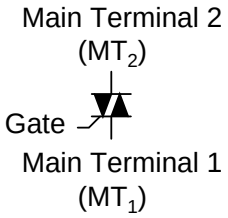
- If the diodes in a rectifier circuit are replaced by SCRs, you have the makings of a *controlled* rectifier circuit, whereby DC power to a load may be time-proportioned by triggering the SCRs at different points along the AC power waveform.

7.6 The TRIAC

SCRs are unidirectional (one-way) current devices, making them useful for controlling DC only. If two SCRs are joined in back-to-back parallel fashion just like two Shockley diodes were joined together to form a DIAC, we have a new device known as the *TRIAC*:

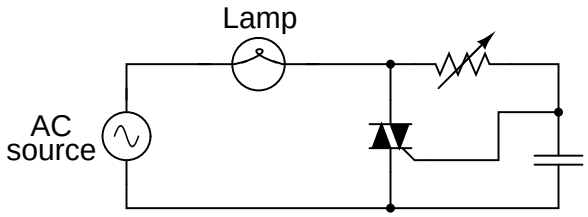


TRIAC equivalent circuit



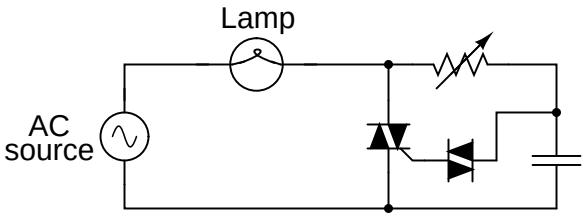
TRIAC schematic symbol

Because individual SCRs are more flexible to use in advanced control systems, they are more commonly seen in circuits like motor drives, while TRIACs are usually seen in simple, low-power applications like household dimmer switches. A simple lamp dimmer circuit is shown here, complete with the phase-shifting resistor-capacitor network necessary for after-peak firing.



TRIACs are notorious for not firing *symmetrically*. This means they usually won't trigger at the exact same gate voltage level for one polarity as for the other. Generally speaking, this is undesirable, because unsymmetrical firing results in a current waveform with a greater variety of harmonic frequencies. Waveforms that are symmetrical above and below their average centerlines are comprised of only odd-numbered harmonics. Unsymmetrical waveforms, on the other hand, contain even-numbered harmonics (which may or may not be accompanied by odd-numbered harmonics as well).

In the interest of reducing total harmonic content in power systems, the fewer and less diverse the harmonics, the better – one more reason why individual SCRs are favored over TRIACs for complex, high-power control circuits. One way to make the TRIAC's current waveform more symmetrical is to use a device external to the TRIAC to time the triggering pulse. A DIAC placed in series with the gate does a fair job of this:

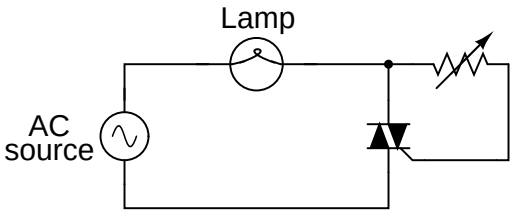


DIAC breakover voltages tend to be much more symmetrical (the same in one polarity as the other) than TRIAC triggering voltage thresholds. Since the DIAC prevents any gate current until the triggering voltage has reached a certain, repeatable level in either direction, the firing point of the TRIAC from one half-cycle to the next tends to be more consistent, and the waveform more symmetrical above and below its centerline.

Practically all the characteristics and ratings of SCRs apply equally to TRIACs, except that TRIACs of course are bidirectional (can handle current in both directions). Not much more needs to be said about this device except for an important caveat concerning its terminal designations.

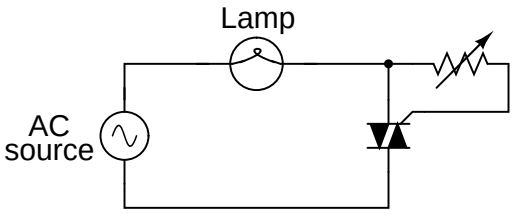
From the equivalent circuit diagram shown earlier, one might think that main terminals 1 and 2 were interchangeable. They are not! Although it is helpful to imagine the TRIAC as being composed of two SCRs joined together, it in fact is constructed from a single piece of semiconducting material, appropriately doped and layered. The actual operating characteristics may differ slightly from that of the equivalent model.

This is made most evident by contrasting two simple circuit designs, one that works and one that doesn't. The following two circuits are a variation of the lamp dimmer circuit shown earlier, the phase-shifting capacitor and DIAC removed for simplicity's sake. Although the resulting circuit lacks the fine control ability of the more complex version (with capacitor and DIAC), it *does* function:



Suppose we were to swap the two main terminals of the TRIAC around. According to the equivalent circuit diagram shown earlier in this section, the swap should make no difference. The circuit ought to work:

This circuit ***will not work****!*



However, if this circuit is built, it will be found that it does not work! The load will receive no power, the TRIAC refusing to fire at all, no matter how low or high a resistance value the control resistor is set to. The key to successfully triggering a TRIAC is to make sure the gate receives its triggering current from the *main terminal 2* side of the circuit (the main terminal on the opposite side of the TRIAC symbol from the gate terminal). Identification of the MT_1 and MT_2 terminals must be done via the TRIAC's part number with reference to a data sheet or book.

- **REVIEW:**
- A *TRIAC* acts much like two SCRs connected back-to-back for bidirectional (AC) operation.
- TRIAC controls are more often seen in simple, low-power circuits than complex, high-power circuits. In large power control circuits, multiple SCRs tend to be favored.
- When used to control AC power to a load, TRIACs are often accompanied by DIACs connected in series with their gate terminals. The DIAC helps the TRIAC fire more symmetrically (more consistently from one polarity to another).
- Main terminals 1 and 2 on a TRIAC are *not* interchangeable.
- To successfully trigger a TRIAC, gate current must come from the *main terminal 2* (MT_2) side of the circuit!

7.7 Optothyristors

Like bipolar transistors, SCRs and TRIACs are also manufactured as light-sensitive devices, the action of impinging light replacing the function of triggering voltage.

Optically-controlled SCRs are often known by the acronym *LASCR*, or **L**ight **A**ctivated **SCR**. Its symbol, not surprisingly, looks like this:

Light **A**ctivated **SCR**

LASCR

Optically-controlled TRIACs don't receive the honor of having their own acronym, but instead are humbly known as opto-TRIACs. Their schematic symbol looks like this:

Opto-TRIAC

Optothyristors (a general term for either the LASCR or the opto-TRIAC) are commonly found inside sealed "optoisolator" modules.

7.8 The Unijunction Transistor (UJT) – PENDING

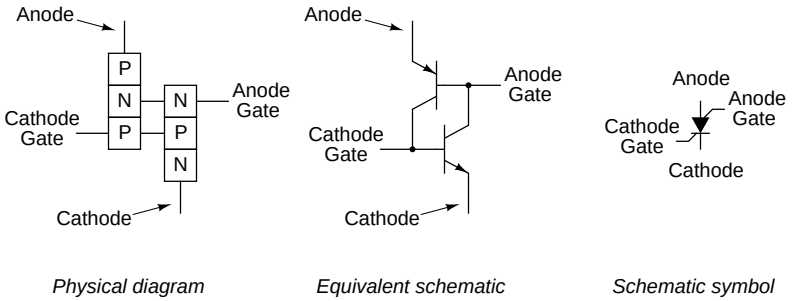
Programmable Unijunction Transistors (PUTs).

- **REVIEW:**
-
-
-

7.9 The Silicon-Controlled Switch (SCS)

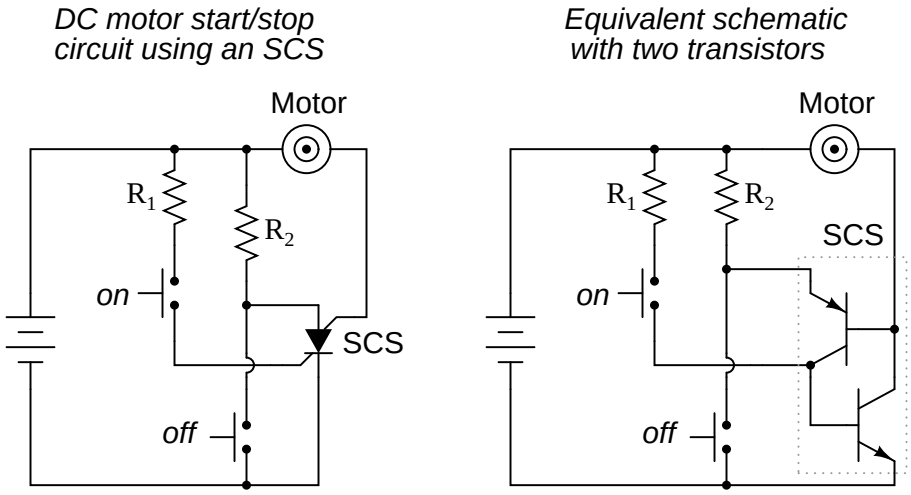
If we take the equivalent circuit for an SCR and add another external terminal, connected to the base of the top transistor and the collector of the bottom transistor, we have a device known as a *silicon-controlled-switch*, or *SCS*:

The Silicon-Controlled Switch (SCS)



Physical diagram *Equivalent schematic* *Schematic symbol*

This extra terminal allows more control to be exerted over the device, particularly in the mode of *forced commutation*, where an external signal forces it to turn off while the main current through the device has not yet fallen below the holding current value. Consider the following circuit:



When the "on" pushbutton switch is actuated, there is a voltage applied between the cathode gate and the cathode, forward-biasing the lower transistor's base-emitter junction, and turning it on. The top transistor of the SCS is ready to conduct, having been supplied with a current path from its emitter terminal (the SCS's anode terminal) through resistor R_2 to the positive side of the power supply. As in the case of the SCR, both transistors turn on and maintain each other in the "on" mode. When the lower transistor turns on, it conducts the motor's load current, and the motor starts and runs.

The motor may be stopped by interrupting the power supply, as with an SCR, and this is called *natural commutation*. However, the SCS provides us with another means of turning off: *forced commutation* by shorting the anode terminal to the cathode. If this is done (by actuating the "off" pushbutton switch), the upper transistor within the SCS will lose its emitter current, thus halting current through the base of the lower transistor. When the lower transistor turns off, it breaks the circuit for base current through the top transistor (securing its "off" state), and the motor (making it stop). The SCS will remain in the off condition until such time that the "on" pushbutton switch is re-actuated.

- **REVIEW:**

- A *silicon-controlled switch*, or *SCS*, is essentially an SCR with an extra gate terminal.

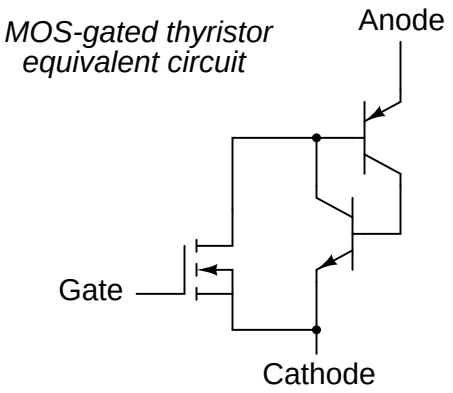
- Typically, the load current through an SCS is carried by the *anode gate* and *cathode* terminals, with the *cathode gate* and *anode* terminals sufficing as control leads.

- An SCS is turned on by applying a positive voltage between the *cathode gate* and *cathode* terminals. It may be turned off (forced commutation) by applying a negative voltage between the *anode* and *cathode* terminals, or simply by shorting those two terminals together. The *anode* terminal must be kept positive with respect to the cathode in order for the SCS to latch.

7.10 Field-effect-controlled thyristors

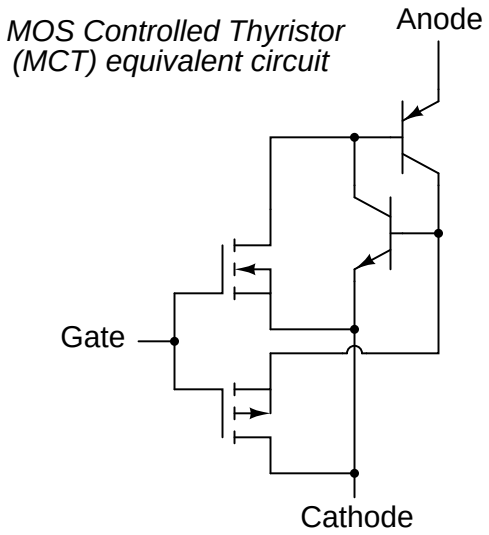
Two relatively recent technologies designed to reduce the "driving" (gate trigger current) requirements of classic thyristor devices are the *MOS-gated thyristor* and the *MOS Controlled Thyristor*, or *MCT*.

The MOS-gated thyristor uses a MOSFET to initiate conduction through the upper (PNP) transistor of a normal thyristor structure, thus triggering the device. Since a MOSFET requires negligible current to "drive" (cause it to saturate), this makes the thyristor as a whole very easy to trigger:



Given the fact that ordinary SCRs are quite easy to "drive" as it is, the practical advantage of using an even more sensitive device (a MOSFET) to initiate triggering is debatable. Also, placing a MOSFET at the gate input of the thyristor now makes it *impossible* to turn it off by a reverse-triggering signal. Only low-current dropout can make this device stop conducting after it has been latched.

A device of arguably greater value would be a fully-controllable thyristor, whereby a small gate signal could both trigger the thyristor and force it to turn off. Such a device does exist, and it is called the *MOS Controlled Thyristor*, or *MCT*. It uses a pair of MOSFETs connected to a common gate terminal, one to trigger the thyristor and the other to "untrigger" it:



A positive gate voltage (with respect to the cathode) turns on the upper (N-channel) MOSFET, allowing base current through the upper (PNP) transistor, which latches the transistor pair in an "on" state. Once both transistors are fully latched, there will be little voltage dropped between anode and cathode, and the thyristor will remain latched so long as the controlled current exceeds the minimum (holding) current value. However, if a negative gate voltage is applied (with respect to the anode, which is at nearly the same voltage as the cathode in the latched state), the lower MOSFET will turn on and "short" between the lower (NPN) transistor's base and emitter terminals, thus forcing it into cutoff. Once the NPN transistor cuts off, the PNP transistor will drop out of conduction, and the whole thyristor turns off. Gate voltage has full control over conduction through the MCT: to turn it on and to turn it off.

This device is still a thyristor, though. If there is zero voltage applied between gate and cathode, neither MOSFET will turn on. Consequently, the bipolar transistor pair will remain in whatever state it was last in (hysteresis). So, a brief positive pulse to the gate turns the MCT on, a brief negative pulse forces it off, and no applied gate voltage lets it remain in whatever state it is already in. In essence, the MCT is a latching version of the IGBT (Insulated Gate Bipolar Transistor).

- **REVIEW:**
- A *MOS-gated thyristor* uses an N-channel MOSFET to trigger a thyristor, resulting in an extremely low gate current requirement.
- A *MOS Controlled Thyristor*, or *MCT*, uses two MOSFETS to exert full control over the thyristor. A positive gate voltage triggers the device, while a negative gate voltage forces it to turn off. Zero gate voltage allows the thyristor to remain in whatever state it was previously in (off, or latched on).

Chapter 8

OPERATIONAL AMPLIFIERS

Contents

8.1 Introduction

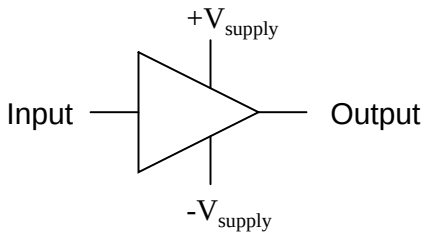
The operational amplifier is arguably the most useful single device in analog electronic circuitry. With only a handful of external components, it can be made to perform a wide variety of analog signal processing tasks. It is also quite affordable, most general-purpose amplifiers selling for under a 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.

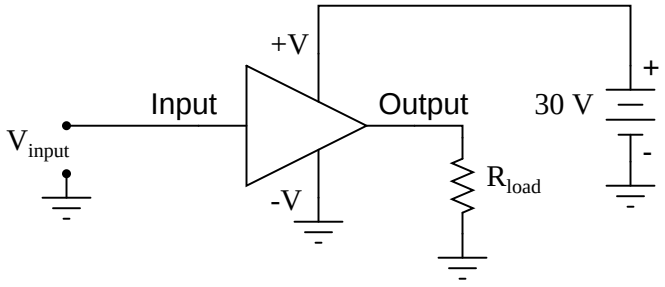
8.2 Single-ended and differential amplifiers

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



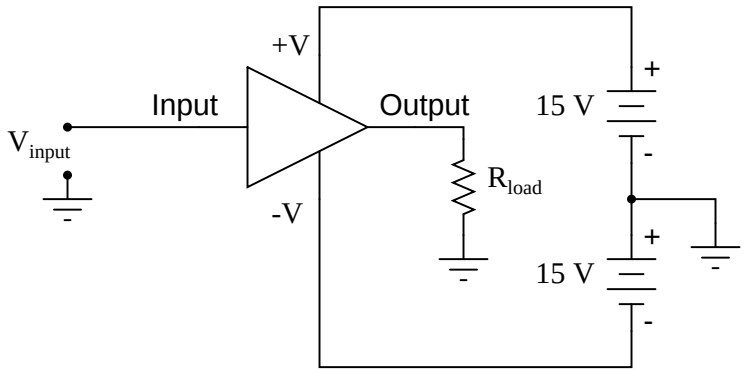
The +V and -V connections denote the positive and negative sides of the DC power supply, respectively. The input and output voltage connections are shown as single conductors, because it is assumed that all signal voltages are referenced to a common connection in the circuit called *ground*. Often (but 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"

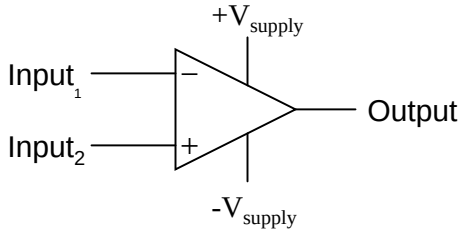
between +V and -V. Sometimes the split power supply configuration is referred to as a *dual* power supply.



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

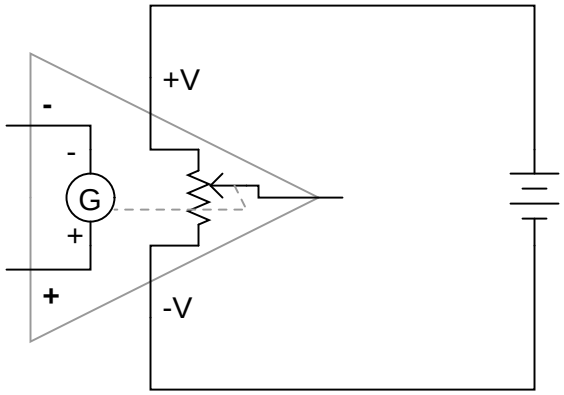
Voltage output equation: $V_{out} = A_V(Input_2 - Input_1)$

or

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

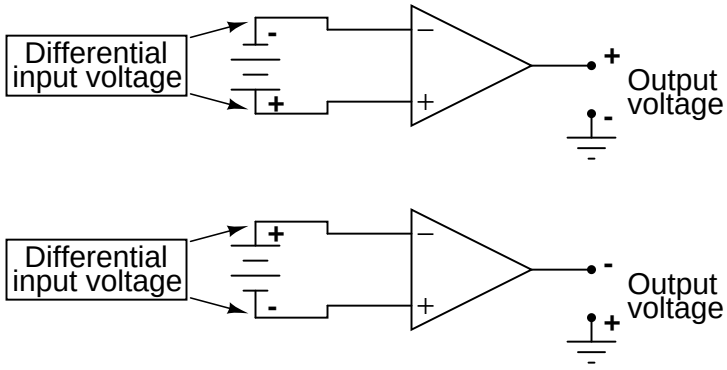
An increasingly positive voltage on the (+) input tends to drive the output voltage more positive, and an increasingly positive voltage on the (-) input tends to drive the output voltage more negative. Likewise, an increasingly negative voltage on the (+) input tends to drive the output negative as well, 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:

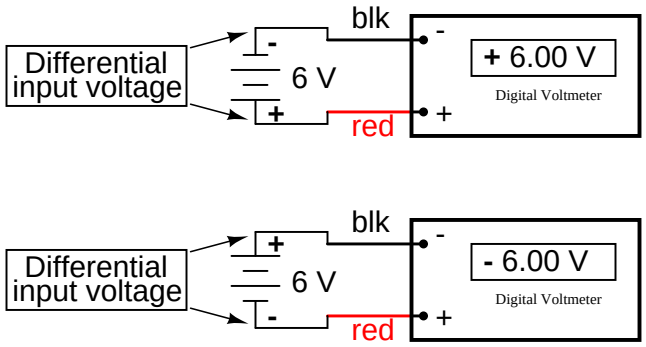


Bear in mind that the above illustration is only a *model* to aid in understanding the behavior of a differential amplifier. It is not a realistic schematic of its actual design. The "G" symbol represents a galvanometer, a sensitive voltmeter movement. The potentiometer connected between +V and -V provides 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: $V_{out} = A_V(V_{noninv} - V_{inv})$

8.3 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:

$$i_C = C \frac{dv}{dt} \qquad F = m \frac{dv}{dt}$$

Where,

i_C = Instantaneous current through capacitor

C = Capacitance in farads

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

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.

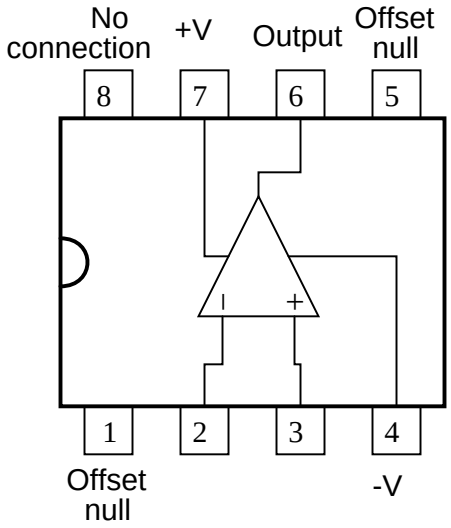
Electronic circuits are very easy and inexpensive to create compared to complex physical systems, so this kind of analog electronic simulation was widely used in the research and development of mechanical systems. For realistic simulation, though, amplifier circuits of high accuracy and easy configurability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

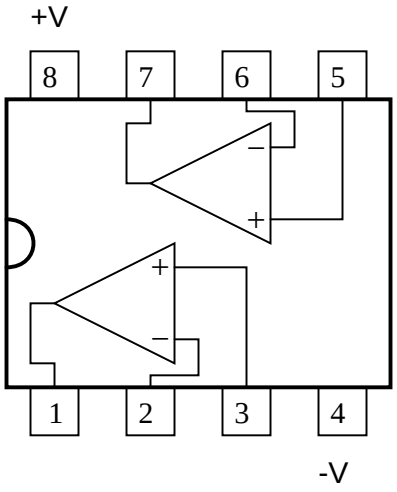
The following diagram shows the pin connections for single op-amps (741 included) when housed in an 8-pin DIP (**D**ual **I**nline **P**ackage) integrated circuit:

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:

Dual op-amp in 8-pin DIP

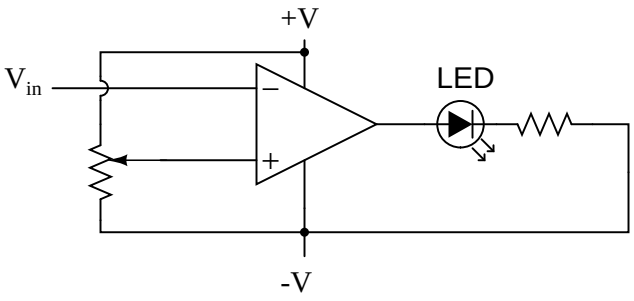


Operational amplifiers are also available four to a package, usually in 14-pin DIP arrangements. Unfortunately, pin assignments aren't as standard for these "quad" op-amps as they are for the "dual" or single units. Consult the manufacturer datasheet(s) for details.

Practical operational amplifier voltage gains are in the range of 200,000 or more, which makes

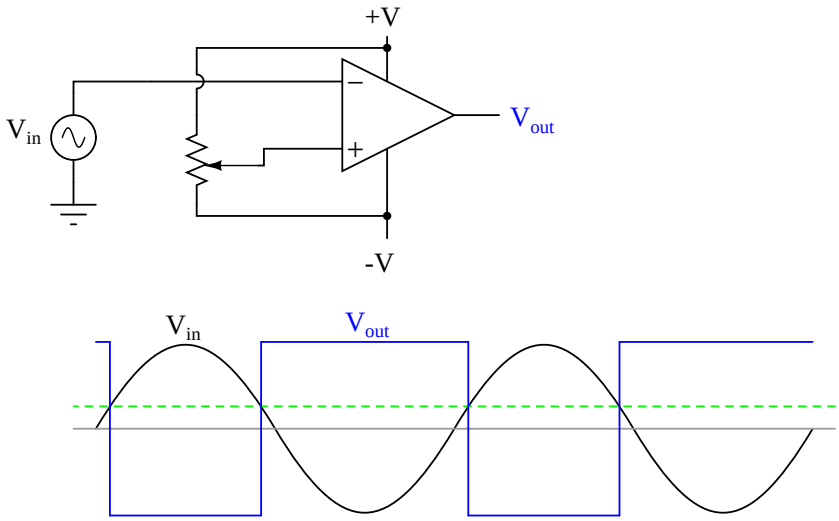
them almost useless as an analog differential amplifier by themselves. For an op-amp with a voltage gain (A_V) of 200,000 and a maximum output voltage swing of +15V/-15V, all it would take is a differential input voltage of 75 μV (microvolts) to drive it to saturation or cutoff! Before we take a look 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.

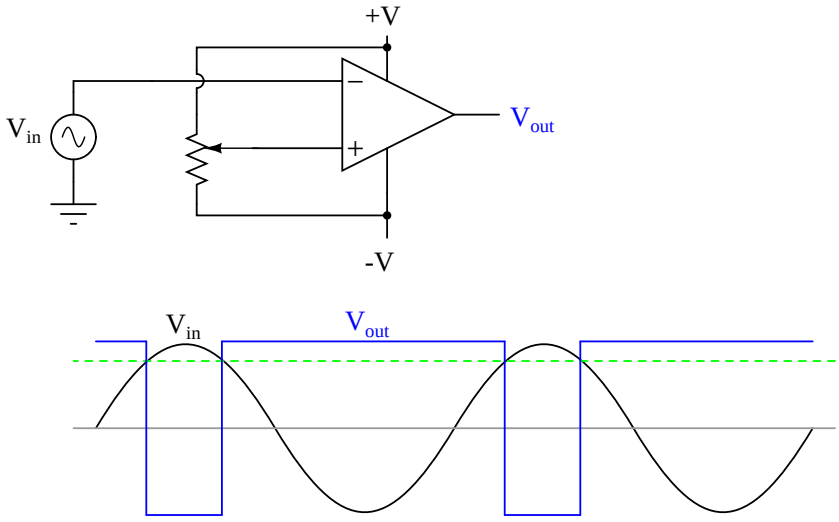


In the above circuit, we have an op-amp connected as a comparator, comparing the input voltage with a reference voltage set by the potentiometer (R_1). If V_{in} drops below the voltage set by R_1, the op-amp's output will saturate to +V, thereby lighting up the LED. Otherwise, if V_{in} is above the 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:



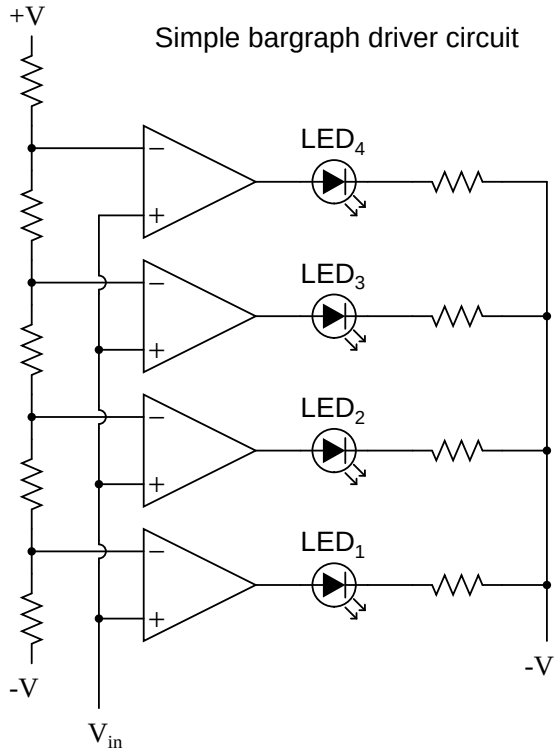
Adjustments to the potentiometer setting would change the reference voltage applied to the noninverting (+) input, which would change the points at which the sine wave would cross, changing the on/off times, or *duty cycle* of the square wave:



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.

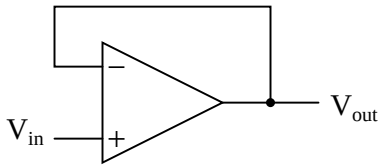
This very same technology is used in some analog-to-digital signal converters, namely the *flash converter*, to translate an analog signal quantity into a series of on/off voltages representing a digital number.

- **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.

8.4 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):



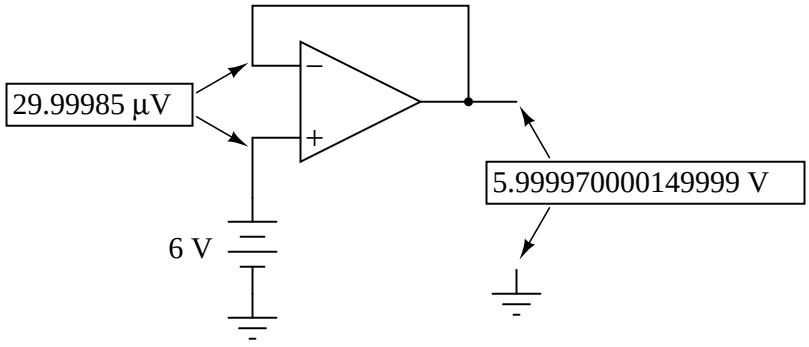
As V_{in} increases, V_{out} will increase in accordance with the differential gain. However, as V_{out} increases, that output voltage is fed back to the inverting input, thereby acting to decrease the voltage differential between inputs, which acts to bring the output down. What will happen for any 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.

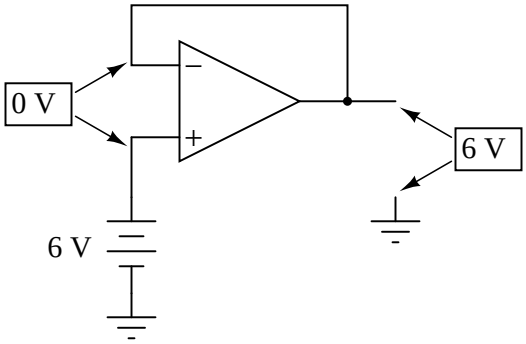
Because the op-amp's gain is so high, the voltage on the inverting input can be maintained almost equal to V_{in}. Let's say that our op-amp has a differential voltage gain of 200,000. If V_{in} equals

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.

The effects of negative feedback



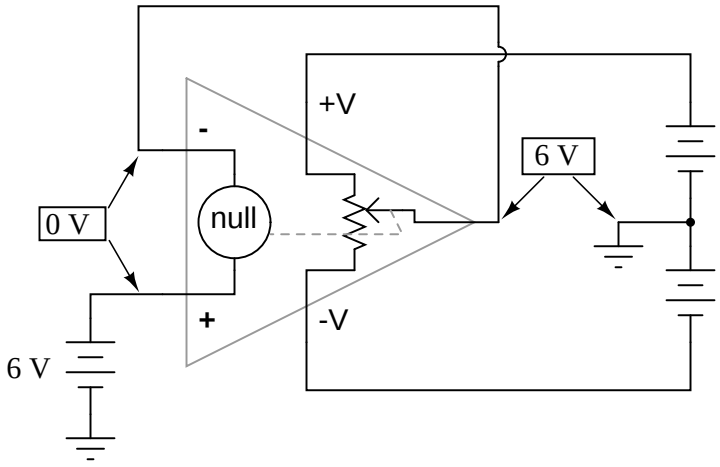
The effects of negative feedback (rounded figures)



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.

