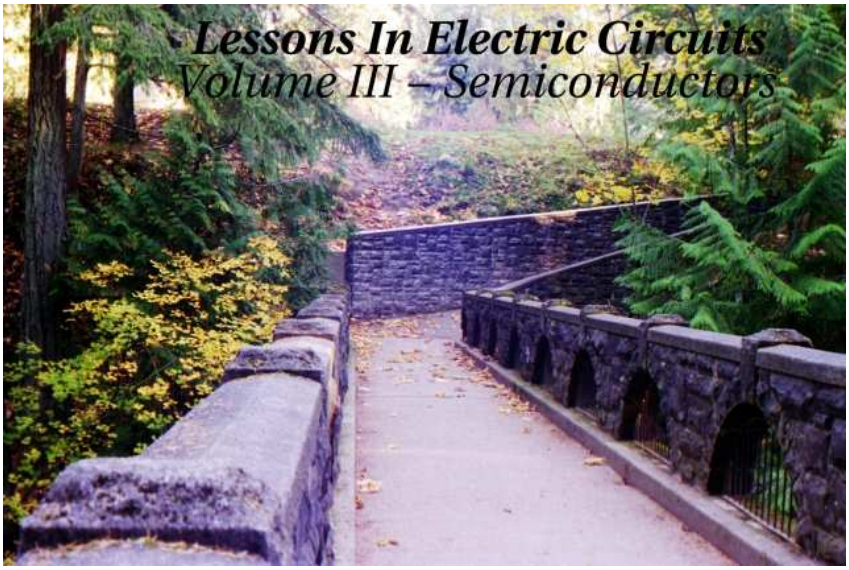


Fifth Edition, last update November 06, 2021

Lessons In Electric Circuits, Volume III – Semiconductors

By Tony R. Kuphaldt

Fifth Edition, last update November 06, 2021

©2000-2023, Tony R. Kuphaldt

This book is published under the terms and conditions of the Creative Commons License. These terms and conditions allow for free copying, distribution, and/or modification of this document by the general public. The full Creative Commons License text is included in the last chapter.

As an open and collaboratively developed text, this book is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the Creative Commons License for more details.

Available in its entirety as part of the Open Book Project collection at:

https://www.ibiblio.org/kuphaldt/electricCircuits

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.
- Fith Edition: Printed in July 2007. New sections added, and error corrections made, format change.

Contents

1 AMPLIFIERS AND ACTIVE DEVICES **1**

1.1 From electric to electronic . . . 1
1.2 Active versus passive devices . . . 3
1.3 Amplifiers . . . 3
1.4 Amplifier gain . . . 6
1.5 Decibels . . . 8
1.6 Absolute dB scales . . . 14
1.7 Attenuators . . . 16

2 SOLID-STATE DEVICE THEORY **27**

2.1 Introduction . . . 27
2.2 Quantum physics . . . 28
2.3 Valence and Crystal structure . . . 41
2.4 Band theory of solids . . . 47
2.5 Electrons and "holes" . . . 50
2.6 The P-N junction . . . 55
2.7 Junction diodes . . . 58
2.8 Bipolar junction transistors . . . 60
2.9 Junction field-effect transistors . . . 65
2.10 Insulated-gate field-effect transistors (MOSFET) . . . 70
2.11 Thyristors . . . 73
2.12 Semiconductor manufacturing techniques . . . 75
2.13 Superconducting devices . . . 80
2.14 Quantum devices . . . 83
2.15 Semiconductor devices in SPICE . . . 91
Bibliography . . . 93

3 DIODES AND RECTIFIERS **97**

3.1 Introduction . . . 98
3.2 Meter check of a diode . . . 103
3.3 Diode ratings . . . 107
3.4 Rectifier circuits . . . 108
3.5 Peak detector . . . 115
3.6 Clipper circuits . . . 117

3.7 Clamper circuits . . . 121
3.8 Voltage multipliers . . . 123
3.9 Inductor commutating circuits . . . 130
3.10 Diode switching circuits . . . 132
3.11 Zener diodes . . . 134
3.12 Special-purpose diodes . . . 143
3.13 Other diode technologies . . . 163
3.14 SPICE models . . . 163
Bibliography . . . 171

4 BIPOLAR JUNCTION TRANSISTORS 173
4.1 Introduction . . . 174
4.2 The transistor as a switch . . . 176
4.3 Meter check of a transistor . . . 179
4.4 Active mode operation . . . 183
4.5 The common-emitter amplifier . . . 189
4.6 The common-collector amplifier . . . 202
4.7 The common-base amplifier . . . 210
4.8 The cascode amplifier . . . 218
4.9 Biasing techniques . . . 222
4.10 Biasing calculations . . . 235
4.11 Input and output coupling . . . 247
4.12 Feedback . . . 256
4.13 Amplifier impedances . . . 263
4.14 Current mirrors . . . 264
4.15 Transistor ratings and packages . . . 269
4.16 BJT quirks . . . 271
Bibliography . . . 277

5 JUNCTION FIELD-EFFECT TRANSISTORS 279
5.1 Introduction . . . 279
5.2 The transistor as a switch . . . 281
5.3 Meter check of a transistor . . . 284
5.4 Active-mode operation . . . 286
5.5 The common-source amplifier – PENDING . . . 295
5.6 The common-drain amplifier – PENDING . . . 296
5.7 The common-gate amplifier – PENDING . . . 296
5.8 Biasing techniques – PENDING . . . 296
5.9 Transistor ratings and packages – PENDING . . . 297
5.10 JFET quirks – PENDING . . . 297

6 INSULATED-GATE FIELD-EFFECT TRANSISTORS 299
6.1 Introduction . . . 299
6.2 Depletion-type IGFETs . . . 300
6.3 Enhancement-type IGFETs – PENDING . . . 309
6.4 Active-mode operation – PENDING . . . 309

6.5 The common-source amplifier – PENDING . . . 310
6.6 The common-drain amplifier – PENDING . . . 310
6.7 The common-gate amplifier – PENDING . . . 310
6.8 Biasing techniques – PENDING . . . 310
6.9 Transistor ratings and packages – PENDING . . . 310
6.10 IGFET quirks – PENDING . . . 311
6.11 MESFETs – PENDING . . . 311
6.12 IGBTs . . . 311

7 THYRISTORS 315
7.1 Hysteresis . . . 315
7.2 Gas discharge tubes . . . 316
7.3 The Shockley Diode . . . 320
7.4 The DIAC . . . 326
7.5 The Silicon-Controlled Rectifier (SCR) . . . 327
7.6 The TRIAC . . . 339
7.7 Optothyristors . . . 342
7.8 The Unijunction Transistor (UJT) . . . 342
7.9 The Silicon-Controlled Switch (SCS) . . . 348
7.10 Field-effect-controlled thyristors . . . 350
Bibliography . . . 352

8 OPERATIONAL AMPLIFIERS 353
8.1 Introduction . . . 353
8.2 Single-ended and differential amplifiers . . . 354
8.3 The ”operational” amplifier . . . 358
8.4 Negative feedback . . . 364
8.5 Divided feedback . . . 367
8.6 An analogy for divided feedback . . . 370
8.7 Voltage-to-current signal conversion . . . 376
8.8 Averager and summer circuits . . . 378
8.9 Building a differential amplifier . . . 380
8.10 The instrumentation amplifier . . . 382
8.11 Differentiator and integrator circuits . . . 383
8.12 Positive feedback . . . 386
8.13 Practical considerations . . . 390
8.14 Operational amplifier models . . . 404
8.15 Data . . . 411

9 PRACTICAL ANALOG SEMICONDUCTOR CIRCUITS 413
9.1 ElectroStatic Discharge . . . 413
9.2 Power Supply circuits . . . 419
9.3 Amplifier circuits – PENDING . . . 432
9.4 Oscillator circuits – INCOMPLETE . . . 433
9.5 Phase-locked loops – PENDING . . . 434
9.6 Radio circuits – INCOMPLETE . . . 434

9.7 Computational circuits 444
9.8 Measurement circuits – INCOMPLETE 465
9.9 Control circuits – PENDING 466
Bibliography 466

10 ACTIVE FILTERS **469**
10.1 Two pole active filters 469

11 DC MOTOR DRIVES **471**
11.1 Pulse Width Modulation 471

12 INVERTERS AND AC MOTOR DRIVES **475**

13 ELECTRON TUBES **477**
13.1 Introduction 477
13.2 Early tube history 478
13.3 The triode 481
13.4 The tetrode 483
13.5 Beam power tubes 484
13.6 The pentode 485
13.7 Combination tubes 486
13.8 Tube parameters 489
13.9 Ionization (gas-filled) tubes 491
13.10 Display tubes 495
13.11 Microwave tubes 498
13.12 Tubes versus Semiconductors 501

A-1 ABOUT THIS BOOK **505**

A-2 CONTRIBUTOR LIST **509**

A-3 CC BY License **517**

INDEX **520**

Chapter 1

AMPLIFIERS AND ACTIVE DEVICES

Contents

1.1 From electric to electronic 1
1.2 Active versus passive devices 3
1.3 Amplifiers 3
1.4 Amplifier gain 6
1.5 Decibels 8
1.6 Absolute dB scales 14
1.7 Attenuators 16
1.7.1 Decibels 17
1.7.2 T-section attenuator 19
1.7.3 PI-section attenuator 20
1.7.4 L-section attenuator 21
1.7.5 Bridged T attenuator 21
1.7.6 Cascaded sections 23
1.7.7 RF attenuators 23

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*.