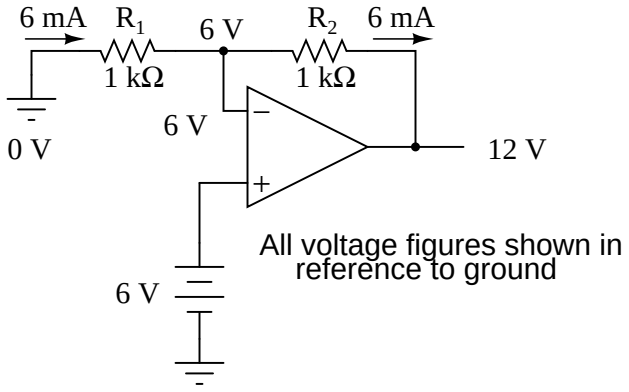
- **REVIEW:**

- 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.

8.5 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

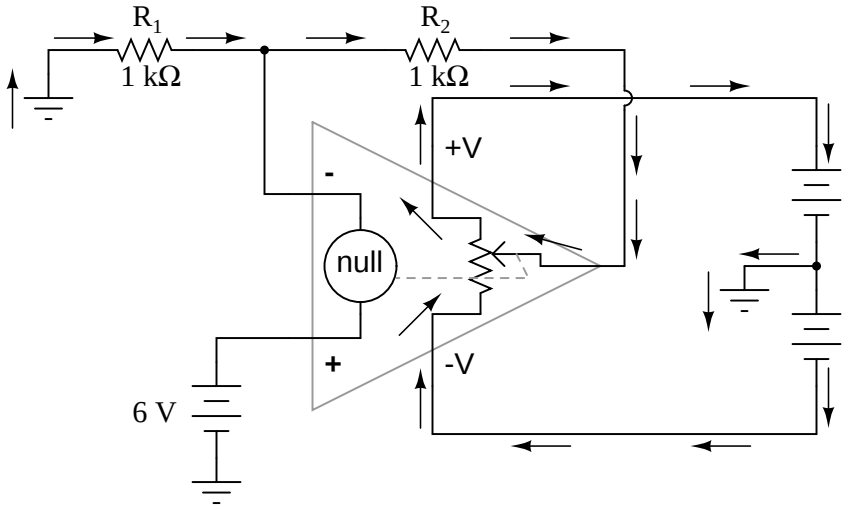


If R_1 and R_2 are both equal and V_{in} is 6 volts, the op-amp will output whatever voltage is needed to drop 6 volts across R_1 (to make the inverting input voltage equal to 6 volts, as well, keeping the voltage difference between the two inputs equal to zero). With the 2:1 voltage divider of R_1 and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

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

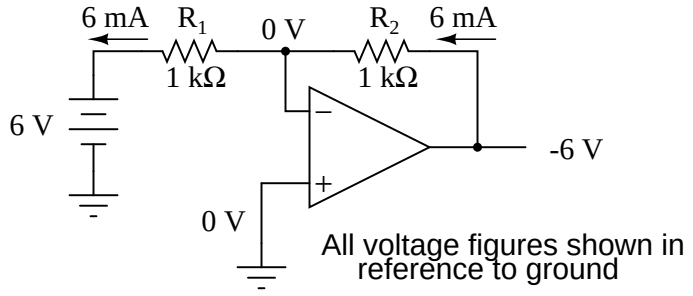
Note that the voltage gain for this design of amplifier circuit can never be less than 1. If we were to lower R_2 to a value of zero ohms, our circuit would be essentially identical to the voltage follower, with the output directly connected to the inverting input. Since the voltage follower has a gain 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

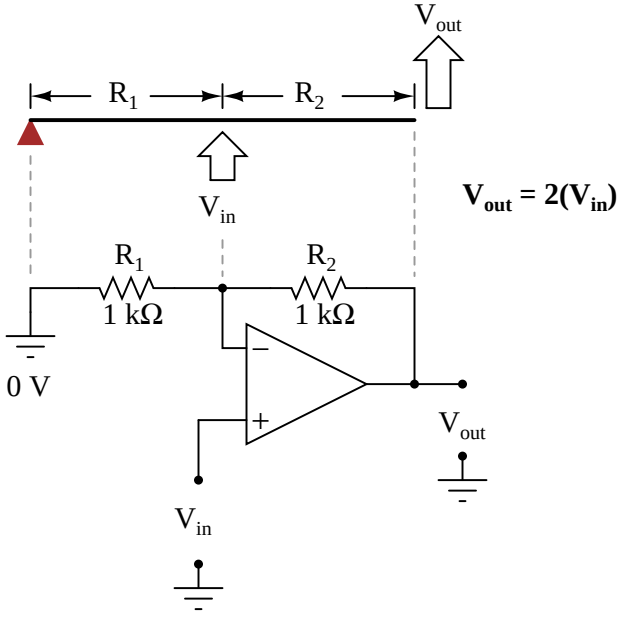
through a resistive voltage divider (feeding back a *fraction* of the output voltage to the inverting input), the output voltage becomes a *multiple* of the input voltage.

- 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: $A_V = R_2/R_1$

8.6 An analogy for divided feedback

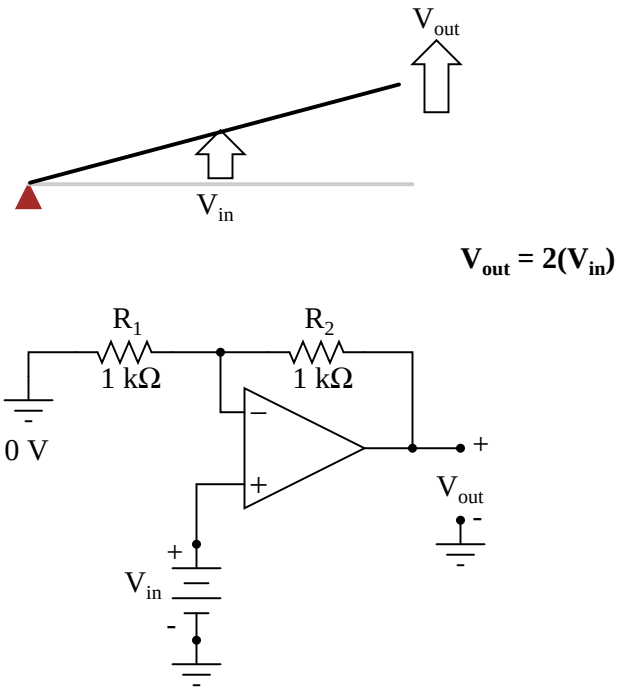
A helpful analogy for understanding divided feedback amplifier circuits is that of a mechanical lever, with relative motion of the lever's ends representing change in input and output voltages, and the fulcrum (pivot point) representing the location of the ground point, real or virtual.

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:

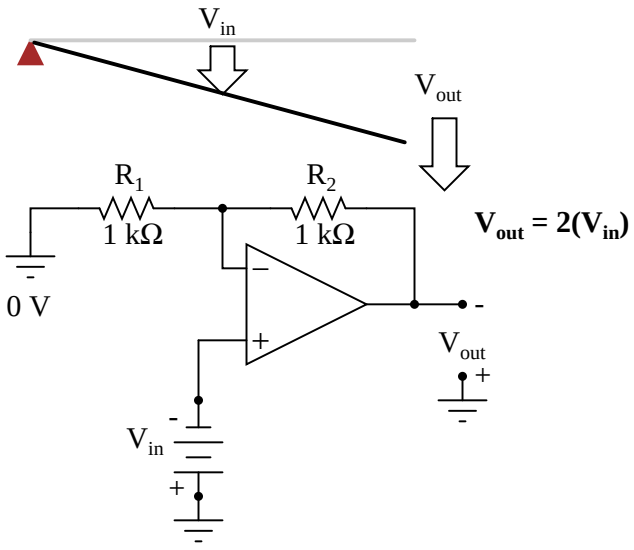


Physicists call this type of lever, with the input force (effort) applied between the fulcrum and output (load), a *third-class* lever. It is characterized by an output displacement (motion) at least as large than the input displacement – a "gain" of at least 1 – and in the same direction. Applying

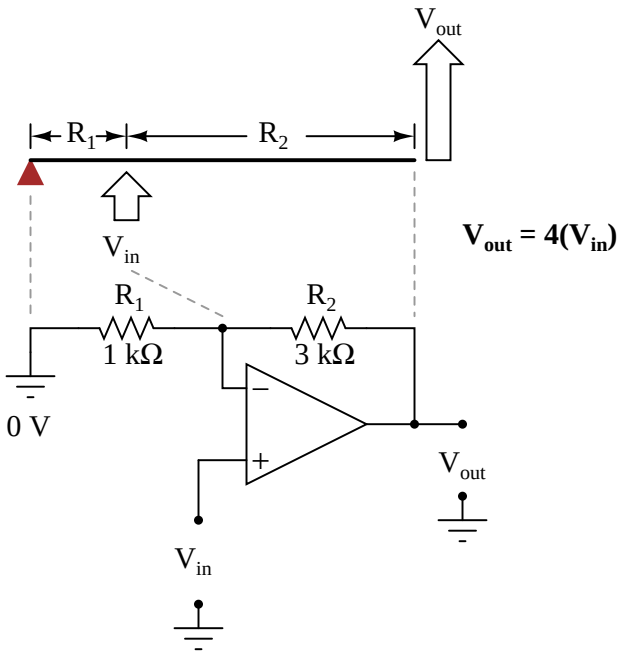
a positive input voltage to this op-amp circuit is analogous to displacing the "input" point on the lever upward:



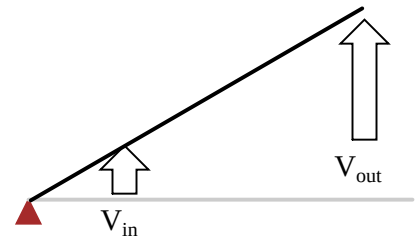
Due to the displacement-amplifying characteristics of the lever, the "output" point will move twice as far as the "input" point, and in the same direction. In the electronic circuit, the output voltage will equal twice the input, with the same polarity. Applying a negative input voltage is analogous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



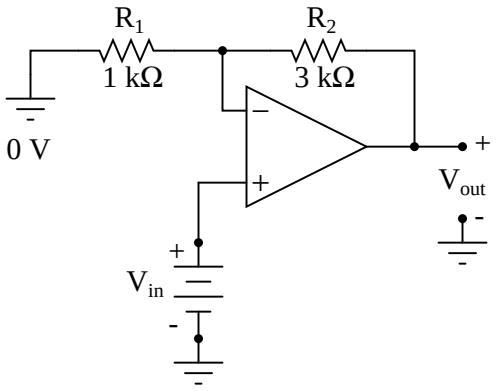
If we alter the resistor ratio R_2/R_1, we change the gain of the op-amp circuit. In lever terms, this means moving the input point in relation to the fulcrum and lever end, which similarly changes the displacement "gain" of the machine:



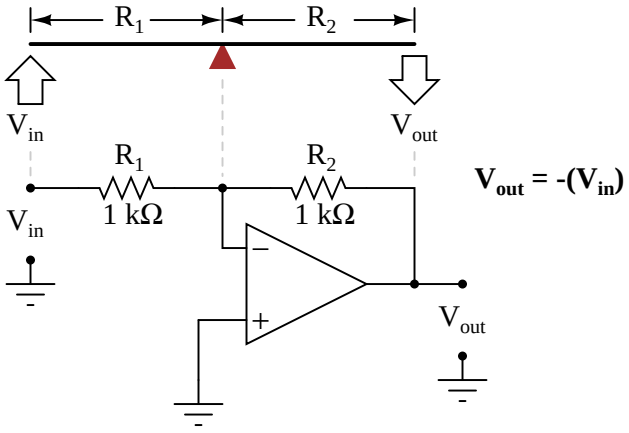
Now, any input signal will become amplified by a factor of four instead of by a factor of two:



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

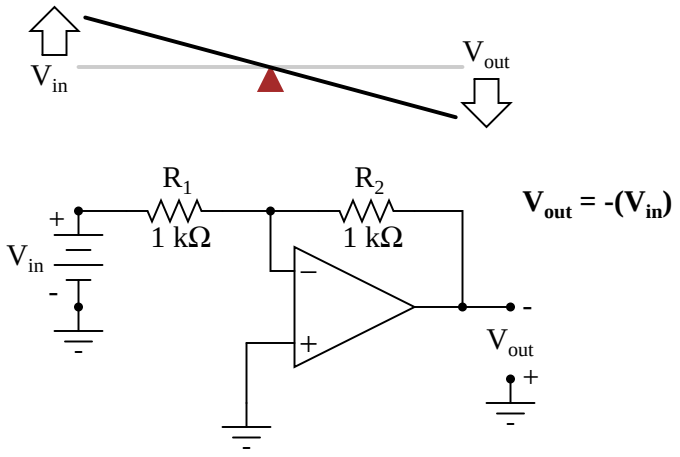


Inverting op-amp circuits may be modeled using the lever analogy as well. With the inverting configuration, the ground point of the feedback voltage divider is the op-amp's inverting input with the input to the left and the output to the right. This is mechanically equivalent to a *first-class* lever,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

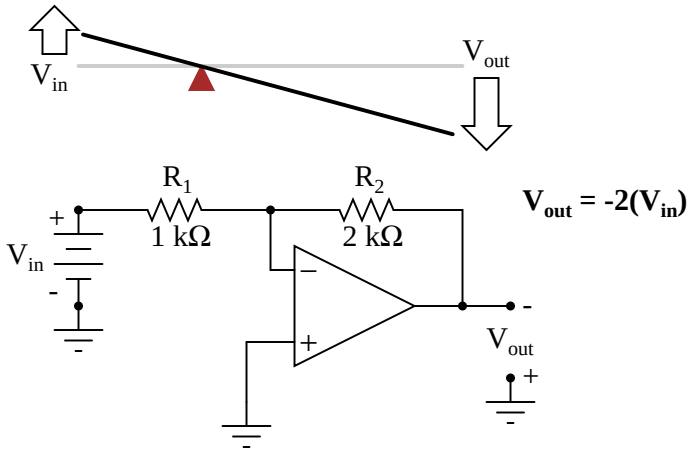


$$\mathbf{V_{out} = -(V_{in})}$$

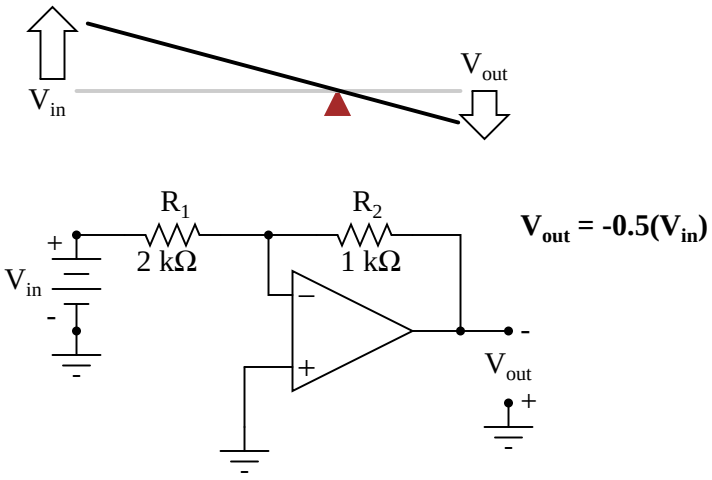
With equal-value resistors (equal-lengths of lever on each side of the fulcrum), the output voltage (displacement) will be equal in magnitude to the input voltage (displacement), but of the opposite polarity (direction). A positive input results in a negative output:



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:



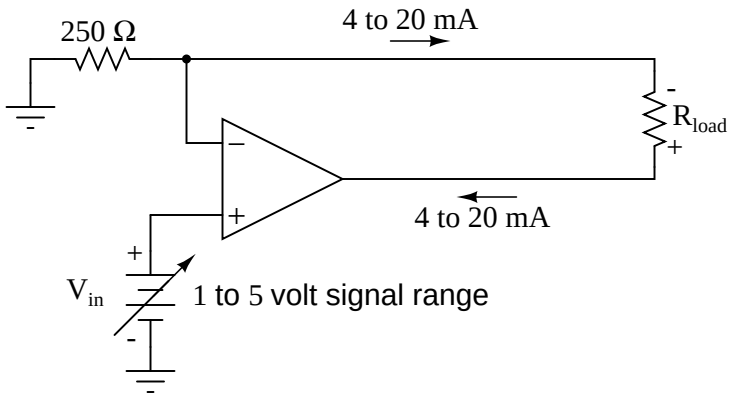
With the inverting amplifier configuration, though, gains of less than 1 are possible, just as with first-class levers. Reversing R_2 and R_1 values is analogous to moving the fulcrum to its complementary position on the lever: one-third of the way from the output end. There, the output displacement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



8.7 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.

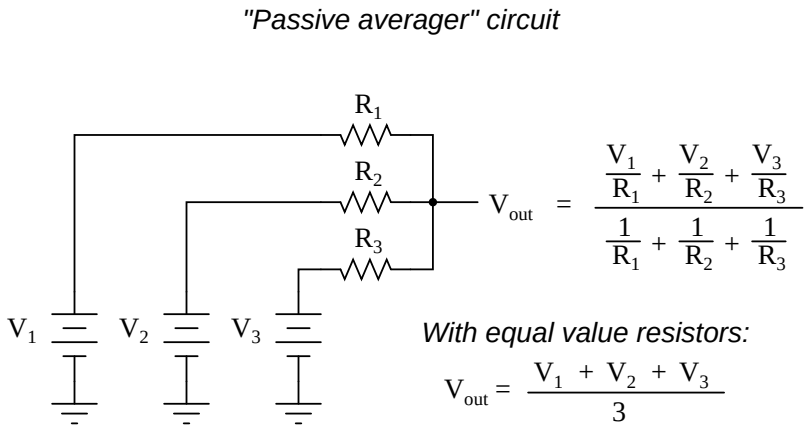
Another name for this circuit is *transconductance amplifier*. In electronics, transconductance is the mathematical ratio of current change divided by voltage change (ΔI / Δ V), and it is measured in the unit of Siemens, the same unit used to express conductance (the mathematical reciprocal of resistance: 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.

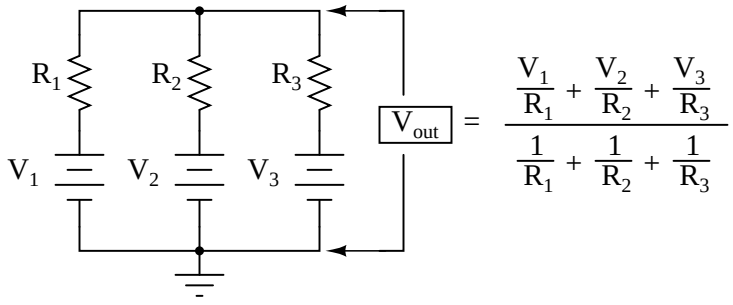
8.8 Averager and summer circuits

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.



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



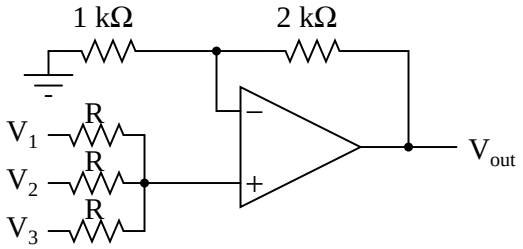
This circuit is commonly known as a *passive averager*, because it generates an average voltage with non-amplifying components. *Passive* simply means that it is an unamplified circuit. The large equation to the right of the averager circuit comes from Millman's Theorem, which describes the voltage produced 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:

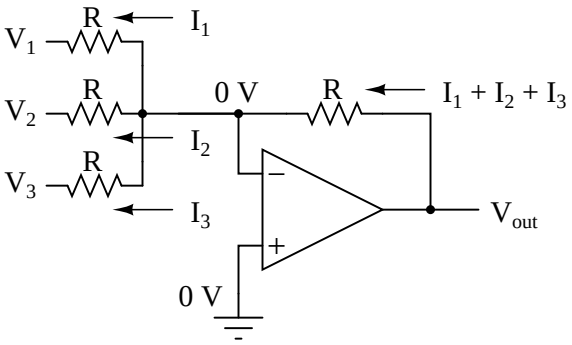


With a voltage divider composed of a 2 kΩ / 1 kΩ combination, the noninverting amplifier circuit will have a voltage gain of 3. By taking the voltage from the passive averager, which is the sum of V_1, V_2, and V_3 divided by 3, and multiplying that average by 3, we arrive at an output voltage equal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

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

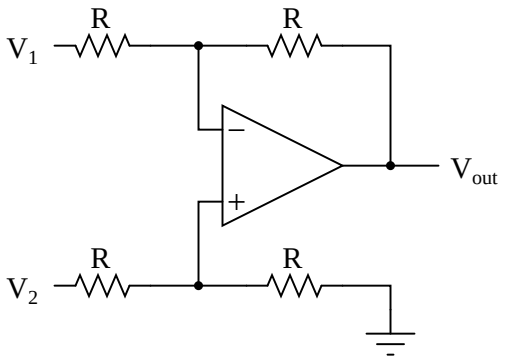
Summer (adder) circuits are quite useful in analog computer design, just as multiplier and divider circuits would be. Again, it is the extremely high differential gain of the op-amp which allows us to build these useful circuits with a bare minimum of components.

- **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.

8.9 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.

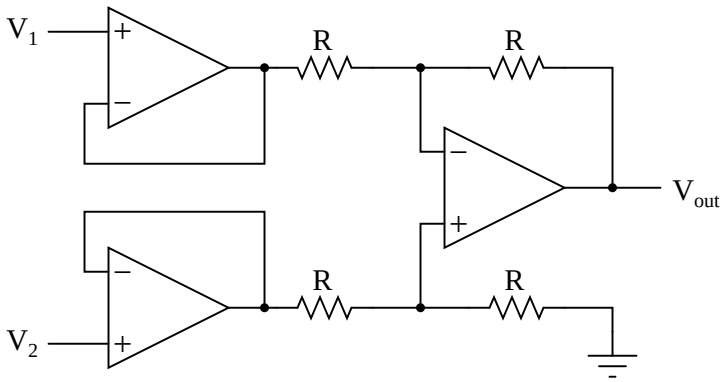


If all the resistor values are equal, this amplifier will have a differential voltage gain of 1. The analysis of this circuit is essentially the same as that of an inverting amplifier, except that the noninverting input (+) of the op-amp is at a voltage equal to a fraction of V_2, rather than being connected 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.

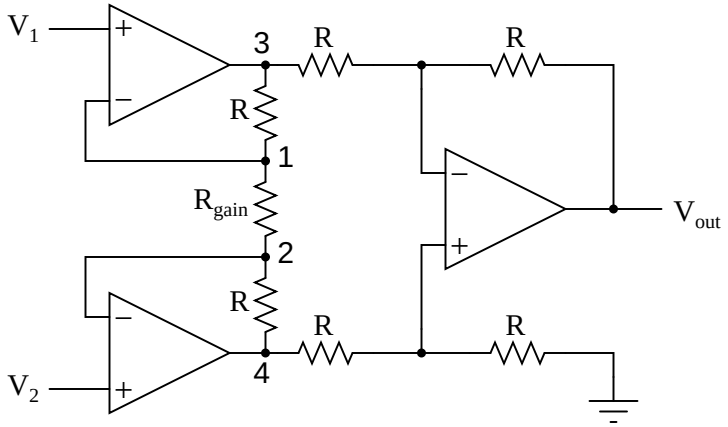
Another limitation of this amplifier design is the fact that its input impedances are rather low compared to that of some other op-amp configurations, most notably the noninverting (single-ended input) amplifier. Each input voltage source has to drive current through a resistance, which constitutes far 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.

8.10 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:

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

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.

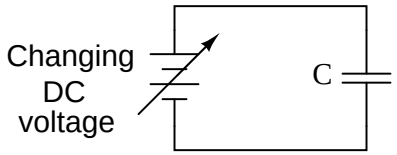
- **REVIEW:**

- 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.

8.11 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:



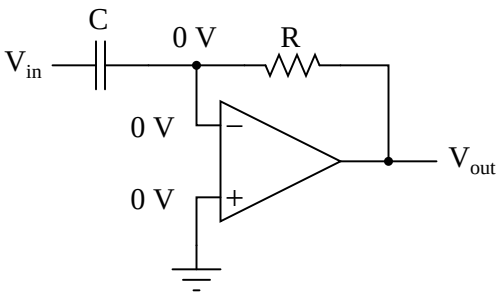
$$i = C \frac{dv}{dt}$$

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:

Differentiator



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:

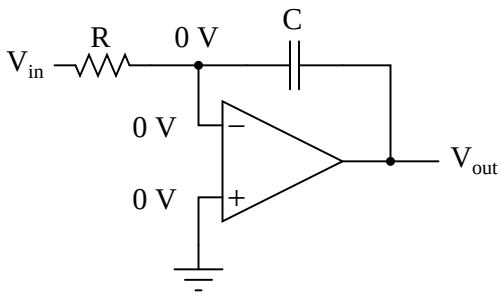
$$V_{out} = -RC \frac{dv_{in}}{dt}$$

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:

Integrator



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.

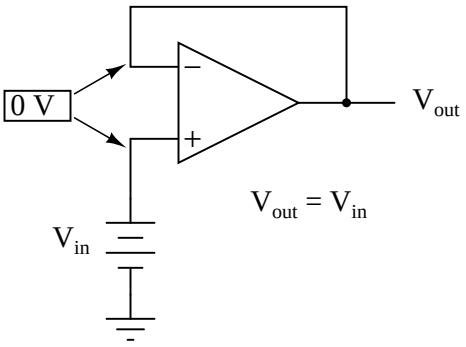
8.12 Positive feedback

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:

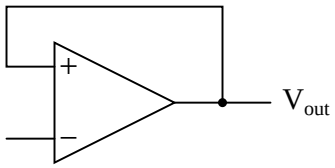
Negative feedback



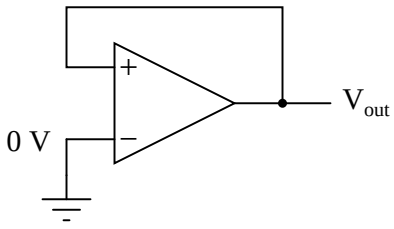
Whether the output is directly fed back to the inverting (-) input or coupled through a set of components, the effect is the same: the extremely high differential voltage gain of the op-amp will be "tamed" and the circuit will respond according to the dictates of the feedback "loop" connecting output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

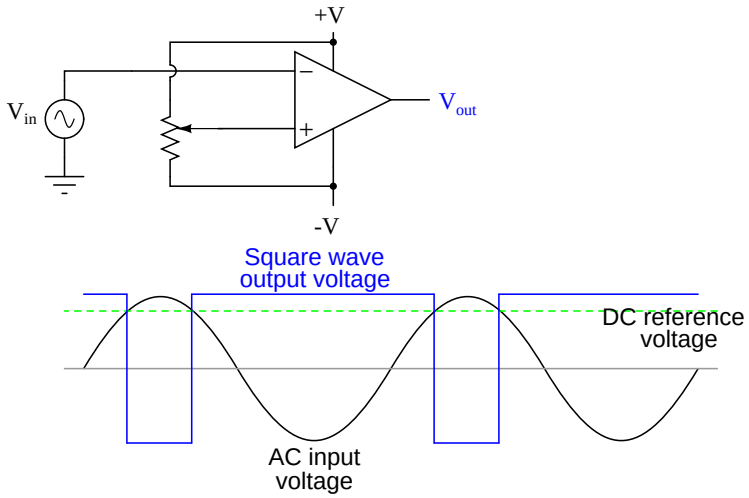


With the inverting input grounded (maintained at zero volts), the output voltage will be dictated by the magnitude and polarity of the voltage at the noninverting input. If that voltage happens to be positive, the op-amp will drive its output positive as well, feeding that positive voltage back to the 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.

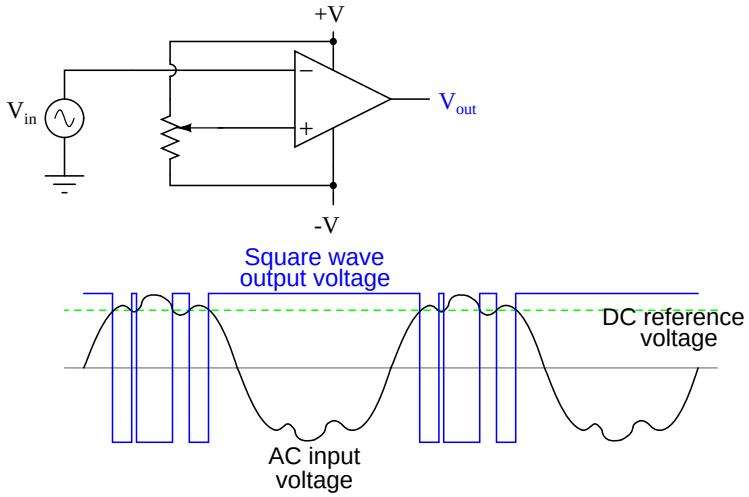
So, an op-amp with positive feedback tends to stay in whatever output state it's already in. It "latches" between one of two states, saturated positive or saturated negative. Technically, this is known as *hysteresis*.

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.



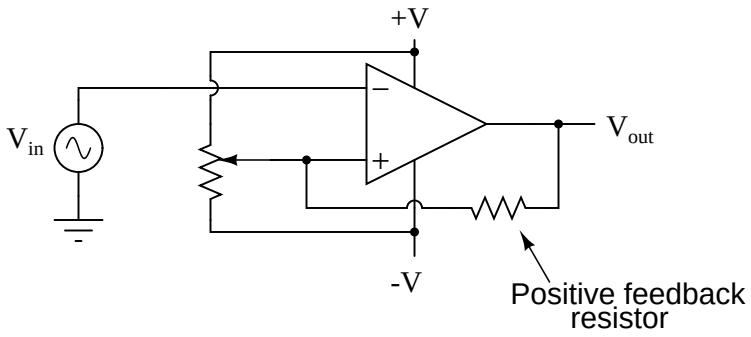
A "clean" AC input waveform produces predictable transition points on the output voltage square wave

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:

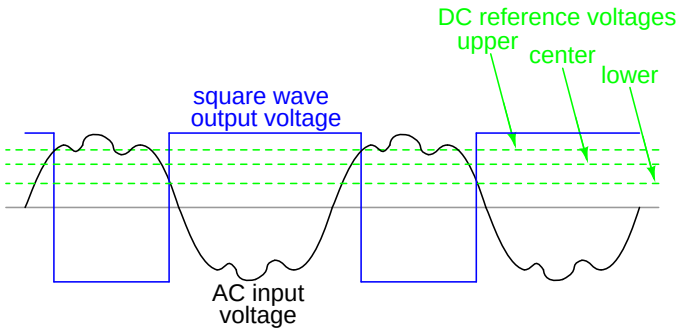


Any time there is a transition through the reference voltage level, no matter how tiny that transition may be, the output of the comparator will switch states, producing a square wave with "glitches."

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.



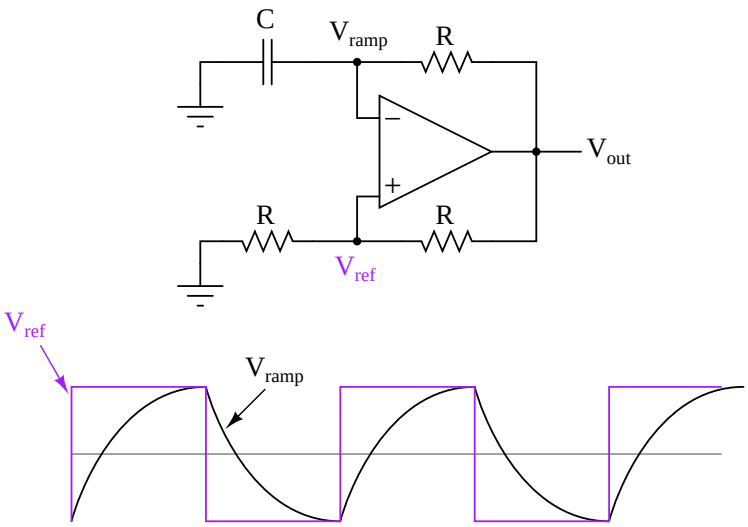
What this feedback resistor creates is a dual-reference for the comparator circuit. The voltage applied to the noninverting (+) input as a reference which to compare with the incoming AC voltage changes depending on the value of the op-amp's output voltage. When the op-amp output is saturated positive, 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



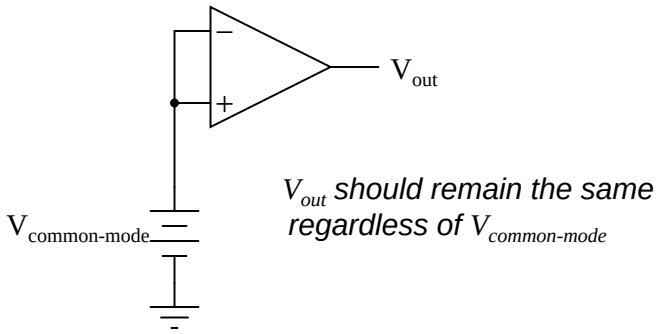
V_{out} is a square wave just like V_{ref}, only taller

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.