The historic precursor to the modern electronics era was invented by Thomas Edison in 1880 while developing the electric incandescent lamp. Edison found that a small current passed from the heated lamp filament to a metal plate mounted inside the vacuum envelop. (Figure 1.1 (a)) Today this is known as the "Edison effect". Note that the battery is only necessary to heat the filament. Electrons would still flow if a non-electrical heat source was used.

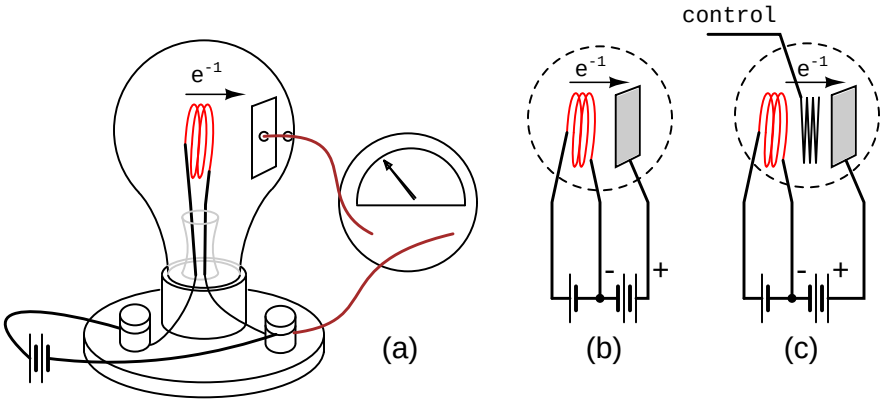


Figure 1.1: *(a) Edison effect, (b) Fleming valve or vacuum diode, (c) DeForest audion triode vacuum tube amplifier.*

By 1904 Marconi Wireless Company adviser John Flemming found that an externally applied current (plate battery) only passed in one direction from filament to plate (Figure 1.1 (b)), but not the reverse direction (not shown). This invention was the vacuum diode, used to convert alternating currents to DC. The addition of a third electrode by Lee DeForest (Figure 1.1 (c)) allowed a small signal to control the larger electron flow from filament to plate.

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 elec-

trons 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

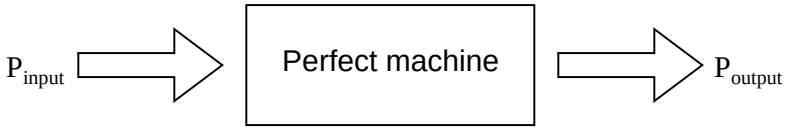
The practical benefit of active devices is their *amplifying* ability. Whether the device in question be voltage-controlled or current-controlled, the amount of power required of the controlling signal is typically far less than the amount of power available in the controlled current. In other words, 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 as in Figure 1.2.

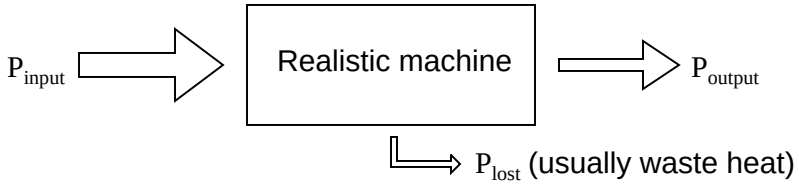
Usually, machines fail even to meet this limit, losing some of their input energy in the form of heat which is radiated into surrounding space and therefore not part of the output energy stream. (Figure 1.3)

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



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

Figure 1.2: *The power output of a machine can approach, but never exceed, the power input for 100% efficiency as an upper limit.*



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

Figure 1.3: *A realistic machine most often loses some of its input energy as heat in transforming it into the output energy stream.*

Law of Conservation of Energy 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". (Figure 1.4)