8.13 Practical considerations: common-mode gain

As stated before, an ideal differential amplifier only amplifies the voltage *difference* between its two inputs. If the two inputs of a differential amplifier were to be shorted together (thus ensuring zero potential difference between them), there should be no change in output voltage for any amount of voltage applied between those two shorted inputs and ground:



Voltage that is common between either of the inputs and ground, as "$V_{common-mode}$" is in this case, is called *common-mode voltage*. As we vary this common voltage, the perfect differential amplifier's output voltage should hold absolutely steady (no change in output for any arbitrary change in common-mode input). This translates to a *common-mode voltage gain* of zero.

$$A_V = \frac{\text{Change in } V_{out}}{\text{Change in } V_{in}}$$

. . . if change in V_{out} = 0 . . .

$$\frac{0}{\text{Change in } V_{in}} = 0$$

$$A_V = 0$$

The operational amplifier, being a differential amplifier with high differential gain, would ideally have zero common-mode gain as well. In real life, however, this is not easily attained. Thus, common-mode voltages will invariably have some effect on the op-amp's output voltage.

The performance of a real op-amp in this regard is most commonly measured in terms of its differential voltage gain (how much it amplifies the difference between two input voltages) versus its common-mode voltage gain (how much it amplifies a common-mode voltage). The ratio of the former to the latter is called the *common-mode rejection ratio*, abbreviated as CMRR:

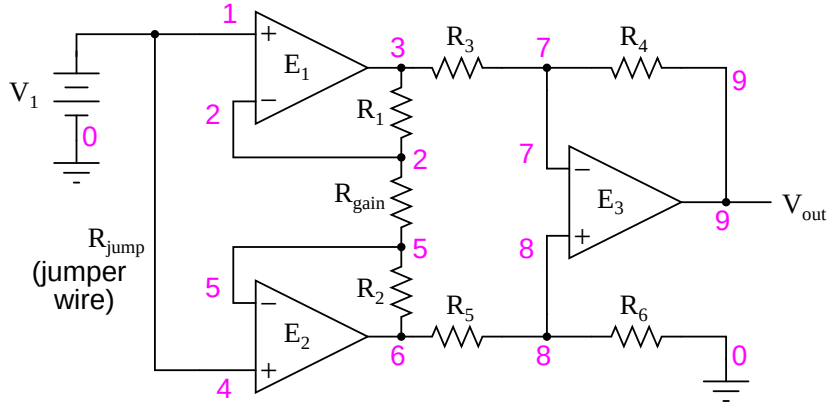
$$\text{CMRR} = \frac{\text{Differential } A_V}{\text{Common-mode } A_V}$$

An ideal op-amp, with zero common-mode gain would have an infinite CMRR. Real op-amps have high CMRRs, the ubiquitous 741 having something around 70 dB, which works out to a little over 3,000 in terms of a ratio.

Because the common mode rejection ratio in a typical op-amp is so high, common-mode gain is usually not a great concern in circuits where the op-amp is being used with negative feedback. If the common-mode input voltage of an amplifier circuit were to suddenly change, thus producing a corresponding change in the output due to common-mode gain, that change in output would be quickly corrected as negative feedback and differential gain (being *much* greater than common-mode

gain) worked to bring the system back to equilibrium. Sure enough, a change might be seen at the output, but it would be a lot smaller than what you might expect.

A consideration to keep in mind, though, is common-mode gain in differential op-amp circuits such as instrumentation amplifiers. Outside of the op-amp's sealed package and extremely high differential gain, we may find common-mode gain introduced by an imbalance of resistor values. To demonstrate this, we'll run a SPICE analysis on an instrumentation amplifier with inputs shorted together (no differential voltage), imposing a common-mode voltage to see what happens. First, we'll run the analysis showing the output voltage of a perfectly balanced circuit. We should expect to see no change in output voltage as the common-mode voltage changes:




```
instrumentation amplifier
v1 1 0
rin1 1 0 9e12
rjump 1 4 1e-12
rin2 4 0 9e12
e1 3 0 1 2 999k
e2 6 0 4 5 999k
e3 9 0 8 7 999k
rload 9 0 10k
r1 2 3 10k
rgain 2 5 10k
r2 5 6 10k
r3 3 7 10k
r4 7 9 10k
r5 6 8 10k
r6 8 0 10k
.dc v1 0 10 1
.print dc v(9)
.end

v1             v(9)
0.000E+00      0.000E+00
```

```
1.000E+00     1.355E-16
2.000E+00     2.710E-16
3.000E+00     0.000E+00   As you can see, the output voltage v(9)
4.000E+00     5.421E-16   hardly changes at all for a common-mode
5.000E+00     0.000E+00   input voltage (v1) that sweeps from 0
6.000E+00     0.000E+00   to 10 volts.
7.000E+00     0.000E+00
8.000E+00     1.084E-15
9.000E+00    -1.084E-15
1.000E+01     0.000E+00
```

Aside from very small deviations (actually due to quirks of SPICE rather than real behavior of the circuit), the output remains stable where it should be: at 0 volts, with zero input voltage differential. However, let's introduce a resistor imbalance in the circuit, increasing the value of R_5 from 10,000 Ω to 10,500 Ω, and see what happens (the netlist has been omitted for brevity – the only thing altered is the value of R_5):

```
v1            v(9)
0.000E+00     0.000E+00
1.000E+00    -2.439E-02
2.000E+00    -4.878E-02
3.000E+00    -7.317E-02   This time we see a significant variation
4.000E+00    -9.756E-02   (from 0 to 0.2439 volts) in output voltage
5.000E+00    -1.220E-01   as the common-mode input voltage sweeps
6.000E+00    -1.463E-01   from 0 to 10 volts as it did before.
7.000E+00    -1.707E-01
8.000E+00    -1.951E-01
9.000E+00    -2.195E-01
1.000E+01    -2.439E-01
```

Our input voltage differential is still zero volts, yet the output voltage changes significantly as the common-mode voltage is changed. This is indicative of a common-mode gain, something we're trying to avoid. More than that, it's a common-mode gain of our own making, having nothing to do with imperfections in the op-amps themselves. With a much-tempered differential gain (actually equal to 3 in this particular circuit) and no negative feedback outside the circuit, this common-mode gain will go unchecked in an instrument signal application.

There is only one way to correct this common-mode gain, and that is to balance all the resistor values. When designing an instrumentation amplifier from discrete components (rather than purchasing one in an integrated package), it is wise to provide some means of making fine adjustments to at least one of the four resistors connected to the final op-amp to be able to "trim away" any such common-mode gain. Providing the means to "trim" the resistor network has additional benefits as well. Suppose that all resistor values are exactly as they should be, but a common-mode gain exists due to an imperfection in one of the op-amps. With the adjustment provision, the resistance could be trimmed to compensate for this unwanted gain.

One quirk of some op-amp models is that of output *latch-up*, usually caused by the common-mode input voltage exceeding allowable limits. If the common-mode voltage falls outside of the

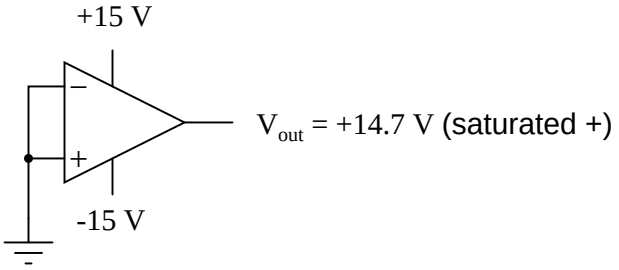
manufacturer's specified limits, the output may suddenly "latch" in the high mode (saturate at full output voltage). In JFET-input operational amplifiers, latch-up may occur if the common-mode input voltage approaches too closely to the negative power supply rail voltage. On the TL082 op-amp, for example, this occurs when the common-mode input voltage comes within about 0.7 volts of the negative power supply rail voltage. Such a situation may easily occur in a single-supply circuit, where the negative power supply rail is ground (0 volts), and the input signal is free to swing to 0 volts.

Latch-up may also be triggered by the common-mode input voltage *exceeding* power supply rail voltages, negative or positive. As a rule, you should never allow either input voltage to rise above the positive power supply rail voltage, or sink below the negative power supply rail voltage, even if the op-amp in question is protected against latch-up (as are the 741 and 1458 op-amp models). At the very least, the op-amp's behavior may become unpredictable. At worst, the kind of latch-up triggered by input voltages exceeding power supply voltages may be destructive to the op-amp.

While this problem may seem easy to avoid, its possibility is more likely than you might think. Consider the case of an operational amplifier circuit during power-up. If the circuit receives full input signal voltage *before* its own power supply has had time enough to charge the filter capacitors, the common-mode input voltage may easily exceed the power supply rail voltages for a short time. If the op-amp receives signal voltage from a circuit supplied by a different power source, and its own power source fails, the signal voltage(s) may exceed the power supply rail voltages for an indefinite amount of time!

8.14 Practical considerations: offset voltage

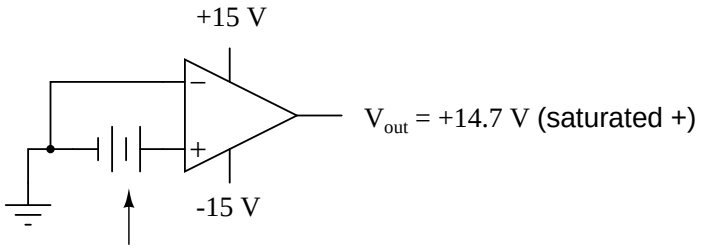
Another practical concern for op-amp performance is *voltage offset*. That is, effect of having the output voltage something other than zero volts when the two input terminals are shorted together. Remember that operational amplifiers are differential amplifiers above all: they're supposed to amplify the difference in voltage between the two input connections and nothing more. When that input voltage difference is exactly zero volts, we would (ideally) expect to have exactly zero volts present on the output. However, in the real world this rarely happens. Even if the op-amp in question has zero common-mode gain (infinite CMRR), the output voltage may not be at zero when both inputs are shorted together. This deviation from zero is called *offset*.



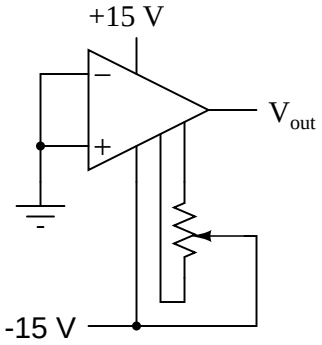
A perfect op-amp would output exactly zero volts with both its inputs shorted together and grounded. However, most op-amps off the shelf will drive their outputs to a saturated level, either negative or positive. In the example shown above, the output voltage is saturated at a value of positive 14.7 volts, just a bit less than +V (+15 volts) due to the positive saturation limit of this particular op-amp. Because the offset in this op-amp is driving the output to a completely saturated

point, there's no way of telling how much voltage offset is present at the output. If the +V/-V split power supply was of a high enough voltage, who knows, maybe the output would be several hundred volts one way or the other due to the effects of offset!

For this reason, offset voltage is usually expressed in terms of the equivalent amount of *input* voltage differential producing this effect. In other words, we imagine that the op-amp is perfect (no offset whatsoever), and a small voltage is being applied in series with one of the inputs to force the output voltage one way or the other away from zero. Being that op-amp differential gains are so high, the figure for "input offset voltage" doesn't have to be much to account for what we see with shorted inputs:



Offset voltage will tend to introduce slight errors in any op-amp circuit. So how do we compensate for it? Unlike common-mode gain, there are usually provisions made by the manufacturer to trim the offset of a packaged op-amp. Usually, two extra terminals on the op-amp package are reserved for connecting an external "trim" potentiometer. These connection points are labeled *offset null* and are used in this general way:



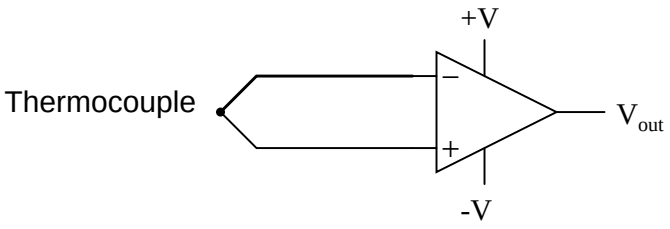
Potentiometer adjusted so that
V_{out} = 0 volts with inputs shorted together

On single op-amps such as the 741 and 3130, the offset null connection points are pins 1 and 5 on the 8-pin DIP package. Other models of op-amp may have the offset null connections located on different pins, and/or require a slightly difference configuration of trim potentiometer connection. Some op-amps don't provide offset null pins at all! Consult the manufacturer's specifications for details.

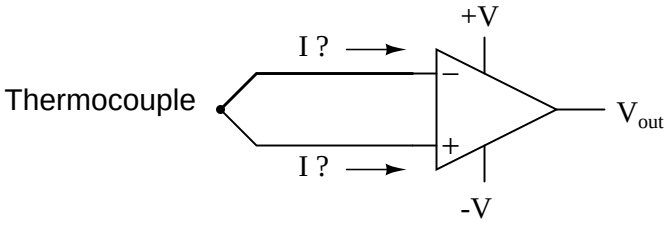
8.15 Practical considerations: bias current

Inputs on an op-amp have extremely high input impedances. That is, the input currents entering or exiting an op-amp's two input signal connections are extremely small. For most purposes of op-amp circuit analysis, we treat them as though they don't exist at all. We analyze the circuit as though there was absolutely zero current entering or exiting the input connections.

This idyllic picture, however, is not entirely true. Op-amps, especially those op-amps with bipolar transistor inputs, have to have some amount of current through their input connections in order for their internal circuits to be properly biased. These currents, logically, are called *bias currents*. Under certain conditions, op-amp bias currents may be problematic. The following circuit illustrates one of those problem conditions:

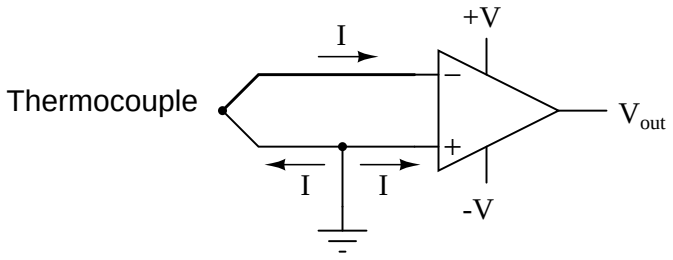


At first glance, we see no apparent problems with this circuit. A thermocouple, generating a small voltage proportional to temperature (actually, a voltage proportional to the *difference* in temperature between the measurement junction and the "reference" junction formed when the alloy thermocouple wires connect with the copper wires leading to the op-amp) drives the op-amp either positive or negative. In other words, this is a kind of comparator circuit, comparing the temperature between the end thermocouple junction and the reference junction (near the op-amp). The problem is this: the wire loop formed by the thermocouple does not provide a path for both input bias currents, because both bias currents are trying to go the same way (either into the op-amp or out of it).



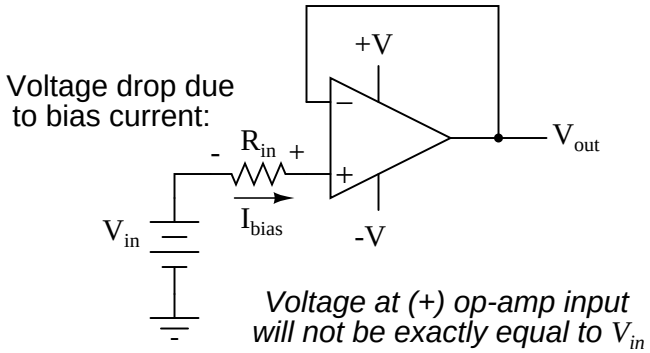
*This comparator circuit **won't** work*

In order for this circuit to work properly, we must ground one of the input wires, thus providing a path to (or from) ground for both currents:



Not necessarily an obvious problem, but a very real one!

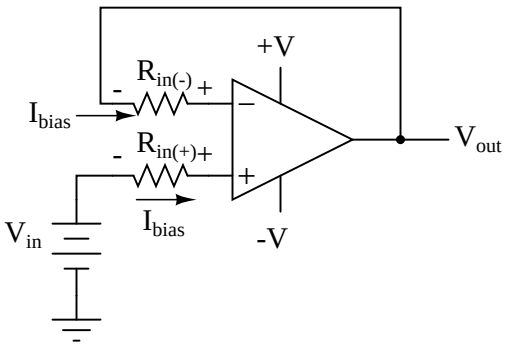
Another way input bias currents may cause trouble is by dropping unwanted voltages across circuit resistances. Take this circuit for example:



We expect a voltage follower circuit such as the one above to reproduce the input voltage precisely at the output. But what about the resistance in series with the input voltage source? If there is any bias current through the noninverting (+) input at all, it will drop some voltage across R_{in}, thus making the voltage at the noninverting input unequal to the actual V_{in} value. Bias currents are usually in the microamp range, so the voltage drop across R_{in} won't be very much, unless R_{in} is very large. One example of an application where the input resistance (R_{in}) *would* be very large is that of pH probe electrodes, where one electrode contains an ion-permeable glass barrier (a very poor conductor, with millions of Ω of resistance).

If we were actually building an op-amp circuit for pH electrode voltage measurement, we'd probably want to use a FET or MOSFET (IGFET) input op-amp instead of one built with bipolar transistors (for less input bias current). But even then, what slight bias currents may remain can cause measurement errors to occur, so we have to find some way to mitigate them through good design.

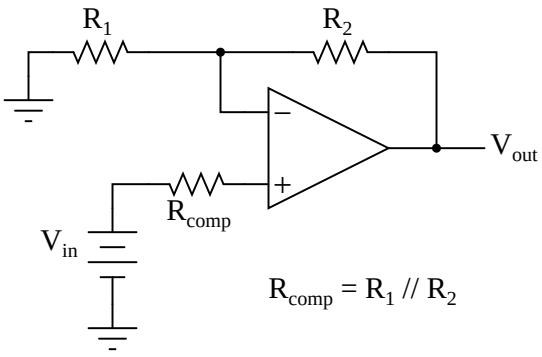
One way to do so is based on the assumption that the two input bias currents will be the same. In reality, they are often close to being the same, the difference between them referred to as the *input offset current*. If they are the same, then we should be able to cancel out the effects of input resistance voltage drop by inserting an equal amount of resistance in series with the other input, like this:



With the additional resistance added to the circuit, the output voltage will be closer to V_{in} than before, even if there is some offset between the two input currents.

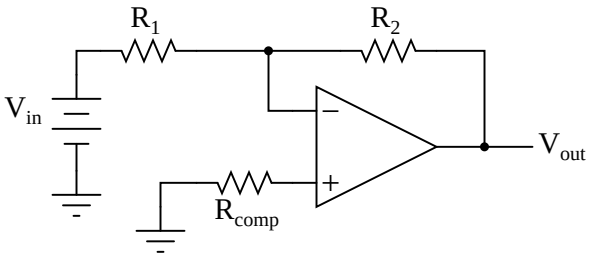
For both inverting and noninverting amplifier circuits, the bias current compensating resistor is placed in series with the noninverting (+) input to compensate for bias current voltage drops in the divider network:

Noninverting amplifier with compensating resistor



$$R_{comp} = R_1 \,//\, R_2$$

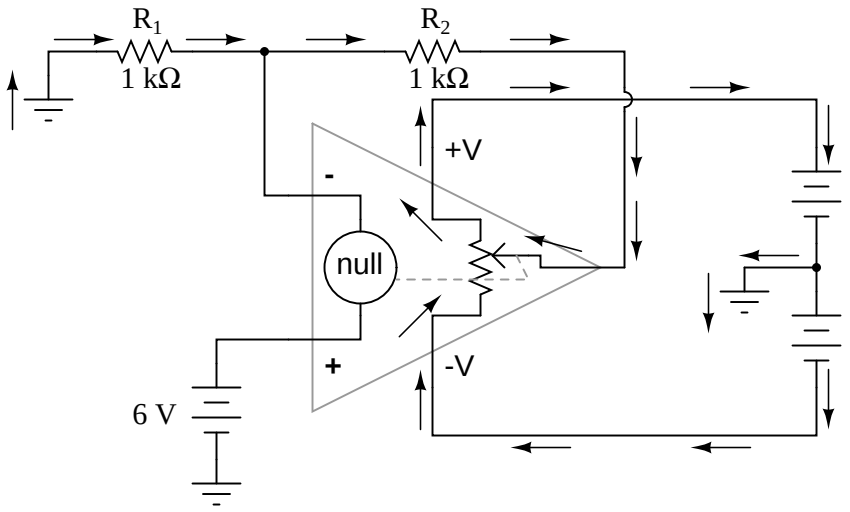
Inverting amplifier with compensating resistor



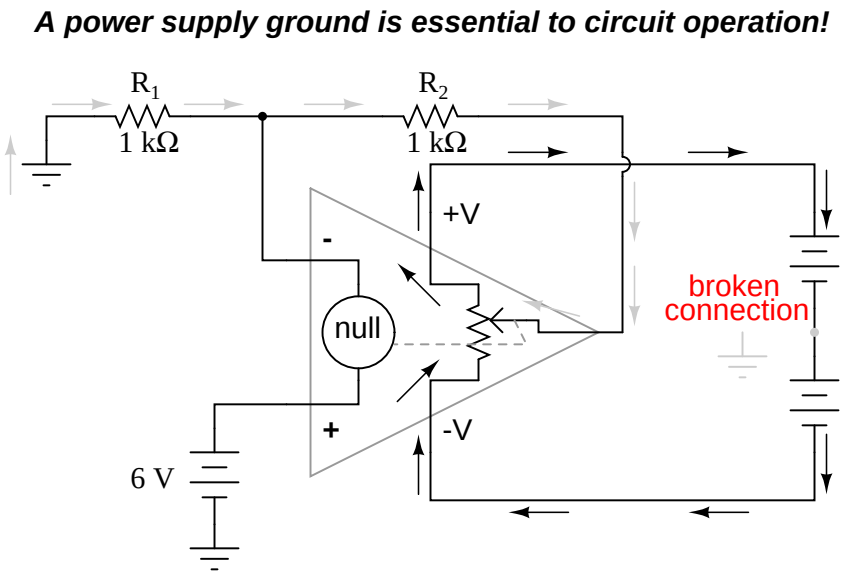
$$R_{comp} = R_1 \,//\, R_2$$

In either case, the compensating resistor value is determined by calculating the parallel resistance value of R_1 and R_2. Why is the value equal to the *parallel* equivalent of R_1 and R_2? When using the Superposition Theorem to figure how much voltage drop will be produced by the inverting (-) input's bias current, we treat the bias current as though it were coming from a current source inside the op-amp and short-circuit all voltage sources (V_{in} and V_{out}). This gives two parallel paths for bias current (through R_1 and through R_2, both to ground). We want to duplicate the bias current's effect on the noninverting (+) input, so the resistor value we choose to insert in series with that input needs to be equal to R_1 in parallel with R_2.

A related problem, occasionally experienced by students just learning to build operational amplifier circuits, is caused by a lack of a common ground connection to the power supply. It is *imperative* to proper op-amp function that some terminal of the DC power supply be common to the "ground" connection of the input signal(s). This provides a complete path for the bias currents, feedback current(s), and for the load (output) current. Take this circuit illustration, for instance, showing a properly grounded power supply:



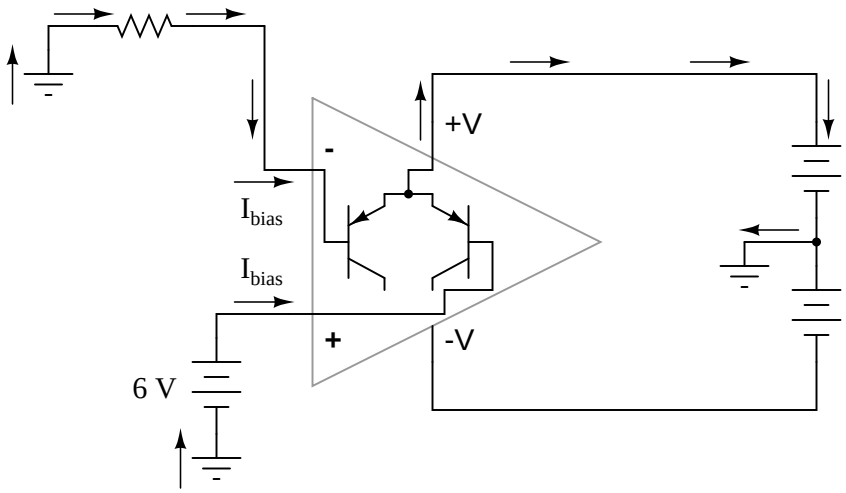
Here, arrows denote the path of electron flow through the power supply batteries, both for powering the op-amp's internal circuitry (the "potentiometer" inside of it that controls output voltage), and for powering the feedback loop of resistors R_1 and R_2. Suppose, however, that the ground connection for this "split" DC power supply were to be removed. The effect of doing this is profound:



No electrons may flow in or out of the op-amp's output terminal, because the pathway to the power supply is a "dead end." Thus, no electrons flow through the ground connection to the left of R_1, neither through the feedback loop. This effectively renders the op-amp useless: it can neither sustain current through the feedback loop, nor through a grounded load, since there is no connection from any point of the power supply to ground.

The bias currents are also stopped, because they rely on a path to the power supply and back to the input source through ground. The following diagram shows the bias currents (only), as they go through the input terminals of the op-amp, through the base terminals of the input transistors, and eventually through the power supply terminal(s) and back to ground.

Bias current paths shown, through power supply



Without a ground reference on the power supply, the bias currents will have no complete path for a circuit, and they will halt. Since bipolar junction transistors are current-controlled devices, this renders the input stage of the op-amp useless as well, as both input transistors will be forced into cutoff by the complete lack of base current.

- **REVIEW:**
- Op-amp inputs usually conduct very small currents, called *bias currents*, needed to properly bias the first transistor amplifier stage internal to the op-amps' circuitry. Bias currents are small (in the microamp range), but large enough to cause problems in some applications.
- Bias currents in both inputs *must* have paths to flow to either one of the power supply "rails" or to ground. It is not enough to just have a conductive path from one input to the other.
- To cancel any offset voltages caused by bias current flowing through resistances, just add an equivalent resistance in series with the other op-amp input (called a *compensating resistor*). This corrective measure is based on the assumption that the two input bias currents will be equal.
- Any inequality between bias currents in an op-amp constitutes what is called an *input offset current*.
- It is essential for proper op-amp operation that there be a ground reference on some terminal of the power supply, to form complete paths for bias currents, feedback current(s), and load current.

8.16 Practical considerations: drift

Being semiconductor devices, op-amps are subject to slight changes in behavior with changes in operating temperature. Any changes in op-amp performance with temperature fall under the category

of op-amp *drift*. Drift parameters can be specified for bias currents, offset voltage, and the like. Consult the manufacturer's data sheet for specifics on any particular op-amp.

To minimize op-amp drift, we can select an op-amp made to have minimum drift, and/or we can do our best to keep the operating temperature as stable as possible. The latter action may involve providing some form of temperature control for the inside of the equipment housing the op-amp(s). This is not as strange as it may first seem. Laboratory-standard precision voltage reference generators, for example, are sometimes known to employ "ovens" for keeping their sensitive components (such as zener diodes) at constant temperatures. If extremely high accuracy is desired over the usual factors of cost and flexibility, this may be an option worth looking at.

- **REVIEW:**

- Op-amps, being semiconductor devices, are susceptible to variations in temperature. Any variations in amplifier performance resulting from changes in temperature is known as *drift*. Drift is best minimized with environmental temperature control.

8.17 Practical considerations: frequency response

With their incredibly high differential voltage gains, op-amps are prime candidates for a phenomenon known as *feedback oscillation*. You've probably heard the equivalent audio effect when the volume (gain) on a public-address or other microphone amplifier system is turned too high: that high pitched squeal resulting from the sound waveform "feeding back" through the microphone to be amplified again. An op-amp circuit can manifest this same effect, with the feedback happening electrically rather than audibly.

A case example of this is seen in the 3130 op-amp, if it is connected as a voltage follower with the bare minimum of wiring connections (the two inputs, output, and the power supply connections). The output of this op-amp will self-oscillate due to its high gain, no matter what the input voltage. To combat this, a small *compensation capacitor* must be connected to two specially-provided terminals on the op-amp. The capacitor provides a high-impedance path for negative feedback to occur within the op-amp's circuitry, thus decreasing the AC gain and inhibiting unwanted oscillations. If the op-amp is being used to amplify high-frequency signals, this compensation capacitor may not be needed, but it is absolutely essential for DC or low-frequency AC signal operation.

Some op-amps, such as the model 741, have a compensation capacitor built in to minimize the need for external components. This improved simplicity is not without a cost: due to that capacitor's presence inside the op-amp, the negative feedback tends to get stronger as the operating frequency increases (that capacitor's reactance decreases with higher frequencies). As a result, the op-amp's differential voltage gain decreases as frequency goes up: it becomes a less effective amplifier at higher frequencies.

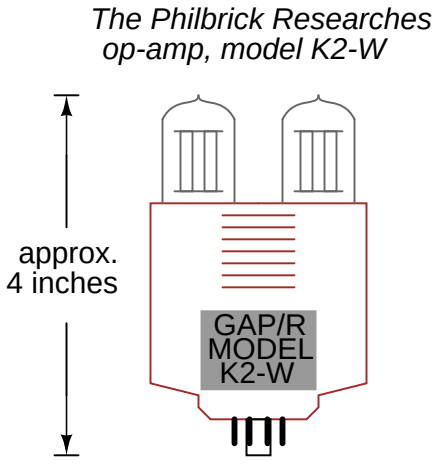
Op-amp manufacturers will publish the frequency response curves for their products. Since a sufficiently high differential gain is absolutely essential to good feedback operation in op-amp circuits, the gain/frequency response of an op-amp effectively limits its "bandwidth" of operation. The circuit designer must take this into account if good performance is to be maintained over the required range of signal frequencies.

- **REVIEW:**

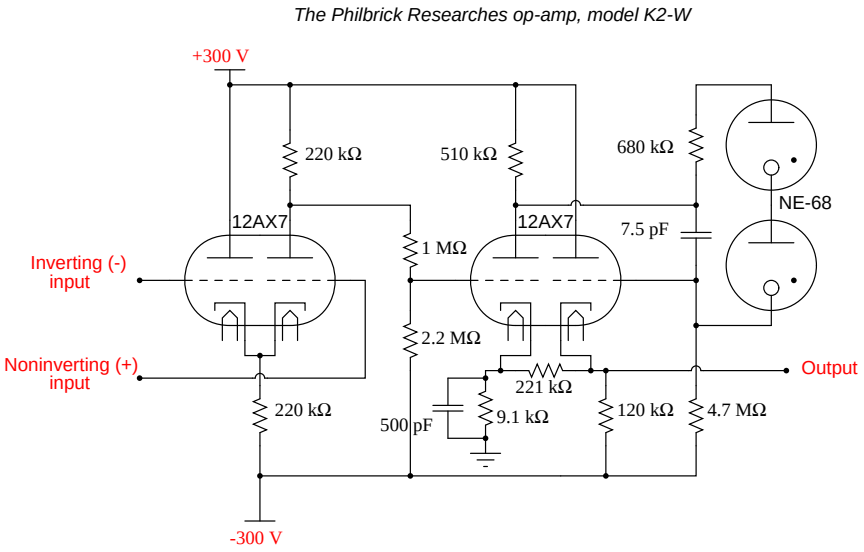
- Due to capacitances within op-amps, their differential voltage gain tends to decrease as the input frequency increases. Frequency response curves for op-amps are available from the manufacturer.

8.18 Operational amplifier models

While mention of operational amplifiers typically provokes visions of semiconductor devices built as integrated circuits on a miniature silicon chip, the first op-amps were actually vacuum tube circuits. The first commercial, general purpose operational amplifier was manufactured by the George A. Philbrick Researches, Incorporated, in 1952. Designated the K2-W, it was built around two twin-triode tubes mounted in an assembly with an octal (8-pin) socket for easy installation and servicing in electronic equipment chassis of that era. The assembly looked something like this:



The schematic diagram shows the two tubes, along with ten resistors and two capacitors, a fairly simple circuit design even by 1952 standards:



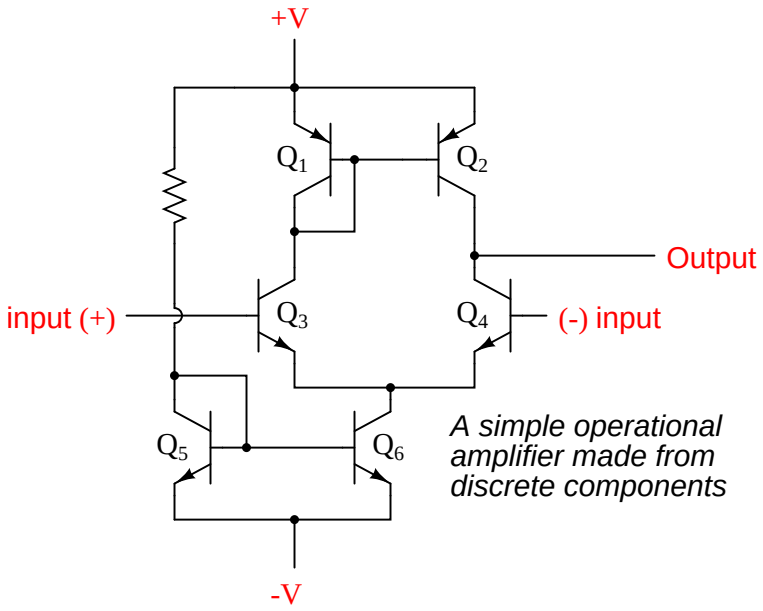
In case you're unfamiliar with the operation of vacuum tubes, they operate similarly to N-channel depletion-type IGFET transistors: that is, they conduct more current when the control grid (the dashed line) is made more positive with respect to the cathode (the bent line near the bottom of the tube symbol), and conduct less current when the control grid is made less positive (or more negative) than the cathode. The twin triode tube on the left functions as a *differential pair*, converting the differential inputs (inverting and noninverting input voltage signals) into a single, amplified voltage signal which is then fed to the control grid of the left triode of the second triode pair through a voltage divider (1 MΩ −− 2.2 MΩ). That triode amplifies and inverts the output of the differential pair for a larger voltage gain, then the amplified signal is coupled to the second triode of the same dual-triode tube in a noninverting amplifier configuration for a larger current gain. The two neon "glow tubes" act as voltage regulators, similar to the behavior of semiconductor zener diodes, to provide a bias voltage in the coupling between the two single-ended amplifier triodes.

With a dual-supply voltage of +300/-300 volts, this op-amp could only swing its output +/- 50 volts, which is very poor by today's standards. It had an open-loop voltage gain of 15,000 to 20,000, a slew rate of +/- 12 volts/μsecond, a maximum output current of 1 mA, a quiescent power consumption of over 3 watts (not including power for the tubes' filaments!), and cost about $24 in 1952 dollars. Better performance could have been attained using a more sophisticated circuit design, but only at the expense of greater power consumption, greater cost, and decreased reliability.

With the advent of solid-state transistors, op-amps with far less quiescent power consumption and increased reliability became feasible, but many of the other performance parameters remained about the same. Take for instance Philbrick's model P55A, a general-purpose solid-state op-amp circa 1966. The P55A sported an open-loop gain of 40,000, a slew rate of 1.5 volt/μsecond and an output swing of +/- 11 volts (at a power supply voltage of +/- 15 volts), a maximum output current of 2.2 mA, and a cost of $49 (or about $21 for the "utility grade" version). The P55A, as well as other op-amps in Philbrick's lineup of the time, was of discrete-component construction, its constituent transistors, resistors, and capacitors housed in a solid "brick" resembling a large integrated circuit package.

It isn't very difficult to build a crude operational amplifier using discrete components. A

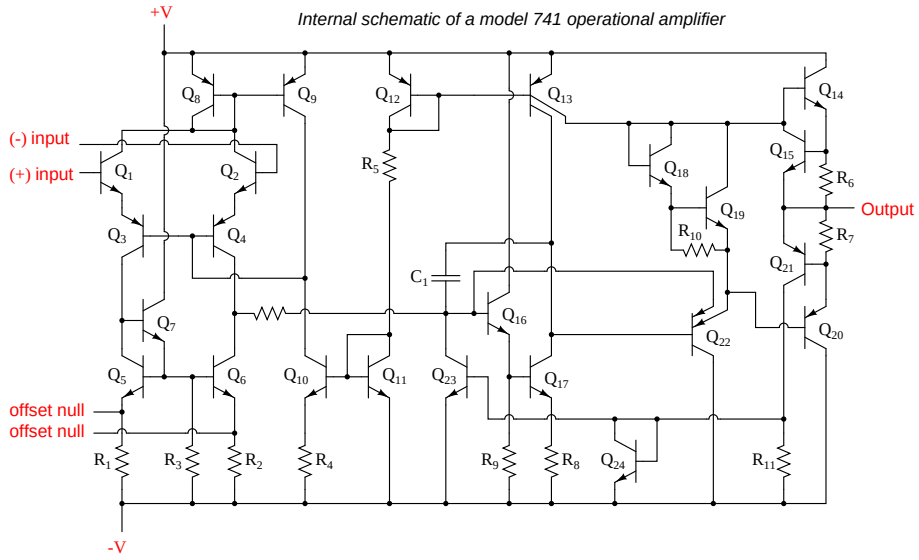
schematic of one such circuit is shown here:



While its performance is rather dismal by modern standards, it demonstrates that complexity is not necessary to create a minimally functional op-amp. Transistors Q_3 and Q_4 form the heart of another differential pair circuit, the semiconductor equivalent of the first triode tube in the K2-W schematic. As it was in the vacuum tube circuit, the purpose of a differential pair is to amplify and convert a differential voltage between the two input terminals to a single-ended output voltage.

With the advent of integrated-circuit (IC) technology, op-amp designs experienced a dramatic increase in performance, reliability, density, and economy. Between the years of 1964 and 1968, the Fairchild corporation introduced three models of IC op-amps: the 702, 709, and the still-popular 741. While the 741 is now considered outdated in terms of performance, it is still a favorite among hobbyists for its simplicity and fault tolerance (short-circuit protection on the output, for instance). Personal experience abusing many 741 op-amps has led me to the conclusion that it is a hard chip to kill . . .

The internal schematic diagram for a model 741 op-amp is as follows:



By integrated circuit standards, the 741 is a very simple device: an example of *small-scale integration*, or *SSI* technology. It would be no small matter to build this circuit using discrete components, so you can see the advantages of even the most primitive integrated circuit technology over discrete components where high parts counts are involved.

For the hobbyist, student, or engineer desiring greater performance, there are literally hundreds of op-amp models to choose from. Many sell for less than a dollar apiece, even retail! Special-purpose instrumentation and radio-frequency (RF) op-amps may be quite a bit more expensive. In this section I will showcase several popular and affordable op-amps, comparing and contrasting their performance specifications. The venerable 741 is included as a "benchmark" for comparison, although it is, as I said before, considered an obsolete design.

OPAMP MODEL NUMBER	NUMBER OF AMPLIFIERS IN PACKAGE	PWR SUPPLY VOLTAGE MIN./MAX.	BAND-WIDTH (MHz)	MAX. BIAS CURRENT (nA)	SLEW RATE (V/us)	MAX. OUT CURRENT (mA)
TL082	2	12 / 36	4	8	13	17
LM301A	1	10 / 36	1	250	0.5	25
LM318	1	10 / 40	15	500	70	20
LM324	4	3 / 32	1	45	0.25	20
LF353	2	12 / 36	4	8	13	20
LF356	1	10 / 36	5	8	12	25
LF411	1	10 / 36	4	20	15	25

OPAMP MODEL NUMBER	NUMBER OF AMPLIFIERS IN PACKAGE	PWR SUPPLY VOLTAGE MIN./MAX.	BAND-WIDTH (MHz)	MAX. BIAS CURRENT (nA)	SLEW RATE (V/us)	MAX. OUT CURRENT (mA)
LM741C	1	10 / 36	1	500	0.5	25
LM833	2	10 / 36	15	1050	7	40
LM1458	2	6 / 36	1	800	10	45
CA3130	1	5 / 16	15	0.05	10	20

These are but a few of the low-cost operational amplifier models widely available from electronics suppliers. Most of them are available through retail supply stores such as Radio Shack. All are under $1.00 cost direct from the manufacturer (year 2001 prices). As you can see, there is substantial variation in performance between some of these units. Take for instance the parameter of input bias current: the CA3130 wins the prize for lowest, at 0.05 nA (or 50 pA), and the LM833 has the highest at slightly over 1 μA. The model CA3130 achieves its incredibly low bias current through the use of MOSFET transistors in its input stage. One manufacturer advertises the 3130's input impedance as 1.5 tera-ohms, or 1.5 x 10^{12} Ω! Other op-amps shown here with low bias current figures use JFET input transistors, while the high bias current models use bipolar input transistors.

While the 741 is specified in many electronic project schematics and showcased in many textbooks, its performance has long been surpassed by other designs in every measure. Even some designs originally based on the 741 have been improved over the years to far surpass original design specifications. One such example is the model 1458, two op-amps in an 8-pin DIP package, which at one time had the exact same performance specifications as the single 741. In its latest incarnation it boasts a wider power supply voltage range, a slew rate 50 times as great, and almost twice the output current capability of a 741, while still retaining the output short-circuit protection feature of the 741. Op-amps with JFET and MOSFET input transistors *far* exceed the 741's performance in terms of bias current, and generally manage to beat the 741 in terms of bandwidth and slew rate as well.

My own personal recommendations for op-amps are as such: when low bias current is a priority (such as in low-speed integrator circuits), choose the 3130. For general-purpose DC amplifier work, the 1458 offers good performance (and you get two op-amps in the space of one package). For an upgrade in performance, choose the model 353, as it is a pin-compatible replacement for the 1458. The 353 is designed with JFET input circuitry for very low bias current, and has a bandwidth 4 times are great as the 1458, although its output current limit is lower (but still short-circuit protected). It may be more difficult to find on the shelf of your local electronics supply house, but it is just as reasonably priced as the 1458.

If low power supply voltage is a requirement, I recommend the model 324, as it functions on as low as 3 volts DC. Its input bias current requirements are also low, and it provides four op-amps in a single 14-pin chip. Its major weakness is speed, limited to 1 MHz bandwidth and an output slew rate of only 0.25 volts per μs. For high-frequency AC amplifier circuits, the 318 is a very good "general purpose" model.

Special-purpose op-amps are available for modest cost which provide better performance specifications. Many of these are tailored for a specific type of performance advantage, such as maximum

bandwidth or minimum bias current. Take for instance these op-amps, both designed for high bandwidth:

OPAMP MODEL NUMBER	NUMBER OF AMPLIFIERS IN PACKAGE	PWR SUPPLY VOLTAGE MIN./MAX.	BAND-WIDTH (MHz)	MAX. BIAS CURRENT (nA)	SLEW RATE (V/us)	MAX. OUT CURRENT (mA)
CLC404	1	10 / 14	232	44,000	2600	70
CLC425	1	5 / 14	1,900	40,000	350	90

The CLC404 lists at $21.80 (almost as much as George Philbrick's first commercial op-amp, albeit without correction for inflation), while the CLC425 is quite a bit less expensive at $3.23 per unit. In both cases high speed is achieved at the expense of high bias currents and restrictive power supply voltage ranges. Here are some other op-amps, designed for high power output:

OPAMP MODEL NUMBER	NUMBER OF AMPLIFIERS IN PACKAGE	PWR SUPPLY VOLTAGE MIN./MAX.	BAND-WIDTH (MHz)	MAX. BIAS CURRENT (nA)	SLEW RATE (V/us)	MAX. OUT CURRENT (mA)
LM12CL	1	15 / 80	0.7	1,000	9	13,000
LM7171	1	5.5 / 36	200	12,000	4100	100

Yes, the LM12CL actually has an output current rating of *13 amps* (13,000 milliamps)! It lists at $14.40, which is not a lot of money, considering the raw power of the device. The LM7171, on the other hand, trades high current output ability for fast voltage output ability (a high slew rate). It lists at $1.19, about as low as some "general purpose" op-amps.

Amplifier packages may also be purchased as complete application circuits as opposed to bare operational amplifiers. The Burr-Brown and Analog Devices corporations, for example, both long known for their precision amplifier product lines, offer instrumentation amplifiers in pre-designed packages as well as other specialized amplifier devices. In designs where high precision and repeatability after repair is important, it might be advantageous for the circuit designer to choose such a pre-engineered amplifier "block" rather than build the circuit from individual op-amps. Of course, these units typically cost quite a bit more than individual op-amps.

8.19 Data

Parametrical data for all semiconductor op-amp models *except* the CA3130 comes from National Semiconductor's online resources, available at this website: `(http://www.national.com)`. Data for the CA3130 comes from Harris Semiconductor's CA3130/CA3130A datasheet (file number 817.4).

Chapter 9

PRACTICAL ANALOG SEMICONDUCTOR CIRCUITS

Contents

***** INCOMPLETE *****

9.1 Power supply circuits – INCOMPLETE

There are three major kinds of power supplies: *unregulated* (also called *brute force*), *linear regulated*, and *switching*. A fourth type of power supply circuit called the *ripple-regulated*, is a hybrid between the "brute force" and "switching" designs, and merits a subsection to itself.

9.1.1 Unregulated

An unregulated power supply is the most rudimentary type, consisting of a transformer, rectifier, and low-pass filter. These power supplies typically exhibit a lot of ripple voltage (i.e. rapidly-varying

instability) and other AC "noise" superimposed on the DC power. If the input voltage varies, the output voltage will vary by a proportional amount. The advantage of an unregulated supply is that it's cheap, simple, and efficient.

9.1.2 Linear regulated

A linear regulated supply is simply a "brute force" (unregulated) power supply followed by a transistor circuit operating in its "active," or "linear" mode, hence the name *linear* regulator. (Obvious in retrospect, isn't it?) A typical linear regulator is designed to output a fixed voltage for a wide range of input voltages, and it simply drops any excess input voltage to allow a maximum output voltage to the load. This excess voltage drop results in significant power dissipation in the form of heat. If the input voltage gets too low, the transistor circuit will lose regulation, meaning that it will fail to keep the voltage steady. It can only drop excess voltage, not make up for a deficiency in voltage from the brute force section of the circuit. Therefore, you have to keep the input voltage at least 1 to 3 volts higher than the desired output, depending on the regulator type. This means the power equivalent of at *least* 1 to 3 volts multiplied by the full load current will be dissipated by the regulator circuit, generating a lot of heat. This makes linear regulated power supplies rather inefficient. Also, to get rid of all that heat they have to use large heat sinks which makes them large, heavy, and expensive.

9.1.3 Switching

A switching regulated power supply ("switcher") is an effort to realize the advantages of both brute force and linear regulated designs (small, efficient, and cheap, but also "clean," stable output voltage). Switching power supplies work on the principle of rectifying the incoming AC power line voltage into DC, re-converting it into high-frequency square-wave AC through transistors operated as on/off switches, stepping that AC voltage up or down by using a lightweight transformer, then rectifying the transformer's AC output into DC and filtering for final output. Voltage regulation is achieved by altering the "duty cycle" of the DC-to-AC inversion on the transformer's primary side. In addition to lighter weight because of a smaller transformer core, switchers have another tremendous advantage over the prior two designs: this type of power supply can be made so totally independent of the input voltage that it can work on any electric power system in the world; these are called "universal" power supplies.

The downside of switchers is that they are more complex, and due to their operation they tend to generate a lot of high-frequency AC "noise" on the power line. Most switchers also have significant ripple voltage on their outputs. With the cheaper types, this noise and ripple can be as bad as for an unregulated power supply; such low-end switchers aren't worthless, because they still provide a stable average output voltage, and there's the "universal" input capability.

Expensive switchers are ripple-free and have noise nearly as low as for some a linear types; these switchers tend to be as expensive as linear supplies. The reason to use an expensive switcher instead of a good linear is if you need universal power system compatibility or high efficiency. High efficiency, light weight, and small size are the reasons switching power supplies are almost universally used for powering digital computer circuitry.

9.1.4 Ripple regulated

A ripple-regulated power supply is an alternative to the linear regulated design scheme: a "brute force" power supply (transformer, rectifier, filter) constitutes the "front end" of the circuit, but a transistor operated strictly in its on/off (saturation/cutoff) modes transfers DC power to a large capacitor as needed to maintain the output voltage between a high and a low setpoint. As in switchers, the transistor in a ripple regulator never passes current while in its "active," or "linear," mode for any substantial length of time, meaning that very little energy will be wasted in the form of heat. However, the biggest drawback to this regulation scheme is the necessary presence of some ripple voltage on the output, as the DC voltage varies between the two voltage control setpoints. Also, this ripple voltage varies in frequency depending on load current, which makes final filtering of the DC power more difficult.

Ripple regulator circuits tend to be quite a bit simpler than switcher circuitry, and they need not handle the high power line voltages that switcher transistors must handle, making them safer to work on.

9.2 Amplifier circuits – PENDING

9.3 Oscillator circuits – PENDING

9.4 Phase-locked loops – PENDING

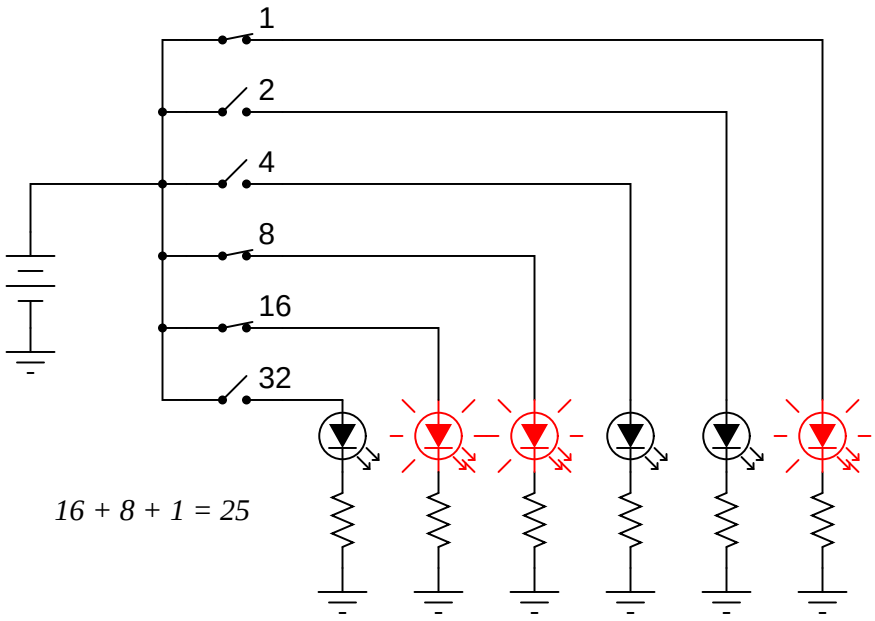
9.5 Radio circuits – PENDING

9.6 Computational circuits

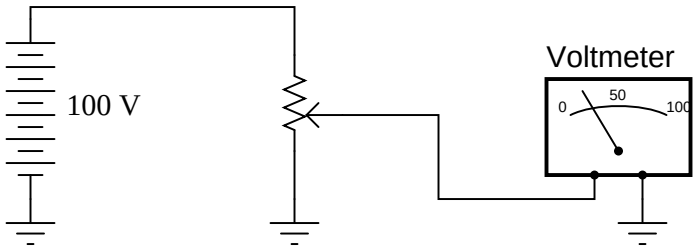
When someone mentions the word "computer," a digital device is what usually comes to mind. Digital circuits represent numerical quantities in *binary* format: patterns of 1's and 0's represented by a multitude of transistor circuits operating in saturated or cutoff states. However, analog circuitry may also be used to represent numerical quantities and perform mathematical calculations, by using variable voltage signals instead of discrete on/off states.

Here is a simple example of binary (digital) representation versus analog representation of the number "twenty-five:"

A digital circuit representing the number 25:



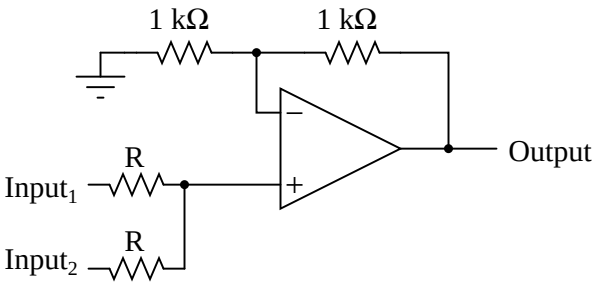
An analog circuit representing the number 25:



Digital circuits are very different from circuits built on analog principles. Digital computational circuits can be incredibly complex, and calculations must often be performed in sequential "steps" to obtain a final answer, much as a human being would perform arithmetical calculations in steps with pencil and paper. Analog computational circuits, on the other hand, are quite simple in comparison, and perform their calculations in continuous, real-time fashion. There is a disadvantage to using analog circuitry to represent numbers, though: imprecision. The digital circuit shown above is representing the number twenty-five, precisely. The analog circuit shown above may or may not be exactly calibrated to 25.000 volts, but is subject to "drift" and error.

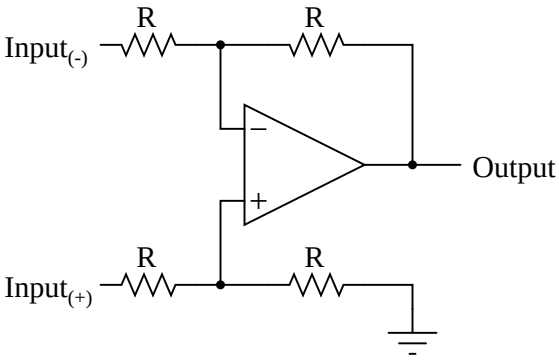
In applications where precision is not critical, analog computational circuits are very practical and elegant. Shown here are a few op-amp circuits for performing analog computation:

Analog summer (adder) circuit



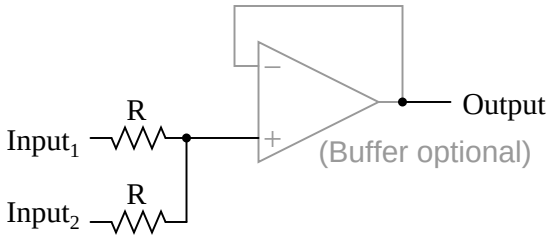
$$\text{Output} = \text{Input}_1 + \text{Input}_2$$

Analog subtractor circuit



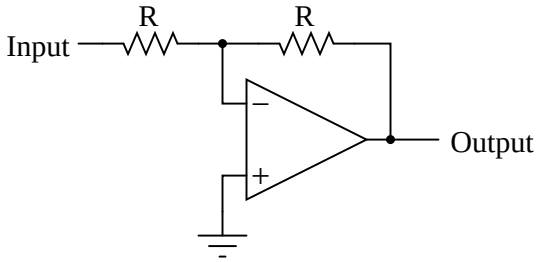
$$\text{Output} = \text{Input}_{(+)} - \text{Input}_{(-)}$$

Analog averager circuit



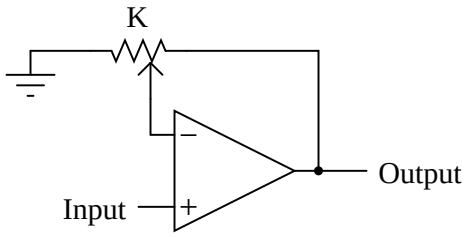
$$\text{Output} = \frac{\text{Input}_1 + \text{Input}_2}{2}$$

Analog inverter (sign reverser) circuit



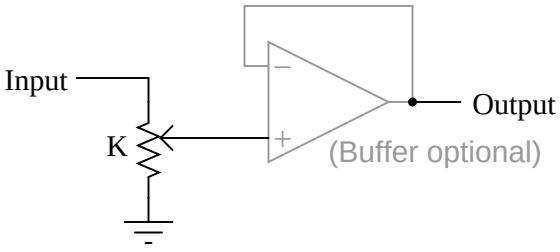
$$\text{Output} = -\text{Input}$$

Analog "multiply-by-constant" circuit



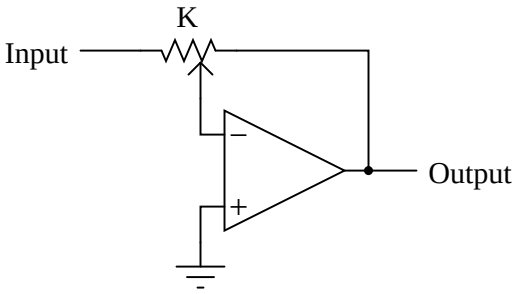
$$\text{Output} = (K)(\text{Input})$$

Analog "divide-by-constant" circuit



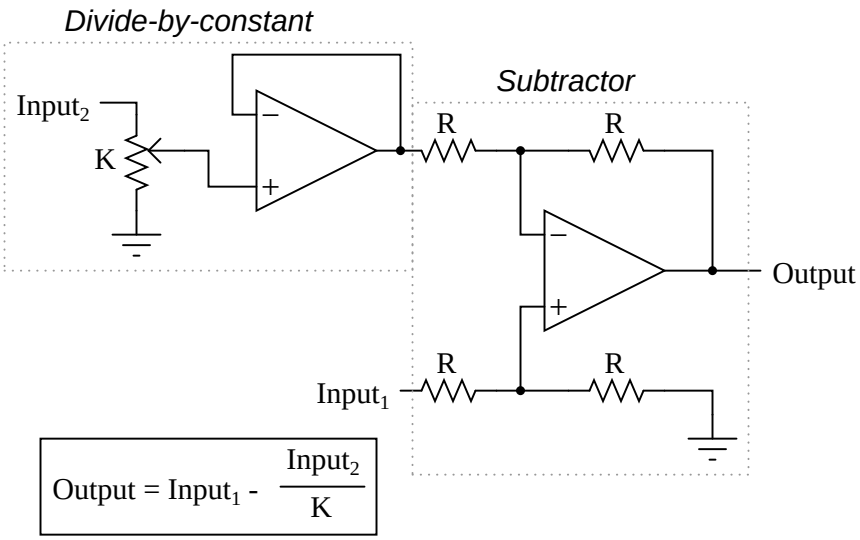
$$\text{Output} = \frac{\text{Input}}{K}$$

Analog inverting "multiply/divide-by-constant" circuit



$$\text{Output} = -(K)(\text{Input})$$

Each of these circuits may be used in modular fashion to create a circuit capable of multiple calculations. For instance, suppose that we needed to subtract a certain fraction of one variable from another variable. By combining a divide-by-constant circuit with a subtractor circuit, we could obtain the required function:



Devices called *analog computers* used to be common in universities and engineering shops, where dozens of op-amp circuits could be "patched" together with removable jumper wires to model mathematical statements, usually for the purpose of simulating some physical process whose underlying equations were known. Digital computers have made analog computers all but obsolete, but analog computational circuitry cannot be beaten by digital in terms of sheer elegance and economy of necessary components.

Analog computational circuitry excels at performing the calculus operations *integration* and *differentiation* with respect to time, by using capacitors in an op-amp feedback loop. To fully understand these circuits' operation and applications, though, we must first grasp the meaning of these fundamental calculus concepts. Fortunately, the application of op-amp circuits to real-world problems involving calculus serves as an excellent means to teach basic calculus. In the words of John I. Smith, taken from his outstanding textbook, *Modern Operational Circuit Design*:

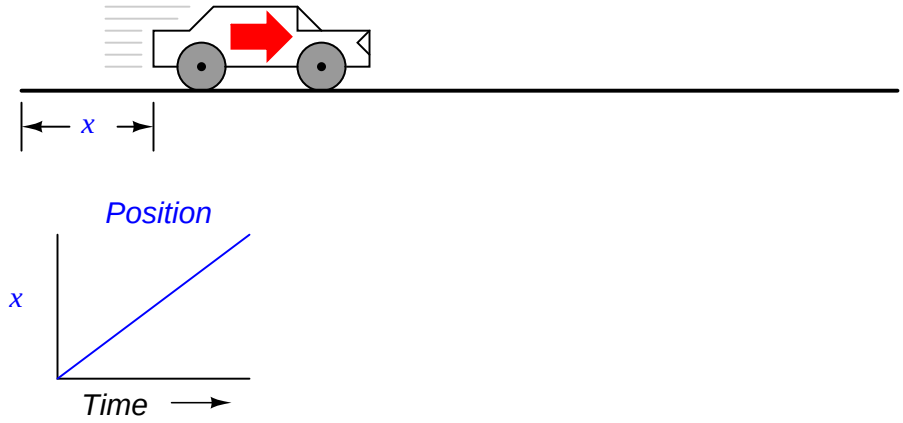
> *"A note of encouragement is offered to certain readers: integral calculus is one of the mathematical disciplines that operational [amplifier] circuitry exploits and, in the process, rather demolishes as a barrier to understanding."* (pg. 4)

Mr. Smith's sentiments on the pedagogical value of analog circuitry as a learning tool for mathematics are not unique. Consider the opinion of engineer George Fox Lang, in an article he wrote for the August 2000 issue of the journal *Sound and Vibration*, entitled, "Analog was <u>not</u> a Computer Trademark!":

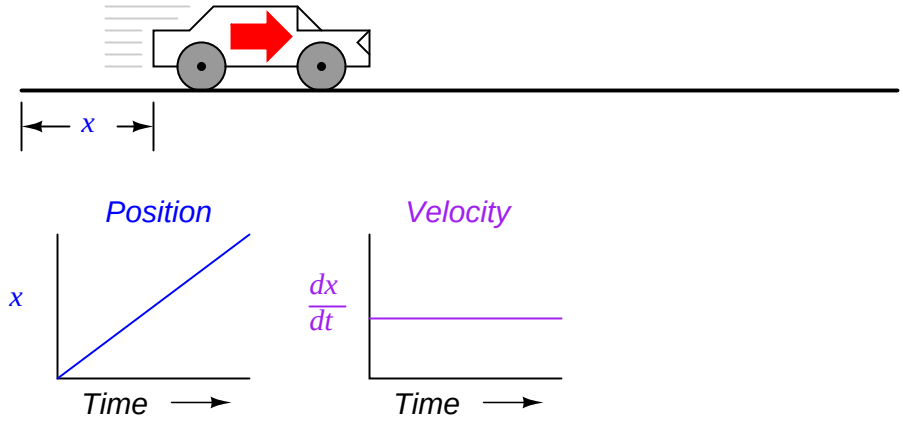
> *"Creating a real physical entity (a circuit) governed by a particular set of equations and interacting with it provides unique insight into those mathematical statements. There is no better way to develop a "gut feel" for the interplay between physics and mathematics than to experience such an interaction. The analog computer was a powerful interdisciplinary teaching tool; its obsolescence is mourned by many educators in a variety of fields."* (pg. 23)

Differentiation is the first operation typically learned by beginning calculus students. Simply put, differentiation is determining the instantaneous rate-of-change of one variable as it relates to another. In analog differentiator circuits, the independent variable is time, and so the rates of change we're dealing with are rates of change for an electronic signal (voltage or current) with respect to time.

Suppose we were to measure the position of a car, traveling in a direct path (no turns), from its starting point. Let us call this measurement, x. If the car moves at a rate such that its distance from "start" increases steadily over time, its position will plot on a graph as a *linear* function (straight line):



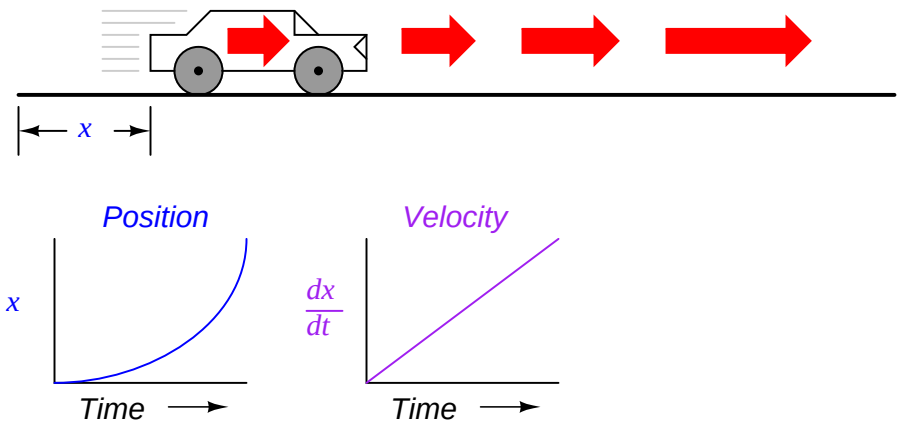
If we were to calculate the *derivative* of the car's position with respect to time (that is, determine the rate-of-change of the car's position with respect to time), we would arrive at a quantity representing the car's velocity. The differentiation function is represented by the fractional notation d/d, so when differentiating position (x) with respect to time (t), we denote the result (the derivative) as dx/dt:



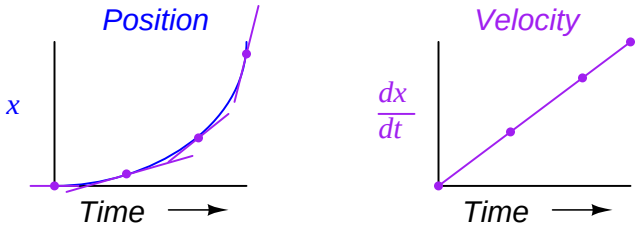
For a linear graph of x over time, the derivate of position (dx/dt), otherwise and more commonly known as *velocity*, will be a flat line, unchanging in value. The derivative of a mathematical function may be graphically understood as its *slope* when plotted on a graph, and here we can see that the

position (x) graph has a constant slope, which means that its derivative (dx/dt) must be constant over time.

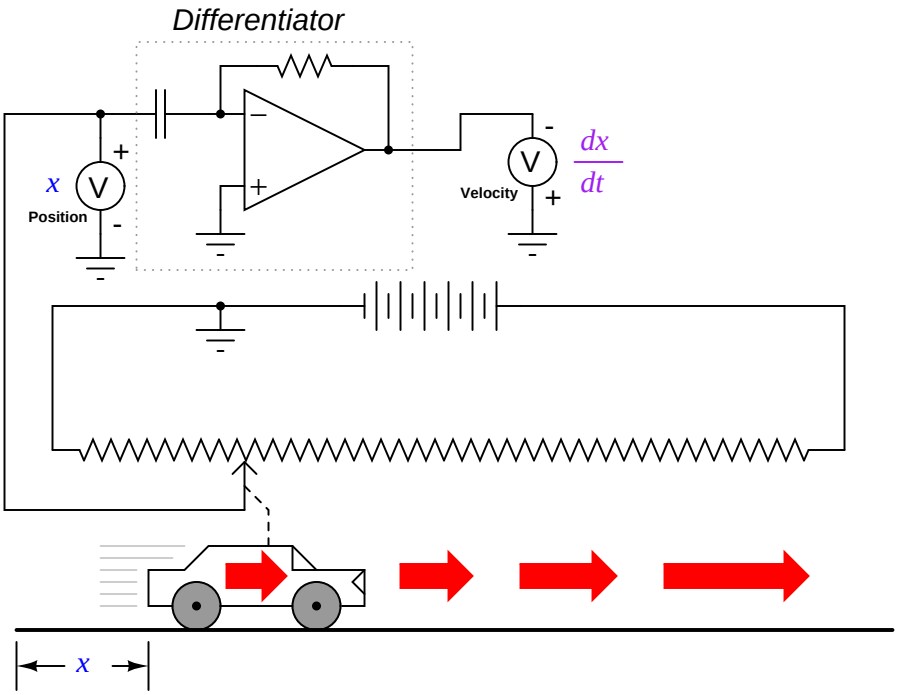
Now, suppose the distance traveled by the car increased exponentially over time: that is, it began its travel in slow movements, but covered more additional distance with each passing period in time. We would then see that the derivative of position (dx/dt), otherwise known as velocity (v), would not be constant over time, but would increase:



The height of points on the velocity graph correspond to the rates-of-change, or slope, of points at corresponding times on the position graph:



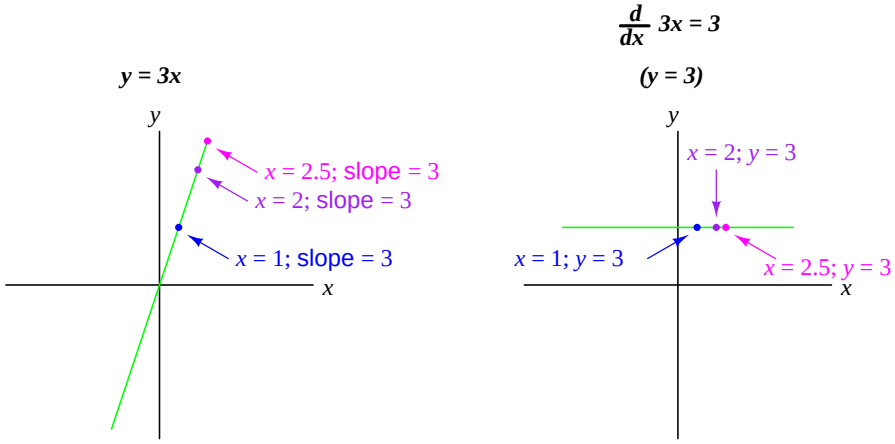
What does this have to do with analog electronic circuits? Well, if we were to have an analog voltage signal represent the car's position (think of a huge potentiometer whose wiper was attached to the car, generating a voltage proportional to the car's position), we could connect a differentiator circuit to this signal and have the circuit continuously *calculate* the car's velocity, displaying the result via a voltmeter connected to the differentiator circuit's output:



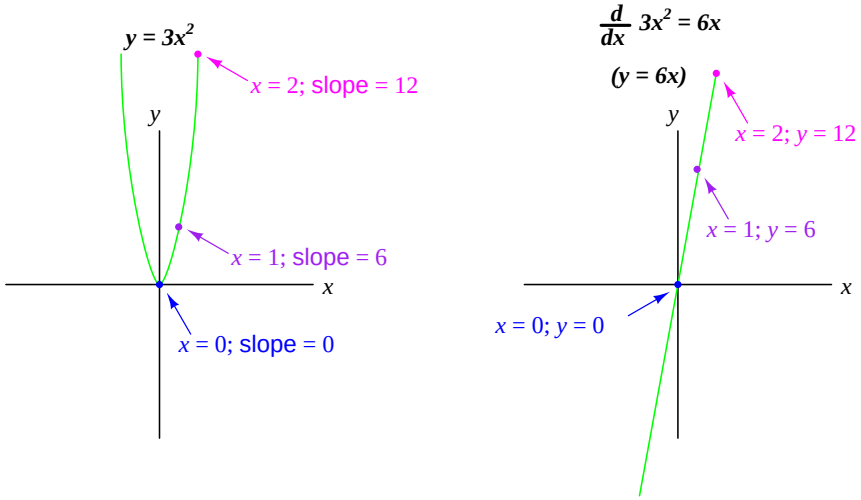
Recall from the last chapter that a differentiator circuit outputs a voltage proportional to the input voltage's *rate-of-change over time* (d/dt). Thus, if the input voltage is changing over time at a constant *rate*, the output voltage will be at a constant value. If the car moves in such a way that its elapsed distance over time builds up at a steady rate, then that means the car is traveling at a constant velocity, and the differentiator circuit will output a constant voltage proportional to that velocity. If the car's elapsed distance over time changes in a non-steady manner, the differentiator circuit's output will likewise be non-steady, but always at a level representative of the input's rate-of-change over time.

Note that the voltmeter registering velocity (at the output of the differentiator circuit) is connected in "reverse" polarity to the output of the op-amp. This is because the differentiator circuit shown is *inverting*: outputting a negative voltage for a positive input voltage rate-of-change. If we wish to have the voltmeter register a positive value for velocity, it will have to be connected to the op-amp as shown. As impractical as it may be to connect a giant potentiometer to a moving object such as an automobile, the concept should be clear: by electronically performing the calculus function of differentiation on a signal representing position, we obtain a signal representing velocity.

Beginning calculus students learn symbolic techniques for differentiation. However, this requires that the equation describing the original graph be known. For example, calculus students learn how to take a function such as $y = 3x$ and find its derivative with respect to x (d/dx), 3, simply by manipulating the equation. We may verify the accuracy of this manipulation by comparing the graphs of the two functions:

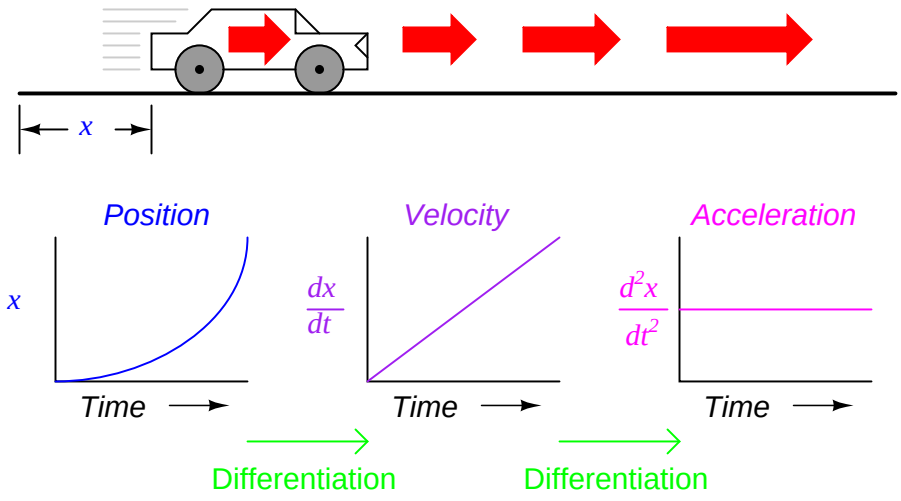


Nonlinear functions such as $y = 3x^2$ may also be differentiated by symbolic means. In this case, the derivative of $y = 3x^2$ with respect to x is $6x$:

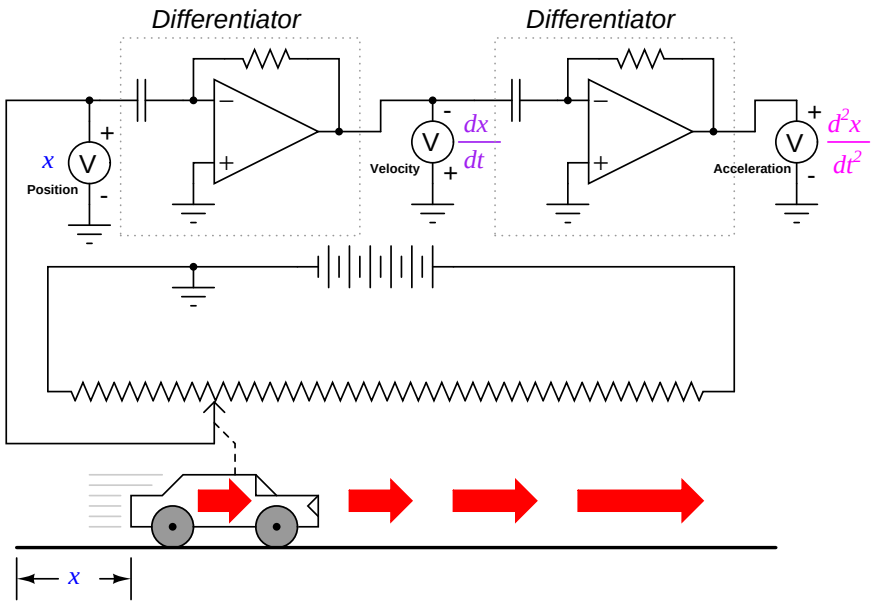


In real life, though, we often cannot describe the behavior of any physical event by a simple equation like $y = 3x$, and so symbolic differentiation of the type learned by calculus students may be impossible to apply to a physical measurement. If someone wished to determine the derivative of our hypothetical car's position (dx/dt = velocity) by symbolic means, they would first have to obtain an equation describing the car's position over time, based on position measurements taken from a real experiment – a nearly impossible task unless the car is operated under carefully controlled conditions leading to a very simple position graph. However, an analog differentiator circuit, by exploiting the behavior of a capacitor with respect to voltage, current, and time $i = C(dv/dt)$, naturally differentiates any real signal in relation to time, and would be able to output a signal corresponding to instantaneous velocity (dx/dt) at any moment. By logging the car's position signal along with the differentiator's output signal using a chart recorder or other data acquisition device, both graphs would naturally present themselves for inspection and analysis.

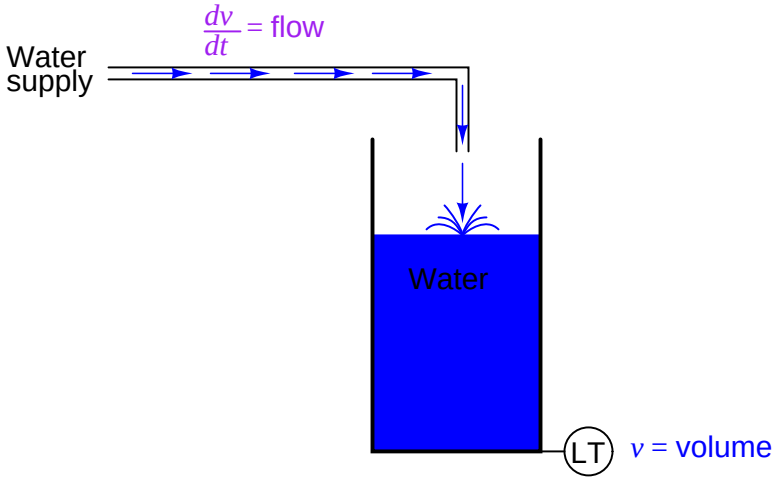
We may take the principle of differentiation one step further by applying it to the velocity signal using another differentiator circuit. In other words, use it to calculate the rate-of-change of velocity, which we know is the rate-of-change of position. What practical measure would we arrive at if we did this? Think of this in terms of the units we use to measure position and velocity. If we were to measure the car's position from its starting point in miles, then we would probably express its velocity in units of miles *per hour* (dx/dt). If we were to differentiate the velocity (measured in miles per hour) with respect to time, we would end up with a unit of miles per hour *per hour*. Introductory physics classes teach students about the behavior of falling objects, measuring position in *meters*, velocity in *meters per second*, and change in velocity over time in *meters per second, per second*. This final measure is called *acceleration*: the rate of change of velocity over time:



The expression d^2x/dt^2 is called the *second derivative* of position (x) with regard to time (t). If we were to connect a second differentiator circuit to the output of the first, the last voltmeter would register acceleration:



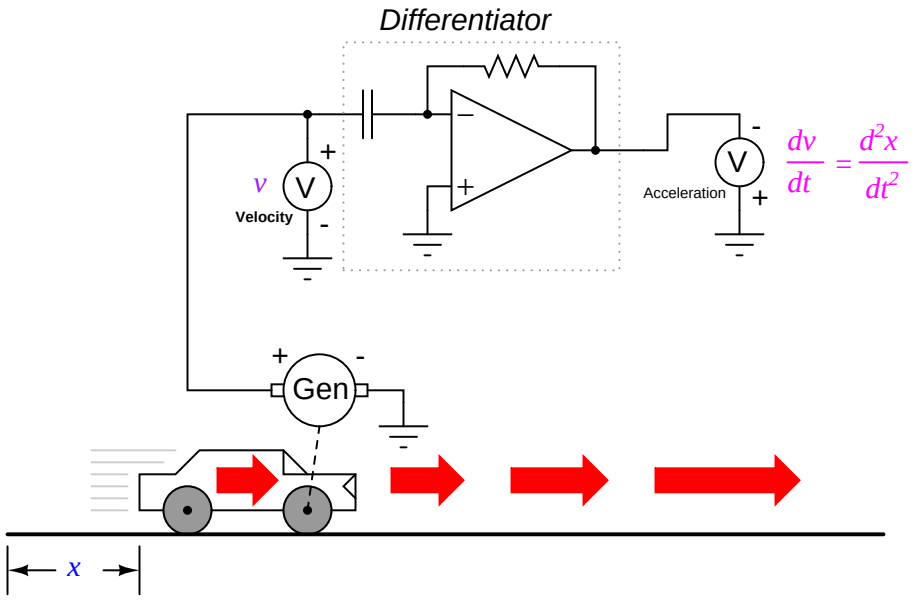
Deriving velocity from position, and acceleration from velocity, we see the principle of differentiation very clearly illustrated. These are not the only physical measurements related to each other in this way, but they are, perhaps, the most common. Another example of calculus in action is the relationship between liquid flow (q) and liquid volume (v) accumulated in a vessel over time:



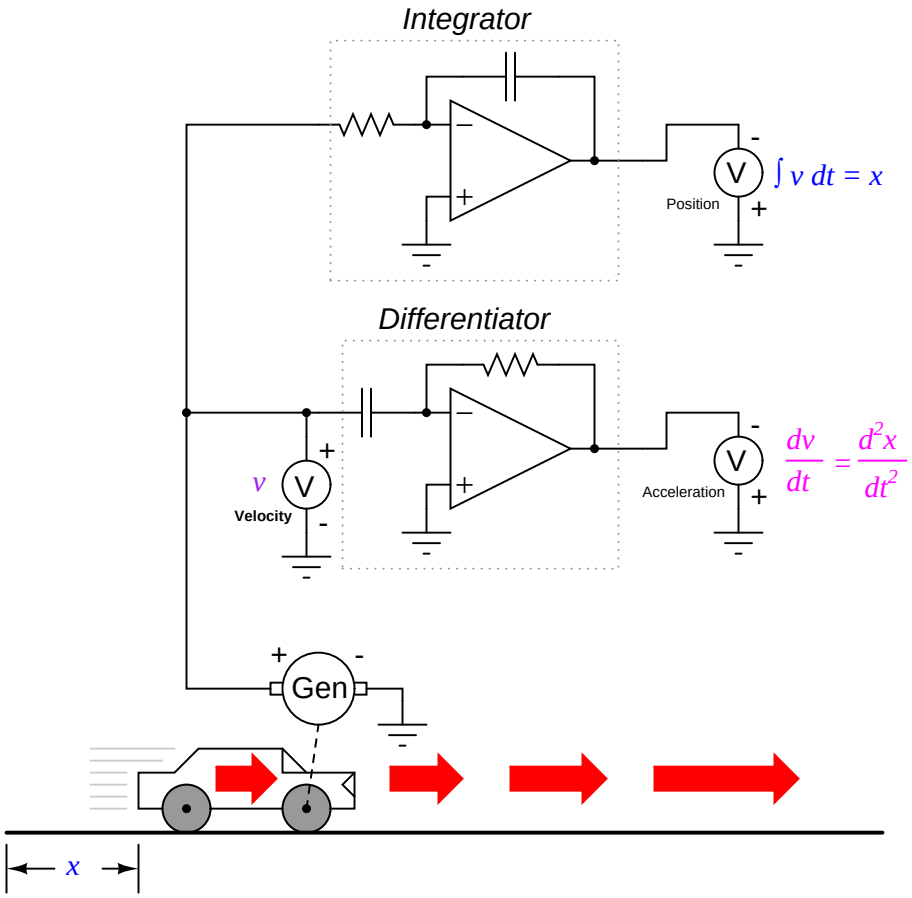
A "Level Transmitter" device mounted on a water storage tank provides a signal directly proportional to water level in the tank, which – if the tank is of constant cross-sectional area throughout its height – directly equates water volume stored. If we were to take this volume signal and differentiate it with respect to time (dv/dt), we would obtain a signal proportional to the water *flow rate* through the pipe carrying water to the tank. A differentiator circuit connected in such a way as to receive this volume signal would produce an output signal proportional to flow, possibly substituting for a

flow-measurement device ("Flow Transmitter") installed in the pipe.

Returning to the car experiment, suppose that our hypothetical car were equipped with a tachogenerator on one of the wheels, producing a voltage signal directly proportional to velocity. We could differentiate the signal to obtain acceleration with one circuit, like this:



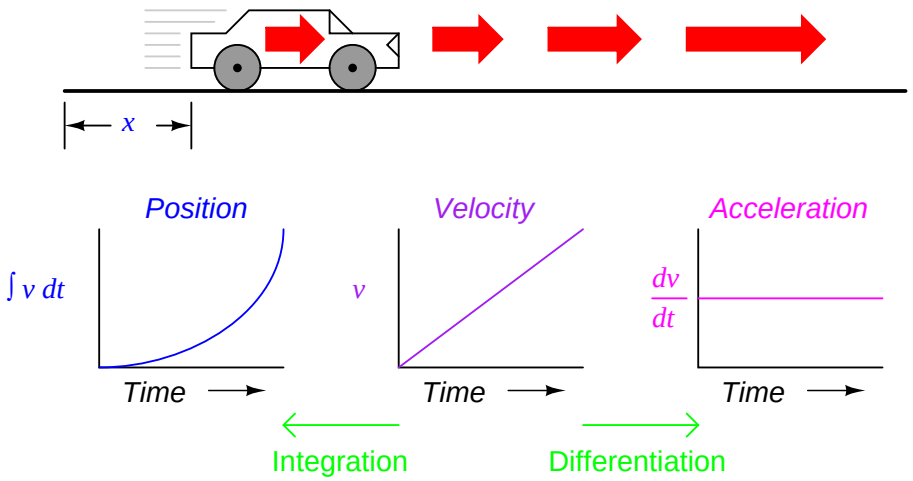
By its very nature, the tachogenerator differentiates the car's position with respect to time, generating a voltage proportional to how rapidly the wheel's angular position changes over time. This provides us with a raw signal already representative of velocity, with only a single step of differentiation needed to obtain an acceleration signal. A tachogenerator measuring velocity, of course, is a far more practical example of automobile instrumentation than a giant potentiometer measuring its physical position, but what we gain in practicality we lose in position measurement. No matter how many times we differentiate, we can never infer the car's position from a velocity signal. If the process of differentiation brought us from position to velocity to acceleration, then somehow we need to perform the "reverse" process of differentiation to go from velocity to position. Such a mathematical process does exist, and it is called *integration*. The "integrator" circuit may be used to perform this function of integration with respect to time:



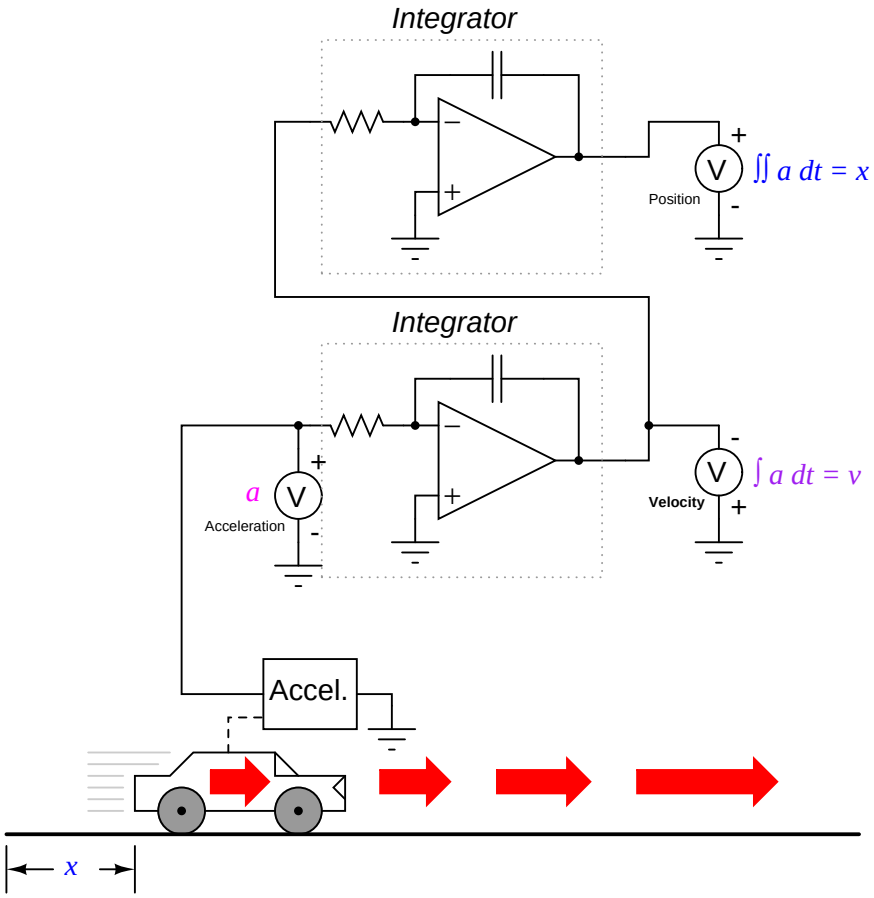
Recall from the last chapter that an integrator circuit outputs a voltage whose rate-of-change over time is proportional to the input voltage's magnitude. Thus, given a constant input voltage, the output voltage will *change* at a constant *rate*. If the car travels at a constant velocity (constant voltage input to the integrator circuit from the tachogenerator), then its distance traveled will increase steadily as time progresses, and the integrator will output a steadily changing voltage proportional to that distance. If the car's velocity is not constant, then neither will the rate-of-change over time be of the integrator circuit's output, but the output voltage *will* faithfully represent the amount of distance traveled by the car at any given point in time.

The symbol for integration looks something like a very narrow, cursive letter "S" ($\int$). The equation utilizing this symbol ($\int v\ dt = x$) tells us that we are integrating velocity (v) with respect to time (dt), and obtaining position (x) as a result.

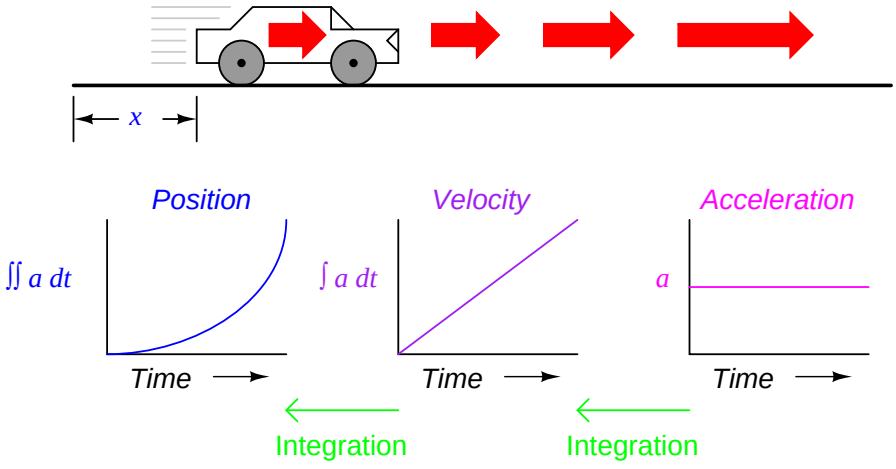
So, we may express three measures of the car's motion (position, velocity, and acceleration) in terms of velocity (v) just as easily as we could in terms of position (x):



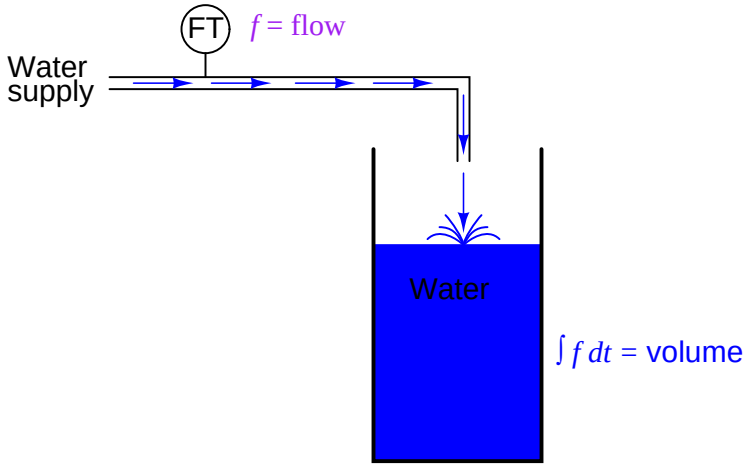
If we had an accelerometer attached to the car, generating a signal proportional to the rate of acceleration or deceleration, we could (hypothetically) obtain a velocity signal with one step of integration, and a position signal with a second step of integration:



Thus, all three measures of the car's motion (position, velocity, and acceleration) may be expressed in terms of acceleration:



As you might have suspected, the process of integration may be illustrated in, and applied to, other physical systems as well. Take for example the water storage tank and flow example shown earlier. If flow rate is the *derivative* of tank volume with respect to time ($q = dv/dt$), then we could also say that volume is the *integral* of flow rate with respect to time:

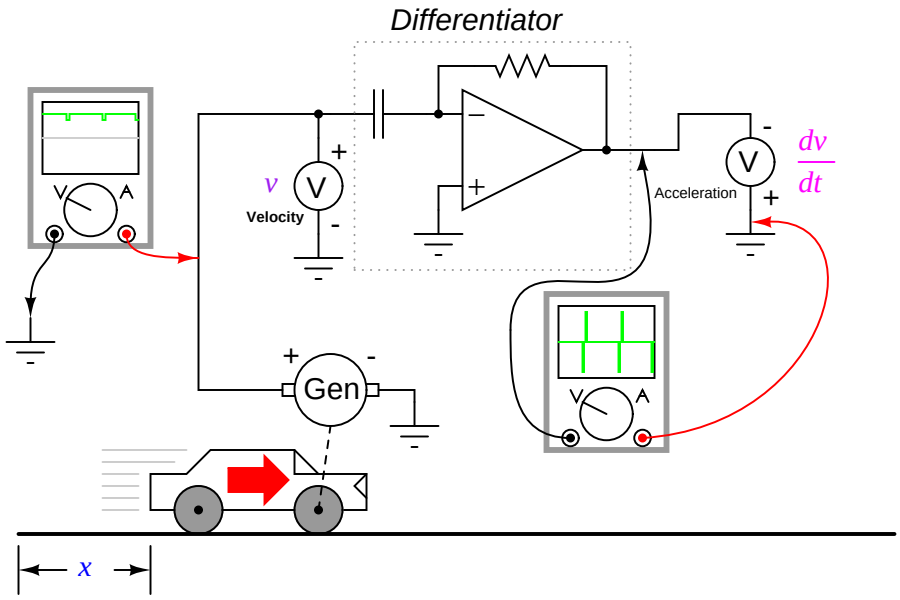


If we were to use a "Flow Transmitter" device to measure water flow, then by time-integration we could calculate the volume of water accumulated in the tank over time. Although it is theoretically possible to use a capacitive op-amp integrator circuit to derive a volume signal from a flow signal, mechanical and digital electronic "integrator" devices are more suitable for integration over long periods of time, and find frequent use in the water treatment and distribution industries.

Just as there are symbolic techniques for differentiation, there are also symbolic techniques for integration, although they tend to be more complex and varied. Applying symbolic integration to a real-world problem like the acceleration of a car, though, is still contingent on the availability of an equation precisely describing the measured signal – often a difficult or impossible thing to derive from measured data. However, electronic integrator circuits perform this mathematical function continuously, in real time, and for *any* input signal profile, thus providing a powerful tool for scientists and engineers.

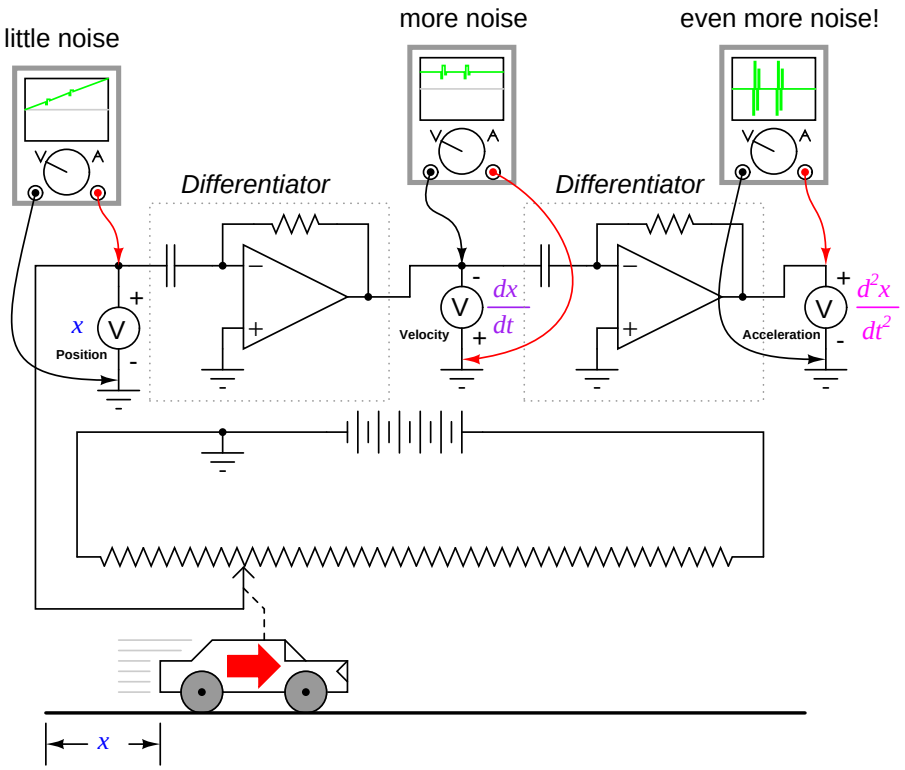
Having said this, there are caveats to the using calculus techniques to derive one type of measurement from another. Differentiation has the undesirable tendency of amplifying "noise" found in the measured variable, since the noise will typically appear as frequencies much higher than the measured variable, and high frequencies by their very nature possess high rates-of-change over time.

To illustrate this problem, suppose we were deriving a measurement of car acceleration from the velocity signal obtained from a tachogenerator with worn brushes or commutator bars. Points of poor contact between brush and commutator will produce momentary "dips" in the tachogenerator's output voltage, and the differentiator circuit connected to it will interpret these dips as very rapid changes in velocity. For a car moving at constant speed – neither accelerating nor decelerating – the acceleration signal should be 0 volts, but "noise" in the velocity signal caused by a faulty tachogenerator will cause the differentiated (acceleration) signal to contain "spikes," falsely indicating brief periods of high acceleration and deceleration:

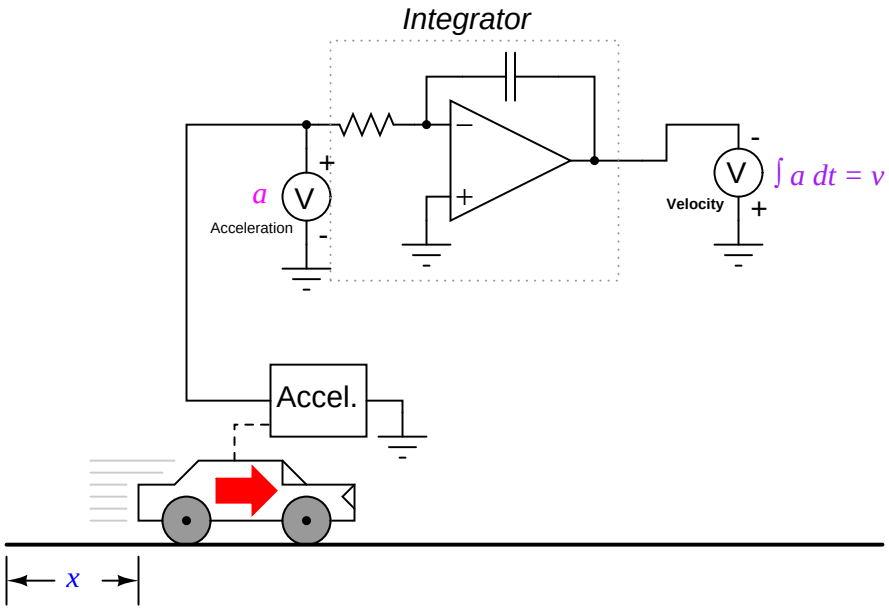


Noise voltage present in a signal to be differentiated need not be of significant amplitude to cause trouble: all that is required is that the noise profile have fast rise or fall times. In other words, any electrical noise with a high dv/dt component will be problematic when differentiated, even if it is of low amplitude.

It should be noted that this problem is not an artifact (an idiosyncratic error of the measuring/computing instrument) of the analog circuitry; rather, it is inherent to the process of differentiation. No matter how we might perform the differentiation, "noise" in the velocity signal will invariably corrupt the output signal. Of course, if we were differentiating a signal twice, as we did to obtain both velocity and acceleration from a position signal, the amplified noise signal output by the first differentiator circuit will be amplified again by the next differentiator, thus compounding the problem:

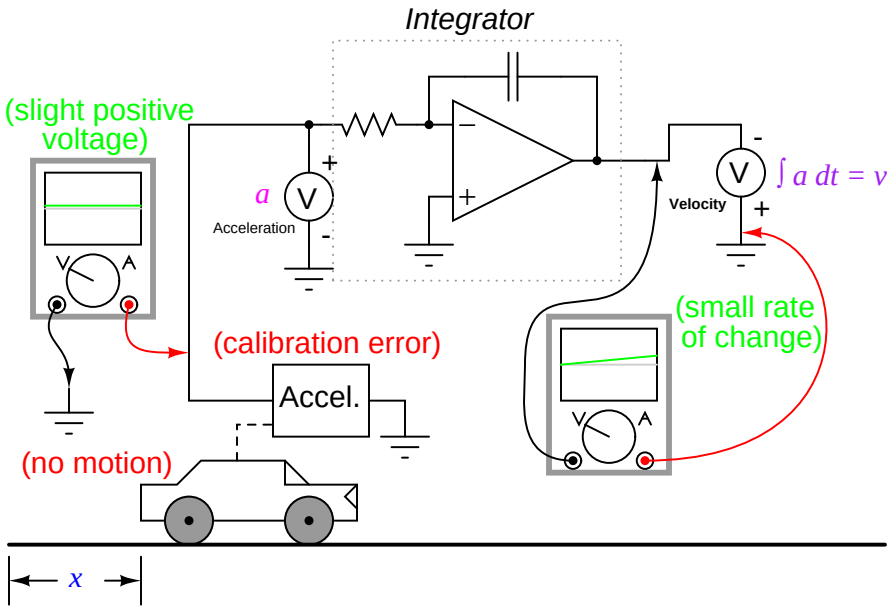


Integration does not suffer from this problem, because integrators act as low-pass filters, attenuating high-frequency input signals. In effect, all the high and low peaks resulting from noise on the signal become averaged together over time, for a diminished net result. One might suppose, then, that we could avoid all trouble by measuring acceleration directly and integrating that signal to obtain velocity; in effect, calculating in "reverse" from the way shown previously:



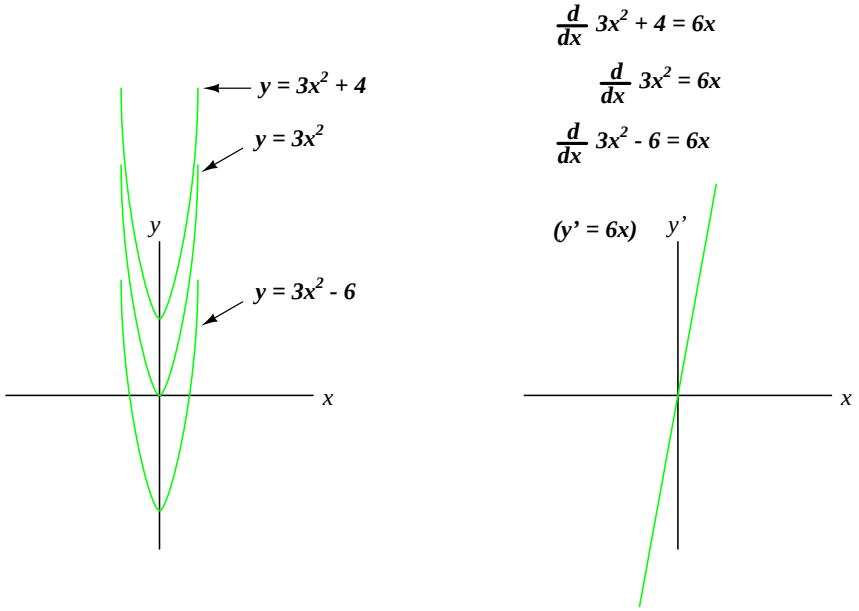
Unfortunately, following this methodology might lead us into other difficulties, one being a common artifact of analog integrator circuits known as *drift*. All op-amps have some amount of input bias current, and this current will tend to cause a charge to accumulate on the capacitor in addition to whatever charge accumulates as a result of the input voltage signal. In other words, all analog integrator circuits suffer from the tendency of having their output voltage "drift" or "creep" even when there is absolutely no voltage input, accumulating error over time as a result. Also, imperfect capacitors will tend to lose their stored charge over time due to internal resistance, resulting in "drift" toward zero output voltage. These problems *are* artifacts of the analog circuitry, and may be eliminated through the use of digital computation.

Circuit artifacts notwithstanding, possible errors may result from the integration of one measurement (such as acceleration) to obtain another (such as velocity) simply because of the way integration works. If the "zero" calibration point of the raw signal sensor is not perfect, it will output a slight positive or negative signal even in conditions when it should output nothing. Consider a car with an imperfectly calibrated accelerometer, or one that is influenced by gravity to detect a slight acceleration unrelated to car motion. Even with a perfect integrating computer, this sensor error will cause the integrator to accumulate error, resulting in an output signal indicating a change of velocity when the car is neither accelerating nor decelerating.



As with differentiation, this error will also compound itself if the integrated signal is passed on to another integrator circuit, since the "drifting" output of the first integrator will very soon present a significant positive or negative signal for the next integrator to integrate. Therefore, care should be taken when integrating sensor signals: if the "zero" adjustment of the sensor is not *perfect*, the integrated result will drift, even if the integrator circuit itself is perfect.

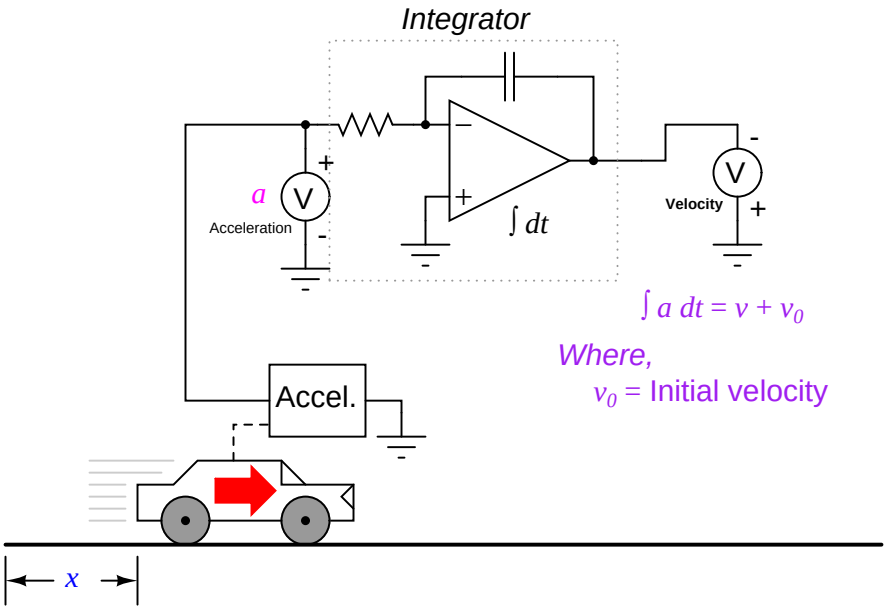
So far, the only integration errors discussed have been artificial in nature: originating from imperfections in the circuitry and sensors. There also exists a source of error inherent to the process of integration itself, and that is the *unknown constant* problem. Beginning calculus students learn that whenever a function is integrated, there exists an unknown constant (usually represented as the variable C) added to the result. This uncertainty is easiest to understand by comparing the derivatives of several functions differing only by the addition of a constant value:



Note how each of the parabolic curves ($y = 3x^2 + C$) share the exact same shape, differing from each other in regard to their vertical offset. However, they all share the exact same derivative function: $y' = (d/dx)(3x^2 + C) = 6x$, because they all share identical *rates of change* (slopes) at corresponding points along the x axis. While this seems quite natural and expected from the perspective of differentiation (different equations sharing a common derivative), it usually strikes beginning students as odd from the perspective of integration, because there are multiple correct answers for the integral of a function. Going from an equation to its derivative, there is only one answer, but going from that derivative back to the original equation leads us to a range of correct solutions. In honor of this uncertainty, the symbolic function of integration is called the *indefinite integral*.

When an integrator performs live signal integration with respect to time, the output is the sum of the integrated input signal over time *and* an initial value of arbitrary magnitude, representing the integrator's pre-existing output at the time integration began. For example, if I integrate the velocity of a car driving in a straight line away from a city, calculating that a constant velocity of 50 miles per hour over a time of 2 hours will produce a distance ($\int v\ dt$) of 100 miles, that does not necessarily mean the car will be 100 miles away from the city after 2 hours. All it tells us is that the car will be 100 miles *further* away from the city after 2 hours of driving. The actual distance from the city after 2 hours of driving depends on how far the car was from the city when integration began. If we do not know this initial value for distance, we cannot determine the car's exact distance from the city after 2 hours of driving.

This same problem appears when we integrate acceleration with respect to time to obtain velocity:



In this integrator system, the calculated velocity of the car will only be valid if the integrator circuit is *initialized* to an output value of zero when the car is stationary ($v = 0$). Otherwise, the integrator could very well be outputting a non-zero signal for velocity (v_0) when the car is stationary, for the accelerometer cannot tell the difference between a stationary state (0 miles per hour) and a state of constant velocity (say, 60 miles per hour, unchanging). This uncertainty in integrator output is inherent to the process of integration, and not an artifact of the circuitry or of the sensor.

In summary, if maximum accuracy is desired for any physical measurement, it is best to measure that variable directly rather than compute it from other measurements. This is not to say that computation is worthless. Quite to the contrary, often it is the only practical means of obtaining a desired measurement. However, the limits of computation must be understood and respected in order that precise measurements be obtained.

9.7 Measurement circuits – PENDING

9.8 Control circuits – PENDING

9.9 Contributors

Contributors to this chapter are listed in chronological order of their contributions, from most recent to first. See Appendix 2 (Contributor List) for dates and contact information.

Warren Young (August 2002): Initial idea and text for "Power supply circuits" section. Paragraphs modified by Tony Kuphaldt (changes in vocabulary, plus inclusion of additional concepts).

Chapter 10

ACTIVE FILTERS

Contents

*** PENDING ***

Chapter 11

DC MOTOR DRIVES

Contents

Chapter 12

INVERTERS AND AC MOTOR DRIVES

Contents

Chapter 13

ELECTRON TUBES

Contents

13.1 Introduction

An often neglected area of study in modern electronics is that of *tubes*, more precisely known as *vacuum tubes* or *electron tubes*. Almost completely overshadowed by semiconductor, or "solid-state" components in most modern applications, tube technology once dominated electronic circuit design.

In fact, the historical transition from "electric" to "electronic" circuits really began with tubes, for it was with tubes that we entered into a whole new realm of circuit function: a way of controlling the flow of electrons (current) in a circuit by means of another electric signal (in the case of most tubes, the controlling signal is a small voltage). The semiconductor counterpart to the tube, of course, is the transistor. Transistors perform much the same function as tubes: controlling the flow of electrons in a circuit by means of another flow of electrons in the case of the bipolar transistor, and controlling the flow of electrons by means of a voltage in the case of the field-effect transistor. In either case, a relatively small electric signal controls a relatively large electric current. This is the

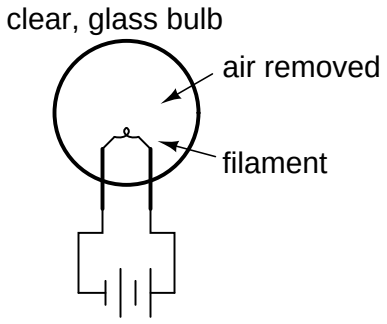
essence of the word "electronic," so as to distinguish it from "electric," which has more to do with how electron flow is regulated by Ohm's Law and the physical attributes of wire and components.

Though tubes are now obsolete for all but a few specialized applications, they are still a worthy area of study. If nothing else, it is fascinating to explore "the way things used to be done" in order to better appreciate modern technology.

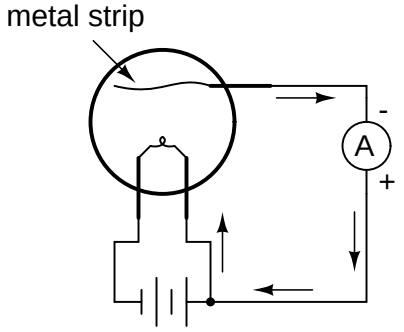
13.2 Early tube history

Thomas Edison, that prolific American inventor, is often credited with the invention of the incandescent lamp. More accurately, it could be said that Edison was the man who *perfected* the incandescent lamp. Edison's successful design of 1879 was actually preceded by 77 years by the British scientist Sir Humphry Davy, who first demonstrated the principle of using electric current to heat a thin strip of metal (called a "filament") to the point of incandescence (glowing white hot).

Edison was able to achieve his success by placing his filament (made of carbonized sewing thread) inside of a clear glass bulb from which the air had been forcibly removed. In this vacuum, the filament could glow at white-hot temperatures without being consumed by combustion:

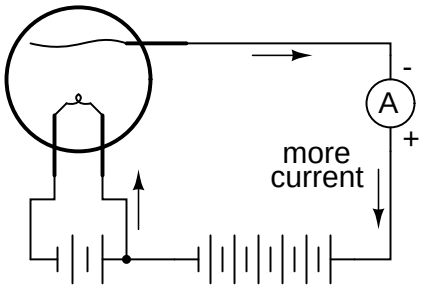


In the course of his experimentation (sometime around 1883), Edison placed a strip of metal inside of an evacuated (vacuum) glass bulb along with the filament. Between this metal strip and one of the filament connections he attached a sensitive ammeter. What he found was that electrons would flow through the meter whenever the filament was hot, but ceased when the filament cooled down:

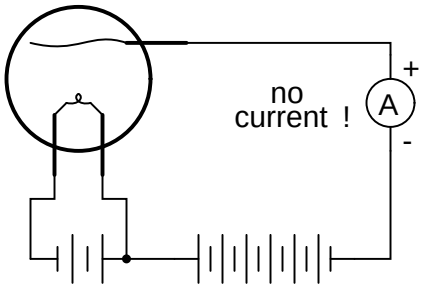


The white-hot filament in Edison's lamp was liberating free electrons into the vacuum of the lamp, those electrons finding their way to the metal strip, through the galvanometer, and back

to the filament. His curiosity piqued, Edison then connected a fairly high-voltage battery in the galvanometer circuit to aid the small current:



Sure enough, the presence of the battery created a much larger current from the filament to the metal strip. However, when the battery was turned around, there was little to no current at all!



In effect, what Edison had stumbled upon was a diode! Unfortunately, he saw no practical use for such a device and proceeded with further refinements in his lamp design.

The one-way electron flow of this device (known as the *Edison Effect*) remained a curiosity until J. A. Fleming experimented with its use in 1895. Fleming marketed his device as a "valve," initiating a whole new area of study in electric circuits. Vacuum tube diodes – Fleming's "valves" being no exception – are not able to handle large amounts of current, and so Fleming's invention was impractical for any application in AC power, only for small electric signals.

Then in 1906, another inventor by the name of Lee De Forest started playing around with the "Edison Effect," seeing what more could be gained from the phenomenon. In doing so, he made a startling discovery: by placing a metal screen between the glowing filament and the metal strip (which by now had taken the form of a plate for greater surface area), the stream of electrons flowing from filament to plate could be regulated by the application of a small voltage between the metal screen and the filament:

The DeForest "Audion" tube

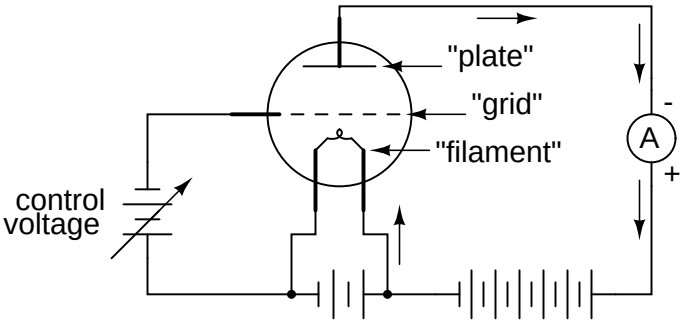


plate current can be controlled by the application of a small control voltage between the grid and filament!

De Forest called this metal screen between filament and plate a *grid*. It wasn't just the amount of voltage between grid and filament that controlled current from filament to plate, it was the polarity as well. A negative voltage applied to the grid with respect to the filament would tend to choke off the natural flow of electrons, whereas a positive voltage would tend to enhance the flow. Although there was some amount of current through the grid, it was very small; much smaller than the current through the plate.

Perhaps most importantly was his discovery that the small amounts of grid voltage and grid current were having large effects on the amount of plate voltage (with respect to the filament) and plate current. In adding the grid to Fleming's "valve," De Forest had made the valve adjustable: it now functioned as an *amplifying* device, whereby a small electrical signal could take control over a larger electrical quantity.

The closest semiconductor equivalent to the Audion tube, and to all of its more modern tube equivalents, is an n-channel D-type MOSFET. It is a voltage-controlled device with a large current gain.

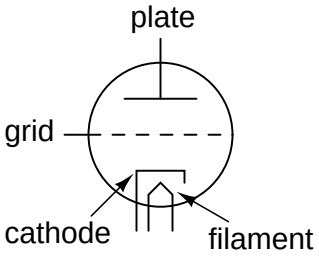
Calling his invention the "Audion," he vigorously applied it to the development of communications technology. In 1912 he sold the rights to his Audion tube as a telephone signal amplifier to the American Telephone and Telegraph Company (AT and T), which made long-distance telephone communication practical. In the following year he demonstrated the use of an Audion tube for generating radio-frequency AC signals. In 1915 he achieved the remarkable feat of broadcasting voice signals via radio from Arlington, Virginia to Paris, and in 1916 inaugurated the first radio news broadcast. Such accomplishments earned De Forest the title "Father of Radio" in America.

13.3 The triode

De Forest's Audion tube came to be known as the *triode* tube, because it had three elements: filament, grid, and plate (just as the "di" in the name *diode* refers to two elements, filament and plate). Later developments in diode tube technology led to the refinement of the electron emitter: instead of using the filament directly as the emissive element, another metal strip called the *cathode* could be heated by the filament.

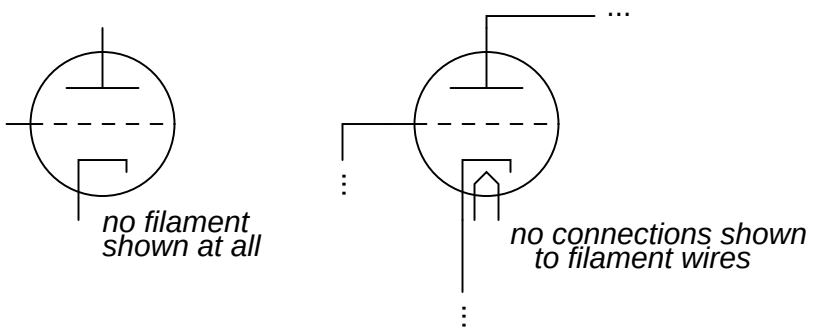
This refinement was necessary in order to avoid some undesired effects of an incandescent filament as an electron emitter. First, a filament experiences a voltage drop along its length, as current overcomes the resistance of the filament material and dissipates heat energy. This meant that the voltage potential between different points along the length of the filament wire and other elements in the tube would not be constant. For this and similar reasons, alternating current used as a power source for heating the filament wire would tend to introduce unwanted AC "noise" in the rest of the tube circuit. Furthermore, the surface area of a thin filament was limited at best, and limited surface area on the electron emitting element tends to place a corresponding limit on the tube's current-carrying capacity.

The cathode was a thin metal cylinder fitting snugly over the twisted wire of the filament. The cathode cylinder would be heated by the filament wire enough to freely emit electrons, without the undesirable side effects of actually carrying the heating current as the filament wire had to. The tube symbol for a triode with an indirectly-heated cathode looks like this:

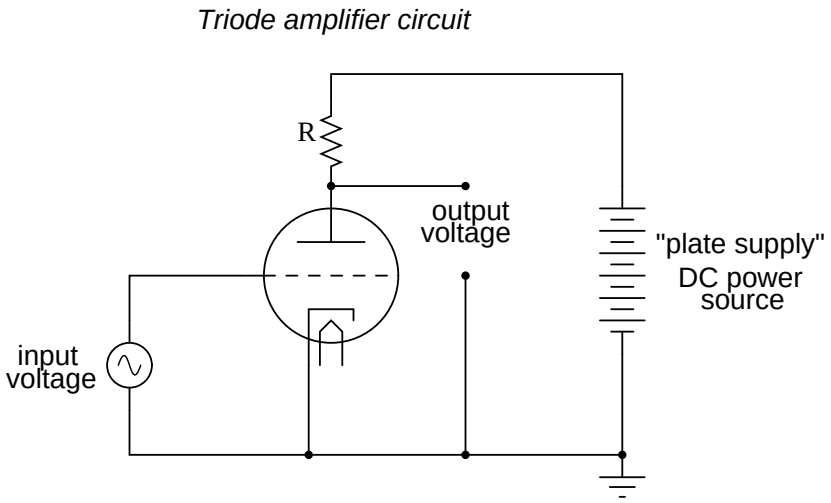


Since the filament is necessary for all but a few types of vacuum tubes, it is often omitted in the

symbol for simplicity, or it may be included in the drawing but with no power connections drawn to it:



A simple triode circuit is shown to illustrate its basic operation as an amplifier:



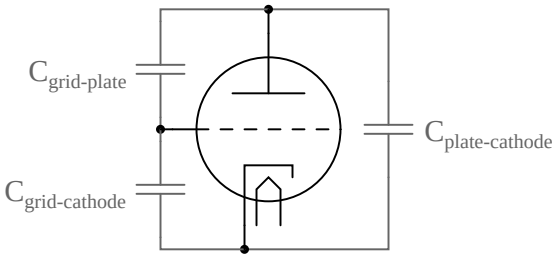
The low-voltage AC signal connected between the grid and cathode alternately suppresses, then enhances the electron flow between cathode and plate. This causes a change in voltage on the output of the circuit (between plate and cathode). The AC voltage and current magnitudes on the tube's grid are generally quite small compared with the variation of voltage and current in the plate circuit. Thus, the triode functions as an amplifier of the incoming AC signal (taking high-voltage, high-current DC power supplied from the large DC source on the right and "throttling" it by means of the tube's controlled conductivity).

In the triode, the amount of current from cathode to plate (the "controlled" current is a function both of grid-to-cathode voltage (the controlling signal) and the plate-to-cathode voltage (the electromotive force available to push electrons through the vacuum). Unfortunately, neither of these independent variables have a purely linear effect on the amount of current through the device (often referred to simply as the "plate current"). That is, triode current does not necessarily respond in a direct, proportional manner to the voltages applied.

In this particular amplifier circuit the nonlinearities are compounded, as plate voltage (with respect to cathode) changes along with the grid voltage (also with respect to cathode) as plate

current is throttled by the tube. The result will be an output voltage waveform that doesn't precisely resemble the waveform of the input voltage. In other words, the quirkiness of the triode tube and the dynamics of this particular circuit will *distort* the waveshape. If we really wanted to get complex about how we stated this, we could say that the tube introduces *harmonics* by failing to exactly reproduce the input waveform.

Another problem with triode behavior is that of stray capacitance. Remember that any time we have two conductive surfaces separated by an insulating medium, a capacitor will be formed. Any voltage between those two conductive surfaces will generate an electric field within that insulating region, potentially storing energy and introducing reactance into a circuit. Such is the case with the triode, most problematically between the grid and the plate. It is as if there were tiny capacitors connected between the pairs of elements in the tube:

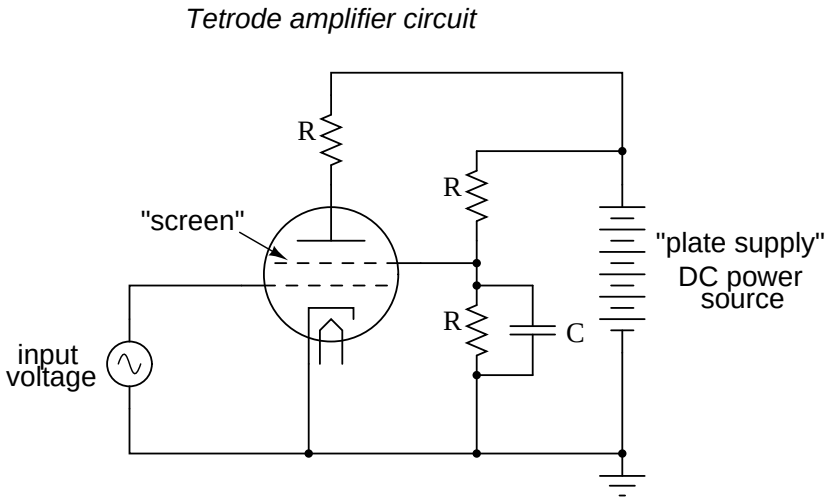


Now, this stray capacitance is quite small, and the reactive impedances usually high. Usually, that is, unless radio frequencies are being dealt with. As we saw with De Forest's Audion tube, radio was probably the prime application for this new technology, so these "tiny" capacitances became more than just a potential problem. Another refinement in tube technology was necessary to overcome the limitations of the triode.

13.4 The tetrode

As the name suggests, the *tetrode* tube contained four elements: cathode (with the implicit filament, or "heater"), grid, plate, and a new element called the *screen*. Similar in construction to the grid, the screen was a wire mesh or coil positioned between the grid and plate, connected to a source of positive DC potential (with respect to the cathode, as usual) equal to a fraction of the plate voltage. When connected to ground through an external capacitor, the screen had the effect of electrostatically shielding the grid from the plate. Without the screen, the capacitive linking between the plate and the grid could cause significant signal feedback at high frequencies, resulting in unwanted oscillations.

The screen, being of less surface area and lower positive potential than the plate, didn't attract many of the electrons passing through the grid from the cathode, so the vast majority of electrons in the tube still flew by the screen to be collected by the plate:



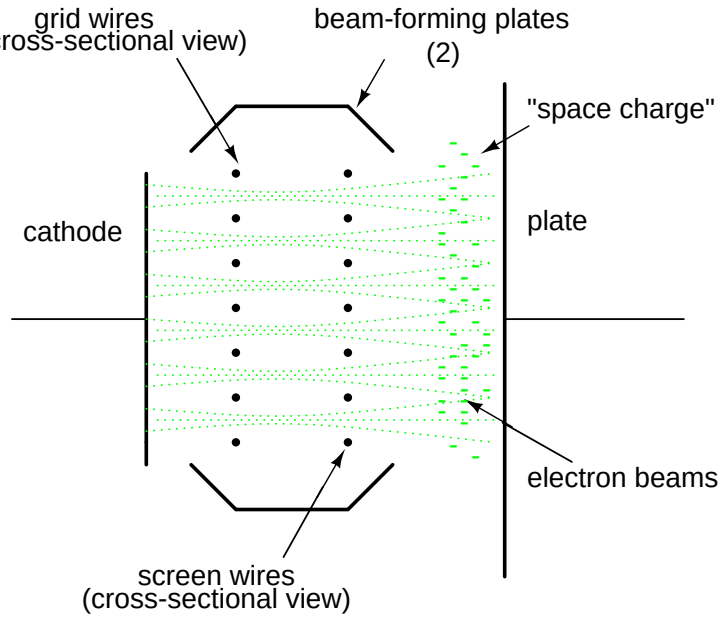
With a constant DC screen voltage, electron flow from cathode to plate became almost exclusively dependent upon grid voltage, meaning the plate voltage could vary over a wide range with little effect on plate current. This made for more stable gains in amplifier circuits, and better linearity for more accurate reproduction of the input signal waveform.

Despite the advantages realized by the addition of a screen, there were some disadvantages as well. The most significant disadvantage was related to something known as *secondary emission*. When electrons from the cathode strike the plate at high velocity, they can cause free electrons to be jarred loose from atoms in the metal of the plate. These electrons, knocked off the plate by the impact of the cathode electrons, are said to be "secondarily emitted." In a triode tube, secondary emission is not that great a problem, but in a tetrode with a positively-charged screen grid in close proximity, these secondary electrons will be attracted to the screen rather than the plate from which they came, resulting in a loss of plate current. Less plate current means less gain for the amplifier, which is not good.

Two different strategies were developed to address this problem of the tetrode tube: *beam power* tubes and *pentodes*. Both solutions resulted in new tube designs with approximately the same electrical characteristics.

13.5 Beam power tubes

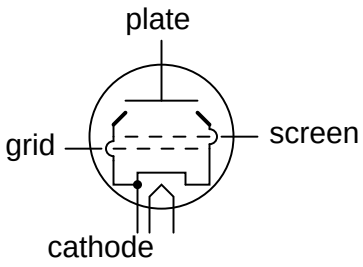
In the beam power tube, the basic four-element structure of the tetrode was maintained, but the grid and screen wires were carefully arranged along with a pair of auxiliary plates to create an interesting effect: focused beams or "sheets" of electrons traveling from cathode to plate. These electron beams formed a stationary "cloud" of electrons between the screen and plate (called a "space charge") which acted to repel secondary electrons emitted from the plate back to the plate. A set of "beam-forming" plates, each connected to the cathode, were added to help maintain proper electron beam focus. Grid and screen wire coils were arranged in such a way that each turn or wrap of the screen fell directly behind a wrap of the grid, which placed the screen wires in the "shadow" formed by the grid. This precise alignment enabled the screen to still perform its shielding function with minimal interference to the passage of electrons from cathode to plate.



This resulted in lower screen current (and more plate current!) than an ordinary tetrode tube, with little added expense to the construction of the tube.

Beam power tetrodes were often distinguished from their non-beam counterparts by a different schematic symbol, showing the beam-forming plates:

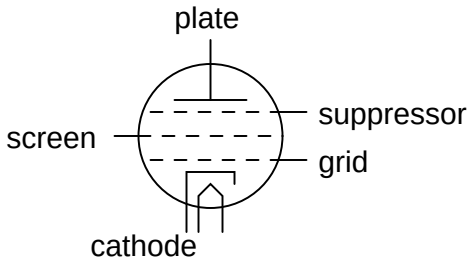
The "Beam power" tetrode tube



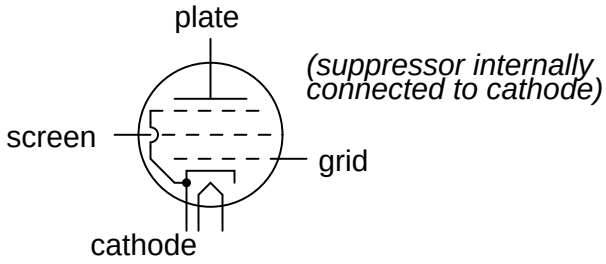
13.6 The pentode

Another strategy for addressing the problem of secondary electrons being attracted by the screen was the addition of a fifth wire element to the tube structure: a *suppressor*. These five-element tubes were naturally called *pentodes*.

The pentode tube



The suppressor was another wire coil or mesh situated between the screen and the plate, usually connected directly to ground potential. In some pentode tube designs, the suppressor was internally connected to the cathode so as to minimize the number of connection pins having to penetrate the tube envelope:



The suppressor's job was to repel any secondarily emitted electrons back to the plate: a structural equivalent of the beam power tube's space charge. This, of course, increased plate current and decreased screen current, resulting in better gain and overall performance. In some instances it allowed for greater operating plate voltage as well.

13.7 Combination tubes

Similar in thought to the idea of the integrated circuit, tube designers tried integrating different tube functions into single tube envelopes to reduce space requirements in more modern tube-type electronic equipment. A common combination seen within a single glass shell was two either diodes or two triodes. The idea of fitting pairs of diodes inside a single envelope makes a lot of sense in light of power supply full-wave rectifier designs, always requiring multiple diodes.

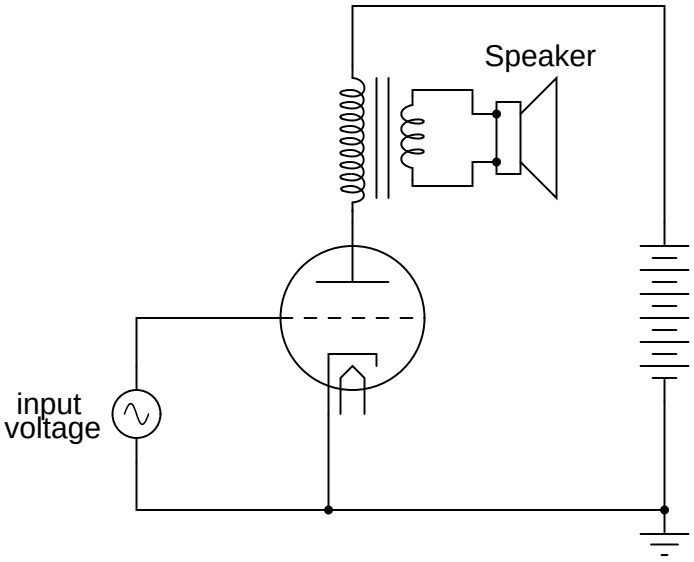
Of course, it would have been quite impossible to combine thousands of tube elements into a single tube envelope the way that thousands of transistors can be etched onto a single piece of silicon, but engineers still did their best to push the limits of tube miniaturization and consolidation. Some of these tubes, whimsically called *compactrons*, held four or more complete tube elements within a single envelope.

Sometimes the functions of two different tubes could be integrated into a single, combination tube in a way that simply worked more elegantly than two tubes ever could. An example of this was the *pentagrid converter*, more generally called a *heptode*, used in some superheterodyne radio designs. These tubes contained seven elements: 5 grids plus a cathode and a plate. Two of the

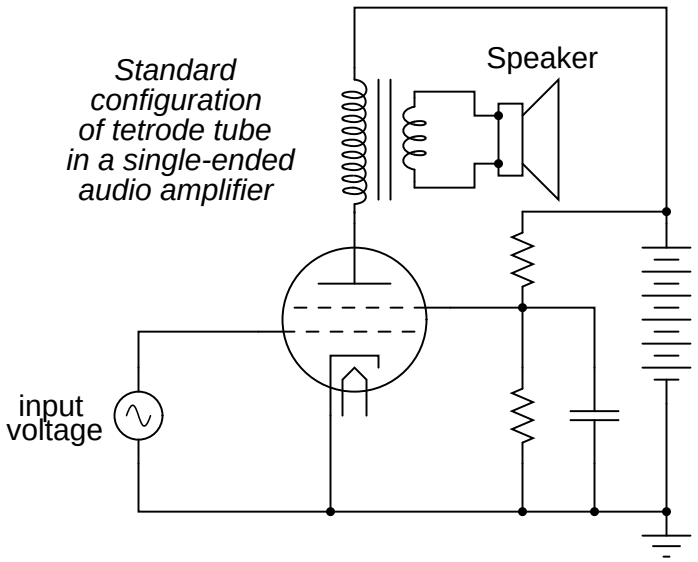
grids were normally reserved for signal input, the other three relegated to screening and suppression (performance-enhancing) functions. Combining the superheterodyne functions of oscillator and signal mixer together in one tube, the signal coupling between these two stages was intrinsic. Rather than having separate oscillator and mixer circuits, the oscillator creating an AC voltage and the mixer "mixing" that voltage with another signal, the pentagrid converter's oscillator section created an electron stream that oscillated in intensity which then directly passed through another grid for "mixing" with another signal.

This same tube was sometimes used in a different way: by applying a DC voltage to one of the control grids, the gain of the tube could be changed for a signal impressed on the other control grid. This was known as *variable-mu* operation, because the "mu" (μ) of the tube (its amplification factor, measured as a ratio of plate-to-cathode voltage change over grid-to-cathode voltage change with a constant plate current) could be altered at will by a DC control voltage signal.

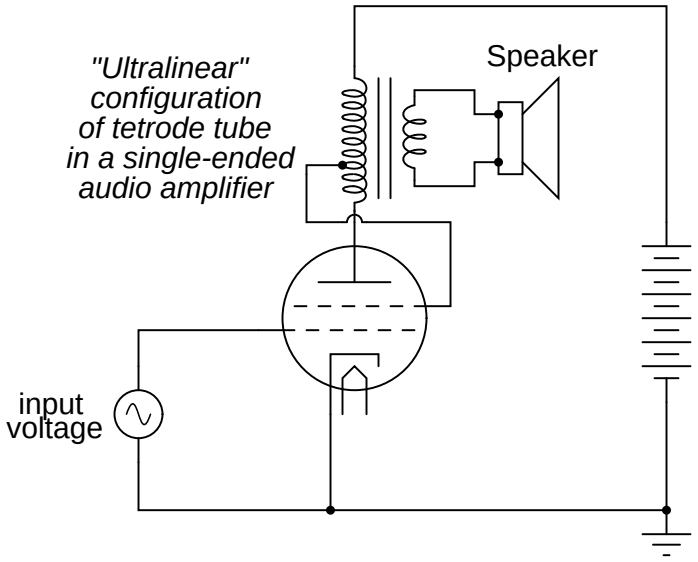
Enterprising electronics engineers also discovered ways to exploit such multi-variable capabilities of "lesser" tubes such as tetrodes and pentodes. One such way was the so-called *ultralinear* audio power amplifier, invented by a pair of engineers named Hafler and Keroes, utilizing a tetrode tube in combination with a "tapped" output transformer to provide substantial improvements in amplifier linearity (decreases in distortion levels). Consider a "single-ended" triode tube amplifier with an output transformer coupling power to the speaker:



If we substitute a tetrode for a triode in this circuit, we will see improvements in circuit gain resulting from the electrostatic shielding offered by the screen, preventing unwanted feedback between the plate and control grid:



However, the tetrode's screen may be used for functions other than merely shielding the grid from the plate. It can also be used as another control element, like the grid itself. If a "tap" is made on the transformer's primary winding, and this tap connected to the screen, the screen will receive a voltage that varies with the signal being amplified (feedback). More specifically, the feedback signal is proportional to the rate-of-change of magnetic flux in the transformer core ($d\Phi/dt$), thus improving the amplifier's ability to reproduce the input signal waveform at the speaker terminals and not just in the primary winding of the transformer:



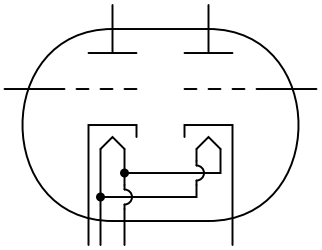
This signal feedback results in significant improvements in amplifier linearity (and consequently, distortion), so long as precautions are taken against "overpowering" the screen with too great a

positive voltage with respect to the cathode. As a concept, the ultralinear (screen-feedback) design demonstrates the flexibility of operation granted by multiple grid-elements inside a single tube: a capability rarely matched by semiconductor components.

Some tube designs combined multiple tube functions in a most economic way: dual plates with a single cathode, the currents for each of the plates controlled by separate sets of control grids. Common examples of these tubes were *triode-heptode* and *triode-hexode* tubes (a hexode tube is a tube with four grids, one cathode, and one plate).

Other tube designs simply incorporated separate tube structures inside a single glass envelope for greater economy. Dual diode (rectifier) tubes were quite common, as were dual triode tubes, especially when the power dissipation of each tube was relatively low.

Dual triode tube



The 12AX7 and 12AU7 models are common examples of dual-triode tubes, both of low-power rating. The 12AX7 is especially common as a preamplifier tube in electric guitar amplifier circuits.

13.8 Tube parameters

For bipolar junction transistors, the fundamental measure of amplification is the Beta ratio (β), defined as the ratio of collector current to base current (I_C/I_B). Other transistor characteristics such as junction resistance, which in some amplifier circuits may impact performance as much as β, are quantified for the benefit of circuit analysis. Electron tubes are no different, their performance characteristics having been explored and quantified long ago by electrical engineers.

Before we can speak meaningfully on these characteristics, we must define several mathematical variables used for expressing common voltage, current, and resistance measurements as well as some of the more complex quantities:

μ = amplification factor, pronounced "*mu*" (unitless)

g_m = mutual conductance, in siemens

E_p = plate-to-cathode voltage

E_g = grid-to-cathode voltage

I_p = plate current

I_k = cathode current

E_s = input signal voltage

r_p = dynamic plate resistance, in ohms

Δ = delta, the Greek symbol for *change*

The two most basic measures of an amplifying tube's characteristics are its amplification factor (μ) and its mutual conductance (g_m), also known as *transconductance*. Transconductance is defined here just the same as it is for field-effect transistors, another category of voltage-controlled devices. Here are the two equations defining each of these performance characteristics:

$$\mu = \frac{\Delta E_p}{\Delta E_g} \quad \text{with constant } I_p \text{ (plate current)}$$

$$g_m = \frac{\Delta I_p}{\Delta E_g} \quad \text{with constant } E_p \text{ (plate voltage)}$$

Another important, though more abstract, measure of tube performance is its *plate resistance*. This is the measurement of plate voltage change over plate current change for a constant value of grid voltage. In other words, this is an expression of how much the tube acts like a resistor for any given amount of grid voltage, analogous to the operation of a JFET in its ohmic mode:

$$r_p = \frac{\Delta E_p}{\Delta I_p} \quad \text{with constant } E_g \text{ (grid voltage)}$$

The astute reader will notice that plate resistance may be determined by dividing the amplification factor by the transconductance:

$$\mu = \frac{\Delta E_p}{\Delta E_g} \qquad g_m = \frac{\Delta I_p}{\Delta E_g}$$

. . . dividing μ by g_m . . .

$$r_p = \frac{\frac{\Delta E_p}{\Delta E_g}}{\frac{\Delta I_p}{\Delta E_g}}$$

$$r_p = \frac{\Delta E_p}{\Delta E_g} \frac{\Delta E_g}{\Delta I_p}$$

$$r_p = \frac{\Delta E_p}{\Delta I_p}$$

These three performance measures of tubes are subject to change from tube to tube (just as β ratios between two "identical" bipolar transistors are never precisely the same) and between different operating conditions. This variability is due partly to the unavoidable nonlinearities of electron tubes and partly due to how they are defined. Even supposing the existence of a perfectly linear tube, it will be impossible for all three of these measures to be constant over the allowable ranges of operation. Consider a tube that *perfectly* regulates current at any given amount of grid voltage (like a bipolar transistor with an absolutely constant β): that tube's plate resistance *must* vary with plate voltage, because plate current will not change even though plate voltage does.

Nevertheless, tubes were (and are) rated by these values at given operating conditions, and may have their characteristic curves published just like transistors.

13.9 Ionization (gas-filled) tubes

So far, we've explored tubes which are totally "evacuated" of all gas and vapor inside their glass envelopes, properly known as *vacuum tubes*. With the addition of certain gases or vapors, however, tubes take on significantly different characteristics, and are able to fulfill certain special roles in electronic circuits.

When a high enough voltage is applied across a distance occupied by a gas or vapor, or when that gas or vapor is heated sufficiently, the electrons of those gas molecules will be stripped away from their respective nuclei, creating a condition of *ionization*. Having freed the electrons from their electrostatic bonds to the atoms' nuclei, they are free to migrate in the form of a current, making the ionized gas a relatively good conductor of electricity. In this state, the gas is more properly referred to as a *plasma*.

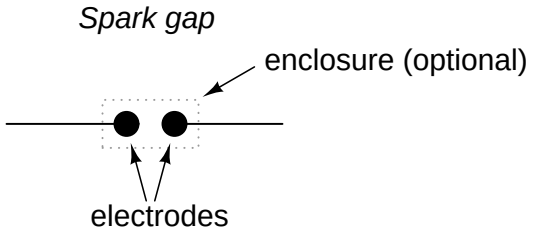
Ionized gas is not a perfect conductor. As such, the flow of electrons through ionized gas will tend to dissipate energy in the form of heat, thereby helping to keep the gas in a state of ionization. The result of this is a tube that will begin to conduct under certain conditions, then tend to stay

in a state of conduction until the applied voltage across the gas and/or the heat-generating current drops to a minimum level.

The astute observer will note that this is precisely the kind of behavior exhibited by a class of semiconductor devices called "thyristors," which tend to stay "on" once turned "on" and tend to stay "off" once turned "off." Gas-filled tubes, it can be said, manifest this same property of *hysteresis*.

Unlike their vacuum counterparts, ionization tubes were often manufactured with no filament (heater) at all. These were called *cold-cathode* tubes, with the heated versions designated as *hot-cathode* tubes. Whether or not the tube contained a source of heat obviously impacted the characteristics of a gas-filled tube, but not to the extent that lack of heat would impact the performance of a hard-vacuum tube.

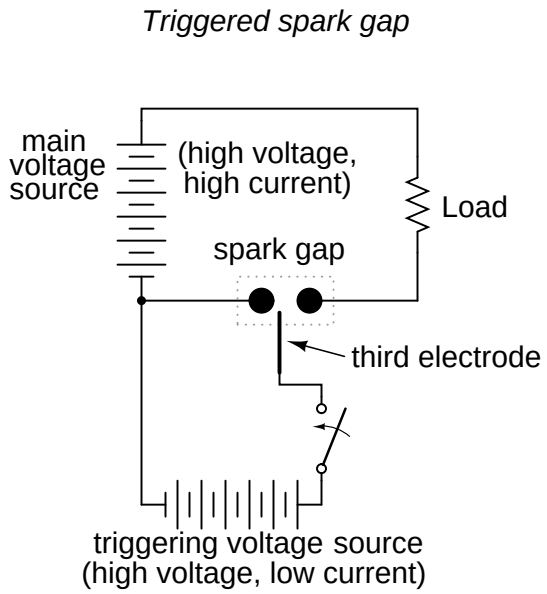
The simplest type of ionization device is not necessarily a tube at all; rather, it is constructed of two electrodes separated by a gas-filled gap. Simply called a *spark gap*, the gap between the electrodes may be occupied by ambient air, other times a special gas, in which case the device must have a sealed envelope of some kind.



A prime application for spark gaps is in overvoltage protection. Engineered not to ionize, or "break down" (begin conducting), with normal system voltage applied across the electrodes, the spark gap's function is to conduct in the event of a significant increase in voltage. Once conducting, it will act as a heavy load, holding the system voltage down through its large current draw and subsequent voltage drop along conductors and other series impedances. In a properly engineered system, the spark gap will stop conducting ("extinguish") when the system voltage decreases to a normal level, well below the voltage required to initiate conduction.

One major caveat of spark gaps is their significantly finite life. The discharge generated by such a device can be quite violent, and as such will tend to deteriorate the surfaces of the electrodes through pitting and/or melting.

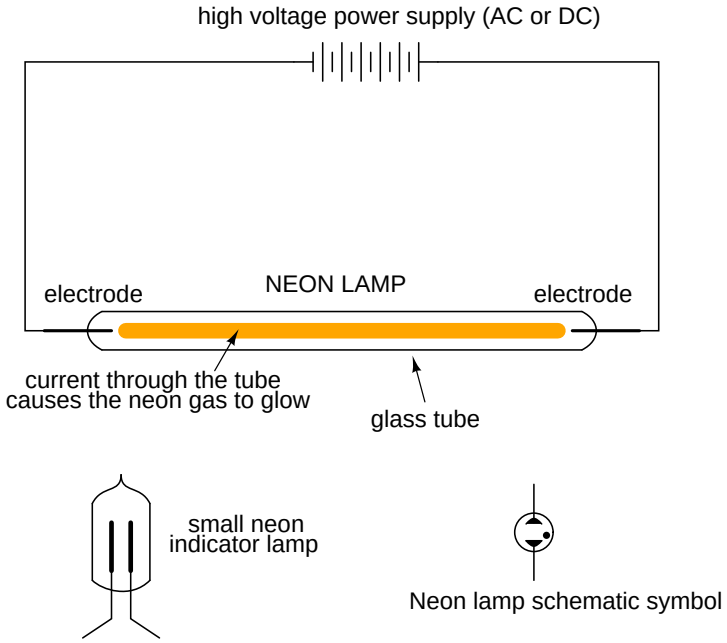
Spark gaps can be made to conduct on command by placing a third electrode (usually with a sharp edge or point) between the other two and applying a high voltage pulse between that electrode and one of the other electrodes. The pulse will create a small spark between the two electrodes, ionizing part of the pathway between the two large electrodes, and enabling conduction between them if the applied voltage is high enough:



Spark gaps of both the triggered and untriggered variety can be built to handle huge amounts of current, some even into the range of mega-amps (millions of amps)! Physical size is the primary limiting factor to the amount of current a spark gap can safely and reliably handle.

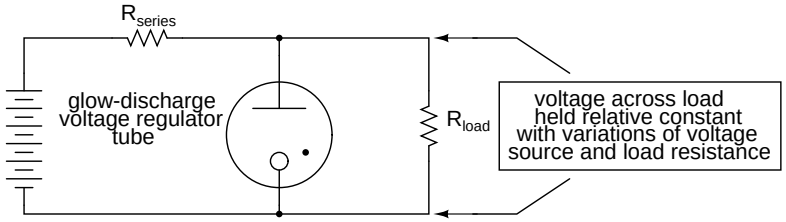
When the two main electrodes are placed in a sealed tube filled with a special gas, a *discharge tube* is formed. The most common type of discharge tube is the neon light, used popularly as a source of colorful illumination, the color of the light emitted being dependent on the type of gas filling the tube.

Construction of neon lamps closely resembles that of spark gaps, but the operational characteristics are quite different:



By controlling the spacing of the electrodes and the type of gas in the tube, neon lights can be made to conduct without drawing the excessive currents that spark gaps do. They still exhibit hysteresis in that it takes a higher voltage to initiate conduction than it does to make them "extinguish," and their resistance is definitely nonlinear (the more voltage applied across the tube, the more current, thus more heat, thus lower resistance). Given this nonlinear tendency, the voltage across a neon tube must not be allowed to exceed a certain limit, lest the tube be damaged by excessive temperatures.

This nonlinear tendency gives the neon tube an application other than colorful illumination: it can act somewhat like a zener diode, "clamping" the voltage across it by drawing more and more current if the voltage decreases. When used in this fashion, the tube is known as a *glow tube*, or *voltage-regulator tube*, and was a popular means of voltage regulation in the days of electron tube circuit design.



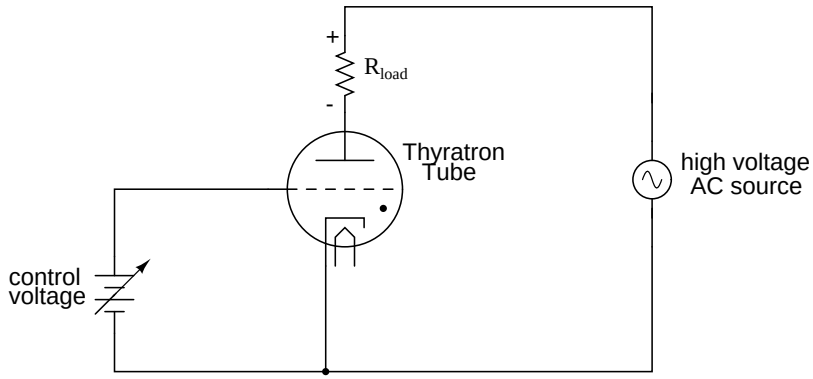
Please take note of the black dot found in the tube symbol shown above (and in the neon lamp symbol shown before that). That marker indicates the tube is gas-filled. It is a common marker used in all gas-filled tube symbols.

One example of a glow tube designed for voltage regulation was the VR-150, with a nominal regulating voltage of 150 volts. Its resistance throughout the allowable limits of current could vary

from 5 kΩ to 30 kΩ, a 6:1 span. Like zener diode regulator circuits of today, glow tube regulators could be coupled to amplifying tubes for better voltage regulation and higher load current ranges.

If a regular triode was filled with gas instead of a hard vacuum, it would manifest all the hysteresis and nonlinearity of other gas tubes with one major advantage: the amount of voltage applied between grid and cathode would determine the minimum plate-to cathode voltage necessary to initiate conduction. In essence, this tube was the equivalent of the semiconductor SCR (Silicon-Controlled Rectifier), and was called the *thyratron*.

(Simple) Thyratron control circuit



It should be noted that the schematic shown above is greatly simplified for most purposes and thyratron tube designs. Some thyratrons, for instance, required that the grid voltage switch polarity between their "on" and "off" states in order to properly work. Also, some thyratrons had more than one grid!

Thyratrons found use in much the same way as SCR's find use today: controlling rectified AC to large loads such as motors. Thyratron tubes have been manufactured with different types of gas fillings for different characteristics: inert (chemically non-reactive) gas, hydrogen gas, and mercury (vaporized into a gas form when activated). Deuterium, a rare isotope of hydrogen, was used in some special applications requiring the switching of high voltages.

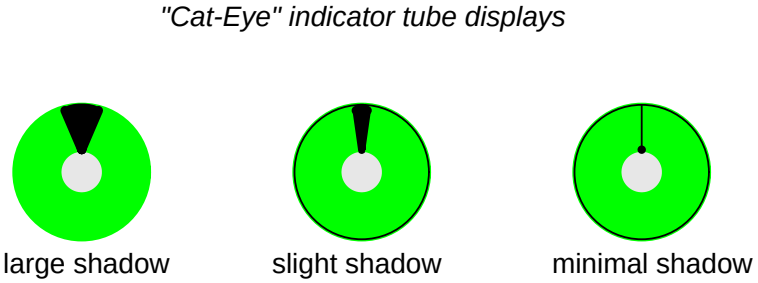
13.10 Display tubes

In addition to performing tasks of amplification and switching, tubes can be designed to serve as display devices.

Perhaps the best-known display tube is the *cathode ray tube*, or *CRT*. Originally invented as an instrument to study the behavior of "cathode rays" (electrons) in a vacuum, these tubes developed into instruments useful in detecting voltage, then later as video projection devices with the advent of television. The main difference between CRTs used in oscilloscopes and CRTs used in televisions is that the oscilloscope variety exclusively use electrostatic (plate) deflection, while televisions use electromagnetic (coil) deflection. Plates function much better than coils over a wider range of signal frequencies, which is great for oscilloscopes but irrelevant for televisions, since a television electron beam sweeps vertically and horizontally at fixed frequencies. Electromagnetic deflection coils are much preferred in television CRT construction because they do not have to penetrate the glass

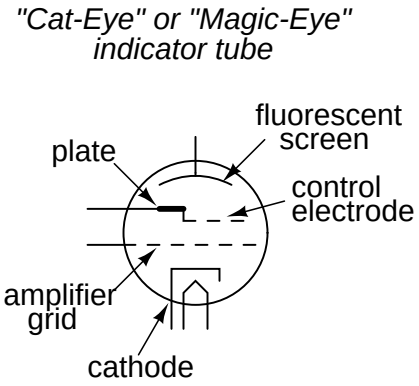
envelope of the tube, thus decreasing the production costs and increasing tube reliability.

An interesting "cousin" to the CRT is the *Cat-Eye* or *Magic-Eye* indicator tube. Essentially, this tube is a voltage-measuring device with a display resembling a glowing green ring. Electrons emitted by the cathode of this tube impinge on a fluorescent screen, causing the green-colored light to be emitted. The shape of the glow produced by the fluorescent screen varies as the amount of voltage applied to a grid changes:



The width of the shadow is directly determined by the potential difference between the control electrode and the fluorescent screen. The control electrode is a narrow rod placed between the cathode and the fluorescent screen. If that control electrode (rod) is significantly more negative than the fluorescent screen, it will deflect some electrons away from the that area of the screen. The area of the screen "shadowed" by the control electrode will appear darker when there is a significant voltage difference between the two. When the control electrode and fluorescent screen are at equal potential (zero voltage between them), the shadowing effect will be minimal and the screen will be equally illuminated.

The schematic symbol for a "cat-eye" tube looks something like this:

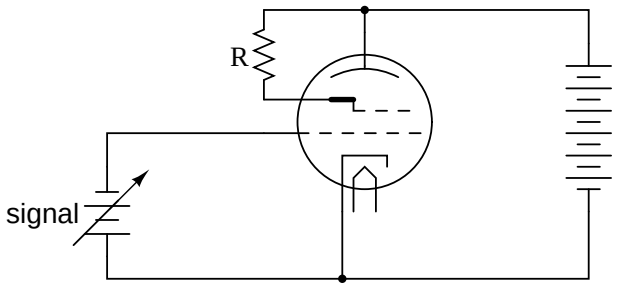


Here is a photograph of a cat-eye tube, showing the circular display region as well as the glass envelope, socket (black, at far end of tube), and some of its internal structure:

Normally, only the end of the tube would protrude from a hole in an instrument panel, so the user could view the circular, fluorescent screen.

In its simplest usage, a "cat-eye" tube could be operated without the use of the amplifier grid. However, in order to make it more sensitive, the amplifier grid *is* used, and it is used like this:

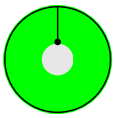
"Cat-Eye" indicator tube circuit



As the signal voltage increases, current through the tube is choked off. This decreases the voltage between the plate and the fluorescent screen, lessening the shadow effect (shadow narrows).

The cathode, amplifier grid, and plate act as a triode to create large changes in plate-to-cathode voltage for small changes in grid-to-cathode voltage. Because the control electrode is internally connected to the plate, it is electrically common to it and therefore possesses the same amount of voltage with respect to the cathode that the plate does. Thus, the large voltage changes induced on the plate due to small voltage changes on the amplifier grid end up causing large changes in the width of the shadow seen by whoever is viewing the tube.

Control electrode negative with respect to the fluorescent screen. This is caused by a positive amplifier grid voltage (with respect to the cathode).



No voltage between control electrode and flourescent screen. This is caused by a negative amplifier grid voltage (with respect to the cathode).

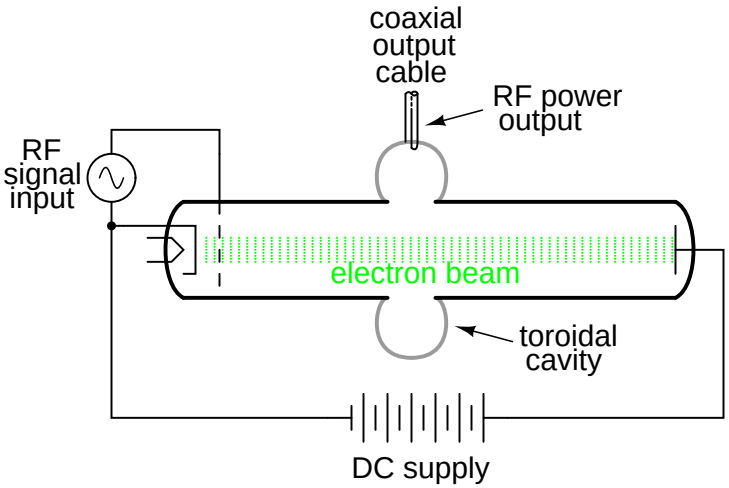
"Cat-eye" tubes were never accurate enough to be equipped with a graduated scale as is the case with CRT's and electromechanical meter movements, but they served well as null detectors in bridge circuits, and as signal strength indicators in radio tuning circuits. An unfortunate limitation to the "cat-eye" tube as a null detector was the fact that it was not directly capable of voltage indication in both polarities.

13.11 Microwave tubes

For extremely high-frequency applications (above 1 GHz), the interelectrode capacitances and transit-time delays of standard electron tube construction become prohibitive. However, there seems to be no end to the creative ways in which tubes may be constructed, and several high-frequency electron tube designs have been made to overcome these challenges.

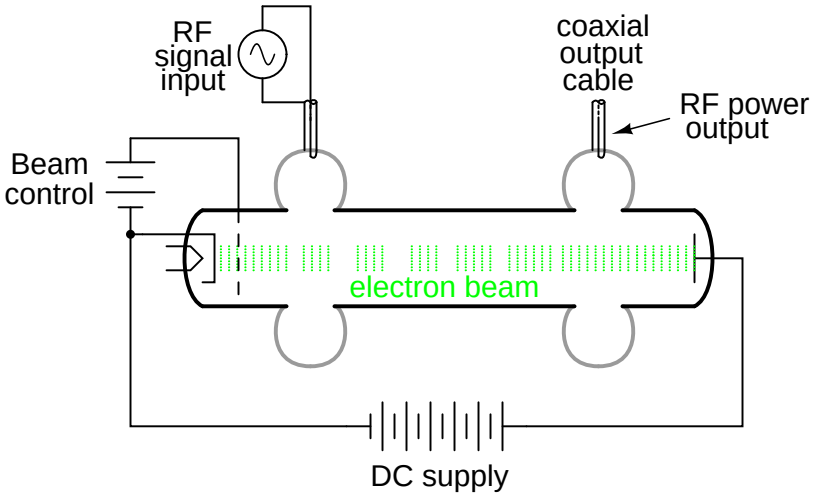
It was discovered in 1939 that a toroidal cavity made of conductive material called a *cavity resonator* surrounding an electron beam of oscillating intensity could extract power from the beam without actually intercepting the beam itself. The oscillating electric and magnetic fields associated with the beam "echoed" inside the cavity, in a manner similar to the sounds of traveling automobiles echoing in a roadside canyon, allowing radio-frequency energy to be transferred from the beam to a waveguide or coaxial cable connected to the resonator with a coupling loop. The tube was called an *inductive output tube*, or *IOT*:

The inductive output tube (IOT)



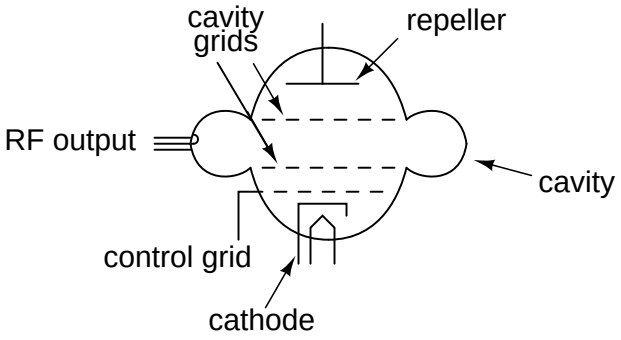
Two of the researchers instrumental in the initial development of the IOT, a pair of brothers named Sigurd and Russell Varian, added a second cavity resonator for signal input to the inductive output tube. This input resonator acted as a pair of inductive grids to alternately "bunch" and release packets of electrons down the drift space of the tube, so the electron beam would be composed of electrons traveling at different velocities. This "velocity modulation" of the beam translated into the same sort of amplitude variation at the output resonator, where energy was extracted from the beam. The Varian brothers called their invention a *klystron*.

The klystron tube



Another invention of the Varian brothers was the *reflex klystron* tube. In this tube, electrons emitted from the heated cathode travel through the cavity grids toward the repeller plate, then are repelled and returned back the way they came (hence the name *reflex*) through the cavity grids. Self-sustaining oscillations would develop in this tube, the frequency of which could be changed by adjusting the repeller voltage. Hence, this tube operated as a voltage-controlled oscillator.

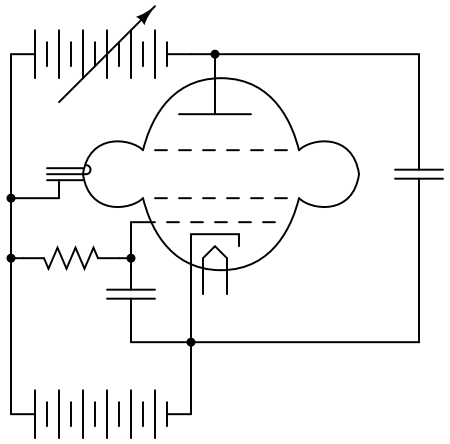
The reflex klystron tube



As a voltage-controlled oscillator, reflex klystron tubes served commonly as "local oscillators"

for radar equipment and microwave receivers:

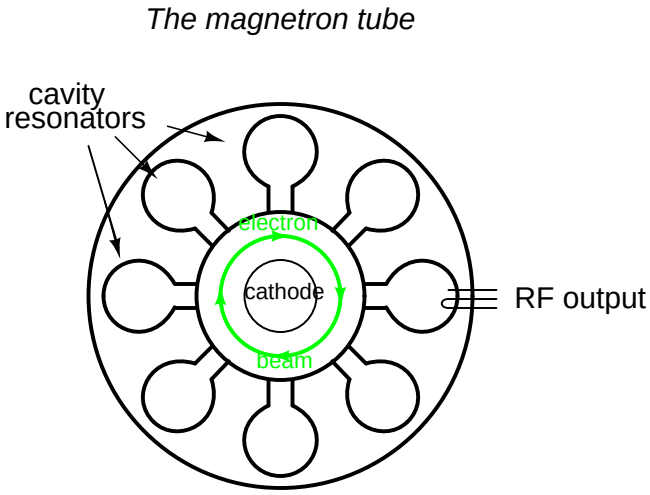
Reflex klystron tube used as a voltage-controlled oscillator



Initially developed as low-power devices whose output required further amplification for radio transmitter use, reflex klystron design was refined to the point where the tubes could serve as power devices in their own right. Reflex klystrons have since been superseded by semiconductor devices in the application of local oscillators, but amplification klystrons continue to find use in high-power, high-frequency radio transmitters and in scientific research applications.

One microwave tube performs its task so well and so cost-effectively that it continues to reign supreme in the competitive realm of consumer electronics: the magnetron tube. This device forms the heart of every microwave oven, generating several hundred watts of microwave RF energy used to heat food and beverages, and doing so under the most grueling conditions for a tube: powered on and off at random times and for random durations.

Magnetron tubes are representative of an entirely different kind of tube than the IOT and klystron. Whereas the latter tubes use a linear electron beam, the magnetron directs its electron beam in a circular pattern by means of a strong magnetic field:



Once again, cavity resonators are used as microwave-frequency "tank circuits," extracting energy from the passing electron beam inductively. Like all microwave-frequency devices using a cavity resonator, at least one of the resonator cavities is tapped with a *coupling loop*: a loop of wire magnetically coupling the coaxial cable to the resonant structure of the cavity, allowing RF power to be directed out of the tube to a load. In the case of the microwave oven, the output power is directed through a waveguide to the food or drink to be heated, the water molecules within acting as tiny load resistors, dissipating the electrical energy in the form of heat.

The magnet required for magnetron operation is not shown in the diagram. Magnetic flux runs perpendicular to the plane of the circular electron path. In other words, from the view of the tube shown in the diagram, you are looking straight at one of the magnetic poles.

13.12 Tubes versus Semiconductors

Devoting a whole chapter in a modern electronics text to the design and function of electron tubes may seem a bit strange, seeing as how semiconductor technology has all but obsoleted tubes in almost every application. However, there is merit in exploring tubes not just for historical purposes, but also for those niche applications that necessitate the qualifying phrase "*almost* every application" in regard to semiconductor supremacy.

In some applications, electron tubes not only continue to see practical use, but perform their respective tasks better than any solid-state device yet invented. In some cases the performance and reliability of electron tube technology is *far* superior.

In the fields of high-power, high-speed circuit switching, specialized tubes such as hydrogen thyratrons and krytrons are able to switch far larger amounts of current, far faster than any semiconductor device designed to date. The thermal and temporal limits of semiconductor physics place limitations on switching ability that tubes – which do not operate on the same principles – are exempt from.

In high-power microwave transmitter applications, the excellent thermal tolerance of tubes alone secures their dominance over semiconductors. Electron conduction through semiconducting materials is greatly impacted by temperature. Electron conduction through a vacuum is not. As a

consequence, the practical thermal limits of semiconductor devices are rather low compared to that of tubes. Being able to operate tubes at far greater temperatures than equivalent semiconductor devices allows tubes to dissipate more thermal energy for a given amount of dissipation area, which makes them smaller and lighter in continuous high power applications.

Another decided advantage of tubes over semiconductor components in high-power applications is their rebuildability. When a large tube fails, it may be disassembled and repaired at far lower cost than the purchase price of a new tube. When a semiconductor component fails, large or small, there is generally no means of repair.

The following photograph shows the front panel of a 1960's vintage 5 kW AM radio transmitter. One of two "Eimac" brand power tubes can be seen in a recessed area, behind the glass door. According to the station engineer who gave the facility tour, the rebuild cost for such a tube is only $800: quite inexpensive compared to the cost of a new tube, and still quite reasonable in contrast to the price of a new, comparable semiconductor component!

Tubes, being less complex in their manufacture than semiconductor components, are potentially cheaper to produce as well, although the huge volume of semiconductor device production in the world greatly offsets this theoretical advantage. Semiconductor manufacture is quite complex, involving many dangerous chemical substances and necessitating super-clean assembly environments. Tubes are essentially nothing more than glass and metal, with a vacuum seal. Physical tolerances are "loose" enough to permit hand-assembly of vacuum tubes, and the assembly work need not be done in a "clean room" environment as is necessary for semiconductor manufacture.

One modern area where electron tubes enjoy supremacy over semiconductor components is in the professional and high-end audio amplifier markets, although this is partially due to musical culture. Many professional guitar players, for example, prefer tube amplifiers over transistor amplifiers because of the specific distortion produced by tube circuits. An electric guitar amplifier is designed to *produce distortion* rather than avoid distortion as is the case with audio-reproduction amplifiers (this is why an electric guitar sounds so much different than an acoustical guitar), and the type of distortion produced by an amplifier is as much a matter of personal taste as it is technical measurement. Since rock music in particular was born with guitarists playing tube-amplifier equipment, there is a significant level of "tube appeal" inherent to the genre itself, and this appeal shows itself in the continuing demand for "tubed" guitar amplifiers among rock guitarists.

As an illustration of the attitude among some guitarists, consider the following quote taken from the technical glossary page of a tube-amplifier website which will remain nameless:

> **Solid State:** *A component that has been specifically designed to make a guitar amplifier sound bad. Compared to tubes, these devices can have a very long lifespan, which guarantees that your amplifier will retain its thin, lifeless, and buzzy sound for a long time to come.*

In the area of audio reproduction amplifiers (music studio amplifiers and home entertainment amplifiers), it is best for an amplifier to reproduce the musical signal with as *little* distortion as possible. Paradoxically, in contrast to the guitar amplifier market where distortion is a design goal, high-end audio is another area where tube amplifiers enjoy continuing consumer demand. Though one might suppose the objective, technical requirement of low distortion would eliminate any subjective bias on the part of audiophiles, one would be very wrong. The market for high-end "tubed" amplifier equipment is quite volatile, changing rapidly with trends and fads, driven by highly subjective claims of "magical" sound from audio system reviewers and salespeople. As in the electric guitar world, there is no small measure of cult-like devotion to tube amplifiers among some quarters of the audiophile world. As an example of this irrationality, consider the design of many ultra-high-end amplifiers, with chassis built to display the working tubes openly, even though this physical exposure of the tubes obviously enhances the undesirable effect of *microphonics* (changes in tube performance as a result of sound waves vibrating the tube structure).

Having said this, though, there is a wealth of technical literature contrasting tubes against semiconductors for audio power amplifier use, especially in the area of distortion analysis. More than a few competent electrical engineers prefer tube amplifier designs over transistors, and are able to produce experimental evidence in support of their choice. The primary difficulty in quantifying audio system performance is the uncertain response of human hearing. *All* amplifiers distort their input signal to some degree, especially when overloaded, so the question is which type of amplifier design distorts the least. However, since human hearing is very nonlinear, people do not interpret all types of acoustic distortion equally, and so some amplifiers will sound "better" than others even if a quantitative distortion analysis with electronic instruments indicates similar distortion levels. To determine what type of audio amplifier will distort a musical signal "the least," we must regard the human ear and brain as part of the whole acoustical system. Since no complete model yet exists for human auditory response, objective assessment is difficult at best. However, some research indicates that the characteristic distortion of tube amplifier circuits (especially when overloaded) is less objectionable than distortion produced by transistors.

Tubes also possess the distinct advantage of low "drift" over a wide range of operating conditions. Unlike semiconductor components, whose barrier voltages, β ratios, bulk resistances, and junction capacitances may change substantially with changes in device temperature and/or other operating conditions, the fundamental characteristics of a vacuum tube remain nearly constant over a wide range in operating conditions, because those characteristics are determined primarily by the physical dimensions of the tube's structural elements (cathode, grid(s), and plate) rather than the interactions of subatomic particles in a crystalline lattice.

This is one of the major reasons solid-state amplifier designers typically engineer their circuits to maximize power-efficiency even when it compromises distortion performance, because a power-inefficient amplifier dissipates a lot of energy in the form of waste heat, and transistor characteristics tend to change substantially with temperature. Temperature-induced "drift" makes it difficult to stabilize "Q" points and other important performance-related measures in an amplifier circuit. Unfortunately, power efficiency and low distortion seem to be mutually exclusive design goals.

For example, class A audio amplifier circuits typically exhibit very low distortion levels, but are

very wasteful of power, meaning that it would be difficult to engineer a solid-state class A amplifier of any substantial power rating due to the consequent drift of transistor characteristics. Thus, most solid-state audio amplifier designers choose class B circuit configurations for greater efficiency, even though class B designs are notorious for producing a type of distortion known as *crossover distortion*. However, with tubes it is easy to design a stable class A audio amplifier circuit because tubes are not as adversely affected by the changes in temperature experienced in a such a power-inefficient circuit configuration.

Tube performance parameters, though, tend to "drift" more than semiconductor devices when measured over long periods of time (years). One major mechanism of tube "aging" appears to be vacuum leaks: when air enters the inside of a vacuum tube, its electrical characteristics become irreversibly altered. This same phenomenon is a major cause of tube mortality, or why tubes typically do not last as long as their respective solid-state counterparts. When tube vacuum is maintained at a high level, though, excellent performance and life is possible. An example of this is a klystron tube (used to produce the high-frequency radio waves used in a radar system) that lasted for 240,000 hours of operation (cited by Robert S. Symons of Litton Electron Devices Division in his informative paper, "Tubes: Still vital after all these years," printed in the April 1998 issue of *IEEE Spectrum* magazine).

If nothing else, the tension between audiophiles over tubes versus semiconductors has spurred a remarkable degree of experimentation and technical innovation, serving as an excellent resource for those wishing to educate themselves on amplifier theory. Taking a wider view, the versatility of electron tube technology (different physical configurations, multiple control grids) hints at the potential for circuit designs of far greater variety than is possible using semiconductors. For this and other reasons, electron tubes will never be "obsolete," but will continue to serve in niche roles, and to foster innovation for those electronics engineers, inventors, and hobbyists who are unwilling to let their minds by stifled by convention.

Appendix A-1

ABOUT THIS BOOK

A-1.1 Purpose

They say that necessity is the mother of invention. At least in the case of this book, that adage is true. As an industrial electronics instructor, I was forced to use a sub-standard textbook during my first year of teaching. My students were daily frustrated with the many typographical errors and obscure explanations in this book, having spent much time at home struggling to comprehend the material within. Worse yet were the many incorrect answers in the back of the book to selected problems. Adding insult to injury was the $100+ price.

Contacting the publisher proved to be an exercise in futility. Even though the particular text I was using had been in print and in popular use for a couple of years, they claimed my complaint was the first they'd ever heard. My request to review the draft for the next edition of their book was met with disinterest on their part, and I resolved to find an alternative text.

Finding a suitable alternative was more difficult than I had imagined. Sure, there were plenty of texts in print, but the really good books seemed a bit too heavy on the math and the less intimidating books omitted a lot of information I felt was important. Some of the best books were out of print, and those that were still being printed were quite expensive.

It was out of frustration that I compiled *Lessons in Electric Circuits* from notes and ideas I had been collecting for years. My primary goal was to put readable, high-quality information into the hands of my students, but a secondary goal was to make the book as affordable as possible. Over the years, I had experienced the benefit of receiving free instruction and encouragement in my pursuit of learning electronics from many people, including several teachers of mine in elementary and high school. Their selfless assistance played a key role in my own studies, paving the way for a rewarding career and fascinating hobby. If only I could extend the gift of their help by giving to other people what they gave to me . . .

So, I decided to make the book freely available. More than that, I decided to make it "open," following the same development model used in the making of free software (most notably the various UNIX utilities released by the Free Software Foundation, and the Linux operating system, whose fame is growing even as I write). The goal was to copyright the text – so as to protect my authorship

– but expressly allow anyone to distribute and/or modify the text to suit their own needs with a minimum of legal encumbrance. This willful and formal revoking of standard distribution limitations under copyright is whimsically termed *copyleft*. Anyone can "copyleft" their creative work simply by appending a notice to that effect on their work, but several Licenses already exist, covering the fine legal points in great detail.

The first such License I applied to my work was the GPL – General Public License – of the Free Software Foundation (GNU). The GPL, however, is intended to copyleft works of computer software, and although its introductory language is broad enough to cover works of text, its wording is not as clear as it could be for that application. When other, less specific copyleft Licenses began appearing within the free software community, I chose one of them (the Design Science License, or DSL) as the official notice for my project.

In "copylefting" this text, I guaranteed that no instructor would be limited by a text insufficient for their needs, as I had been with error-ridden textbooks from major publishers. I'm sure this book in its initial form will not satisfy everyone, but anyone has the freedom to change it, leveraging my efforts to suit variant and individual requirements. For the beginning student of electronics, learn what you can from this book, editing it as you feel necessary if you come across a useful piece of information. Then, if you pass it on to someone else, you will be giving them something better than what you received. For the instructor or electronics professional, feel free to use this as a reference manual, adding or editing to your heart's content. The only "catch" is this: if you plan to distribute your modified version of this text, you must give credit where credit is due (to me, the original author, and anyone else whose modifications are contained in your version), and you must ensure that whoever you give the text to is aware of their freedom to similarly share and edit the text. The next chapter covers this process in more detail.

It must be mentioned that although I strive to maintain technical accuracy in all of this book's content, the subject matter is broad and harbors many potential dangers. Electricity maims and kills without provocation, and deserves the utmost respect. I strongly encourage experimentation on the part of the reader, but only with circuits powered by small batteries where there is no risk of electric shock, fire, explosion, etc. High-power electric circuits should be left to the care of trained professionals! The Design Science License clearly states that neither I nor any contributors to this book bear any liability for what is done with its contents.

A-1.2 The use of SPICE

One of the best ways to learn how things work is to follow the inductive approach: to observe specific instances of things working and derive general conclusions from those observations. In science education, labwork is the traditionally accepted venue for this type of learning, although in many cases labs are designed by educators to reinforce principles previously learned through lecture or textbook reading, rather than to allow the student to learn on their own through a truly exploratory process.

Having taught myself most of the electronics that I know, I appreciate the sense of frustration students may have in teaching themselves from books. Although electronic components are typically inexpensive, not everyone has the means or opportunity to set up a laboratory in their own homes, and when things go wrong there's no one to ask for help. Most textbooks seem to approach the task of education from a deductive perspective: tell the student how things are supposed to work, then apply those principles to specific instances that the student may or may not be able to explore by

themselves. The inductive approach, as useful as it is, is hard to find in the pages of a book.

However, textbooks don't have to be this way. I discovered this when I started to learn a computer program called SPICE. It is a text-based piece of software intended to model circuits and provide analyses of voltage, current, frequency, etc. Although nothing is quite as good as building real circuits to gain knowledge in electronics, computer simulation is an excellent alternative. In learning how to use this powerful tool, I made a discovery: SPICE could be used within a textbook to present circuit simulations to allow students to "observe" the phenomena for themselves. This way, the readers could learn the concepts inductively (by interpreting SPICE's output) as well as deductively (by interpreting my explanations). Furthermore, in seeing SPICE used over and over again, they should be able to understand how to use it themselves, providing a perfectly safe means of experimentation on their own computers with circuit simulations of their own design.

Another advantage to including computer analyses in a textbook is the empirical verification it adds to the concepts presented. Without demonstrations, the reader is left to take the author's statements on faith, trusting that what has been written is indeed accurate. The problem with faith, of course, is that it is only as good as the authority in which it is placed and the accuracy of interpretation through which it is understood. Authors, like all human beings, are liable to err and/or communicate poorly. With demonstrations, however, the reader can immediately see for themselves that what the author describes is indeed true. Demonstrations also serve to clarify the meaning of the text with concrete examples.

SPICE is introduced early in volume I (DC) of this book series, and hopefully in a gentle enough way that it doesn't create confusion. For those wishing to learn more, a chapter in the Reference volume (volume V) contains an overview of SPICE with many example circuits. There may be more flashy (graphic) circuit simulation programs in existence, but SPICE is free, a virtue complementing the charitable philosophy of this book very nicely.

A-1.3 Acknowledgements

First, I wish to thank my wife, whose patience during those many and long evenings (and weekends!) of typing has been extraordinary.

I also wish to thank those whose open-source software development efforts have made this endeavor all the more affordable and pleasurable. The following is a list of various free computer software used to make this book, and the respective programmers:

- *GNU/Linux* Operating System – Linus Torvalds, Richard Stallman, and a host of others too numerous to mention.
- *Vim* text editor – Bram Moolenaar and others.
- *Xcircuit* drafting program – Tim Edwards.
- *SPICE* circuit simulation program – too many contributors to mention.
- TEX text processing system – Donald Knuth and others.
- *Texinfo* document formatting system – Free Software Foundation.
- LATEX document formatting system – Leslie Lamport and others.

- *Gimp* image manipulation program – too many contributors to mention.

Appreciation is also extended to Robert L. Boylestad, whose first edition of *Introductory Circuit Analysis* taught me more about electric circuits than any other book. Other important texts in my electronics studies include the 1939 edition of *The "Radio" Handbook*, Bernard Grob's second edition of *Introduction to Electronics I*, and Forrest Mims' original *Engineer's Notebook*.

Thanks to the staff of the Bellingham Antique Radio Museum, who were generous enough to let me terrorize their establishment with my camera and flash unit.

I wish to specifically thank Jeffrey Elkner and all those at Yorktown High School for being willing to host my book as part of their Open Book Project, and to make the first effort in contributing to its form and content. Thanks also to David Sweet (website: `(http://www.andamooka.org)`) and Ben Crowell (website: `(http://www.lightandmatter.com)`) for providing encouragement, constructive criticism, and a wider audience for the online version of this book.

Thanks to Michael Stutz for drafting his Design Science License, and to Richard Stallman for pioneering the concept of copyleft.

Last but certainly not least, many thanks to my parents and those teachers of mine who saw in me a desire to learn about electricity, and who kindled that flame into a passion for discovery and intellectual adventure. I honor you by helping others as you have helped me.

Tony Kuphaldt, July 2001

"A candle loses nothing of its light when lighting another"
Kahlil Gibran

Appendix A-2

CONTRIBUTOR LIST

A-2.1 How to contribute to this book

As a copylefted work, this book is open to revision and expansion by any interested parties. The only "catch" is that credit must be given where credit is due. This *is* a copyrighted work: it is *not* in the public domain!

If you wish to cite portions of this book in a work of your own, you must follow the same guidelines as for any other copyrighted work. Here is a sample from the Design Science License:

```
The Work is copyright the Author.  All rights to the Work are reserved
by the Author, except as specifically described below. This License
describes the terms and conditions under which the Author permits you
to copy, distribute and modify copies of the Work.

In addition, you may refer to the Work, talk about it, and (as
dictated by "fair use") quote from it, just as you would any
copyrighted material under copyright law.

Your right to operate, perform, read or otherwise interpret and/or
execute the Work is unrestricted; however, you do so at your own risk,
because the Work comes WITHOUT ANY WARRANTY -- see Section 7 ("NO
WARRANTY") below.
```

If you wish to modify this book in any way, you must document the nature of those modifications in the "Credits" section along with your name, and ideally, information concerning how you may be contacted. Again, the Design Science License:

```
Permission is granted to modify or sample from a copy of the Work,
producing a derivative work, and to distribute the derivative work
under the terms described in the section for distribution above,
```

```
provided that the following terms are met:

(a) The new, derivative work is published under the terms of this
License.

(b) The derivative work is given a new name, so that its name or
title can not be confused with the Work, or with a version of
the Work, in any way.

(c) Appropriate authorship credit is given: for the differences
between the Work and the new derivative work, authorship is
attributed to you, while the material sampled or used from
the Work remains attributed to the original Author; appropriate
notice must be included with the new work indicating the nature
and the dates of any modifications of the Work made by you.
```

Given the complexities and security issues surrounding the maintenance of files comprising this book, it is recommended that you submit any revisions or expansions to the original author (Tony R. Kuphaldt). You are, of course, welcome to modify this book directly by editing your own personal copy, but we would all stand to benefit from your contributions if your ideas were incorporated into the online "master copy" where all the world can see it.

A-2.2 Credits

All entries arranged in alphabetical order of surname. Major contributions are listed by individual name with some detail on the nature of the contribution(s), date, contact info, etc. Minor contributions (typo corrections, etc.) are listed by name only for reasons of brevity. Please understand that when I classify a contribution as "minor," it is in no way inferior to the effort or value of a "major" contribution, just smaller in the sense of less text changed. Any and all contributions are gratefully accepted. I am indebted to all those who have given freely of their own knowledge, time, and resources to make this a better book!

A-2.2.1 Dennis Crunkilton

- **Date(s) of contribution(s):** May 2005 to present
- **Nature of contribution:** Nutmeg-spice plots added to chapters 4,5, & 6; 04/2006
- **Nature of contribution:** Mini table of contents, all chapters except appedicies; html, latex, ps, pdf; See Devel/tutorial.html; 01/2006.
- **Contact at:** `dcrunkilton(at)att(dot)net`

A-2.2.2 Tony R. Kuphaldt

- **Date(s) of contribution(s):** 1996 to present
- **Nature of contribution:** Original author.
- **Contact at:** `liec0@lycos.com`

A-2.2.3 Warren Young

- **Date(s) of contribution(s):** August 2002
- **Nature of contribution:** Provided initial text for "Power supply circuits" section of chapter 9.

A-2.2.4 Your name here

- **Date(s) of contribution(s):** Month and year of contribution
- **Nature of contribution:** Insert text here, describing how you contributed to the book.
- **Contact at:** `my_email@provider.net`

A-2.2.5 Typo corrections and other "minor" contributions

- **line-allaboutcircuits.com** (June 2005) Typographical error correction in Volumes 1,2,3,5, various chapters ,(:s/visa-versa/vice versa/).
- *The students of Bellingham Technical College's Instrumentation program.*
- **Clifford Bailey** (July 2005) Correction, Ch 1.4, s/V/I equations in 13026.eps.
- **Colin Barnard** (November 2003) Correction on the nationality of Alexander Graham Bell.
- **Jeff DeFreitas** (March 2006)Improve appearance: replace "/" and "/" Chapters: A1, A2.
- **Sean Donner** (April 2005) Typographical error correction in "Voltage and current" section, Chapter 4: BIPOLAR JUNCTION TRANSISTORS,(by a the/ by the) (no longer loaded/ no longer unloaded).
- **Sulev Eesmaa** (February 2006) Correction, Ch 4, beta by definition- collector and emitter/ collector and base//.
- **Bill Heath** (September-December 2002) Correction on illustration of atomic structure, and corrections of several typographical errors.
- **Maciej Noszczyski** (December 2003) Corrected spelling of Niels Bohr's name.
- **Don Stalkowski** (June 2002) Technical help with PostScript-to-PDF file format conversion.
- **Joseph Teichman** (June 2002) Suggestion and technical help regarding use of PNG images instead of JPEG.

- **Jered Wierzbicki** (December 2002) Correction on diode equation: Boltzmann's constant shown incorrectly.

Appendix A-3

DESIGN SCIENCE LICENSE

A-3.1 0. Preamble

Copyright law gives certain exclusive rights to the author of a work, including the rights to copy, modify and distribute the work (the "reproductive," "adaptative," and "distribution" rights).

The idea of "copyleft" is to willfully revoke the exclusivity of those rights under certain terms and conditions, so that anyone can copy and distribute the work or properly attributed derivative works, while all copies remain under the same terms and conditions as the original.

The intent of this license is to be a general "copyleft" that can be applied to any kind of work that has protection under copyright. This license states those certain conditions under which a work published under its terms may be copied, distributed, and modified.

Whereas "design science" is a strategy for the development of artifacts as a way to reform the environment (not people) and subsequently improve the universal standard of living, this Design Science License was written and deployed as a strategy for promoting the progress of science and art through reform of the environment.

A-3.2 1. Definitions

"License" shall mean this Design Science License. The License applies to any work which contains a notice placed by the work's copyright holder stating that it is published under the terms of this Design Science License.

"Work" shall mean such an aforementioned work. The License also applies to the output of the Work, only if said output constitutes a "derivative work" of the licensed Work as defined by copyright law.

"Object Form" shall mean an executable or performable form of the Work, being an embodiment of the Work in some tangible medium.

"Source Data" shall mean the origin of the Object Form, being the entire, machine-readable, preferred form of the Work for copying and for human modification (usually the language, encoding or format in which composed or recorded by the Author); plus any accompanying files, scripts or other data necessary for installation, configuration or compilation of the Work.

(Examples of "Source Data" include, but are not limited to, the following: if the Work is an image file composed and edited in 'PNG' format, then the original PNG source file is the Source Data; if the Work is an MPEG 1.0 layer 3 digital audio recording made from a 'WAV' format audio file recording of an analog source, then the original WAV file is the Source Data; if the Work was composed as an unformatted plaintext file, then that file is the the Source Data; if the Work was composed in LaTeX, the LaTeX file(s) and any image files and/or custom macros necessary for compilation constitute the Source Data.)

"Author" shall mean the copyright holder(s) of the Work.

The individual licensees are referred to as "you."

A-3.3 2. Rights and copyright

The Work is copyright the Author. All rights to the Work are reserved by the Author, except as specifically described below. This License describes the terms and conditions under which the Author permits you to copy, distribute and modify copies of the Work.

In addition, you may refer to the Work, talk about it, and (as dictated by "fair use") quote from it, just as you would any copyrighted material under copyright law.

Your right to operate, perform, read or otherwise interpret and/or execute the Work is unrestricted; however, you do so at your own risk, because the Work comes WITHOUT ANY WARRANTY – see Section 7 ("NO WARRANTY") below.

A-3.4 3. Copying and distribution

Permission is granted to distribute, publish or otherwise present verbatim copies of the entire Source Data of the Work, in any medium, provided that full copyright notice and disclaimer of warranty, where applicable, is conspicuously published on all copies, and a copy of this License is distributed along with the Work.

Permission is granted to distribute, publish or otherwise present copies of the Object Form of the Work, in any medium, under the terms for distribution of Source Data above and also provided that one of the following additional conditions are met:

(a) The Source Data is included in the same distribution, distributed under the terms of this License; or

(b) A written offer is included with the distribution, valid for at least three years or for as long as the distribution is in print (whichever is longer), with a publicly-accessible address (such as a URL on the Internet) where, for a charge not greater than transportation and media costs, anyone may receive a copy of the Source Data of the Work distributed according to the section above; or

(c) A third party's written offer for obtaining the Source Data at no cost, as described in paragraph (b) above, is included with the distribution. This option is valid only if you are a non-commercial party, and only if you received the Object Form of the Work along with such an offer.

You may copy and distribute the Work either gratis or for a fee, and if desired, you may offer warranty protection for the Work.

The aggregation of the Work with other works which are not based on the Work – such as but not limited to inclusion in a publication, broadcast, compilation, or other media – does not bring the other works in the scope of the License; nor does such aggregation void the terms of the License for the Work.

A-3.5 4. Modification

Permission is granted to modify or sample from a copy of the Work, producing a derivative work, and to distribute the derivative work under the terms described in the section for distribution above, provided that the following terms are met:

(a) The new, derivative work is published under the terms of this License.

(b) The derivative work is given a new name, so that its name or title can not be confused with the Work, or with a version of the Work, in any way.

(c) Appropriate authorship credit is given: for the differences between the Work and the new derivative work, authorship is attributed to you, while the material sampled or used from the Work remains attributed to the original Author; appropriate notice must be included with the new work indicating the nature and the dates of any modifications of the Work made by you.

A-3.6 5. No restrictions

You may not impose any further restrictions on the Work or any of its derivative works beyond those restrictions described in this License.

A-3.7 6. Acceptance

Copying, distributing or modifying the Work (including but not limited to sampling from the Work in a new work) indicates acceptance of these terms. If you do not follow the terms of this License, any rights granted to you by the License are null and void. The copying, distribution or modification of the Work outside of the terms described in this License is expressly prohibited by law.

If for any reason, conditions are imposed on you that forbid you to fulfill the conditions of this License, you may not copy, distribute or modify the Work at all.

If any part of this License is found to be in conflict with the law, that part shall be interpreted in its broadest meaning consistent with the law, and no other parts of the License shall be affected.

A-3.8 7. No warranty

THE WORK IS PROVIDED "AS IS," AND COMES WITH ABSOLUTELY NO WARRANTY, EXPRESS OR IMPLIED, TO THE EXTENT PERMITTED BY APPLICABLE LAW, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

A-3.9 8. Disclaimer of liability

IN NO EVENT SHALL THE AUTHOR OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

END OF TERMS AND CONDITIONS

```
[$Id:  dsl.txt,v 1.25 2000/03/14 13:14:14 m Exp m $]
```

Index