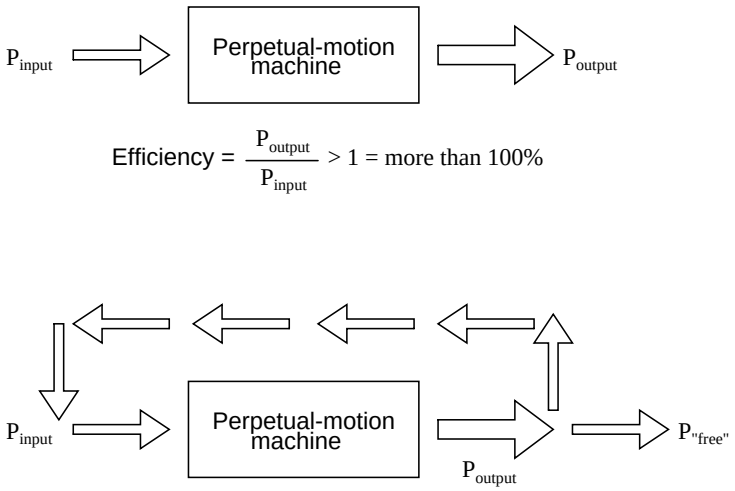


Figure 1.4: *Hypothetical "perpetual motion machine" powers itself?*

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 Conservation of Energy 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 Conservation of Energy 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 as shown in Figure 1.5.

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

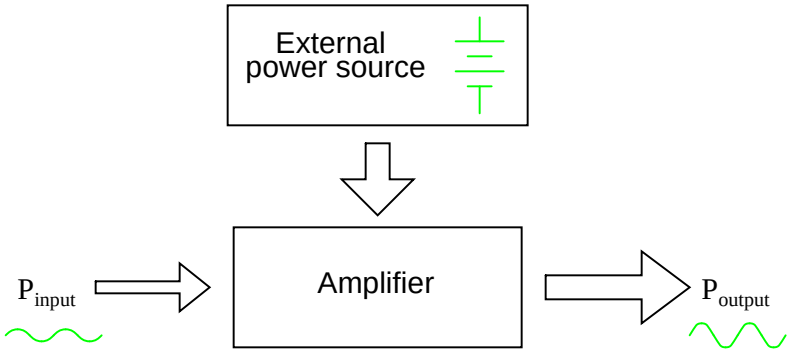


Figure 1.5: *While an amplifier can scale a small input signal to large output, its energy source is an external power supply.*

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:

$$I_{output} = (A_I)(I_{input})$$

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

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

In the last two examples I specifically identified the gains and signal magnitudes in terms of "AC." This was intentional, and illustrates an important concept: electronic amplifiers often respond differently to AC and DC input signals, and may amplify them to different extents. Another way of saying this is that amplifiers often amplify *changes* or *variations* in input signal magnitude (AC) at a different ratio than *steady* input signal magnitudes (DC). The specific reasons for this are too complex to explain at this time, but the fact of the matter is worth mentioning. If gain calculations are to be carried out, it must first be understood what type of signals and gains are being dealt with, AC or DC.

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

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

Δ = *"change in . . ."*

If multiple amplifiers are staged, their respective gains form an overall gain equal to the product (multiplication) of the individual gains. (Figure 1.6) If a 1 V signal were applied to the input of the gain of 3 amplifier in Figure 1.6 a 3 V signal out of the first amplifier would be further amplified by a gain of 5 at the second stage yielding 15 V at the final output.

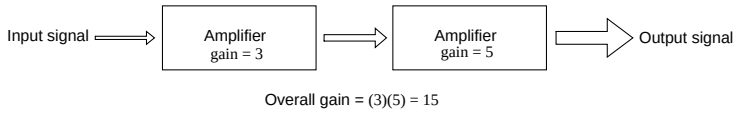


Figure 1.6: *The gain of a chain of cascaded amplifiers is the product of the individual gains.*

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:

Table: Gain / loss in bels

Loss/gain as a ratio	Loss/gain in bels	Loss/gain as a ratio	Loss/gain in bels
$\frac{P_{output}}{P_{input}}$	$\log \frac{P_{output}}{P_{input}}$	$\frac{P_{output}}{P_{input}}$	$\log \frac{P_{output}}{P_{input}}$
1000	3 B	0.1	-1 B
100	2 B	0.01	-2 B
10	1 B	0.001	-3 B
1 (no loss or gain)	0 B	0.0001	-4 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:

Table: Gain / loss in decibels

Loss/gain as a ratio	Loss/gain in decibels	Loss/gain as a ratio	Loss/gain in decibels
$\frac{P_{output}}{P_{input}}$	$10 \log \frac{P_{output}}{P_{input}}$	$\frac{P_{output}}{P_{input}}$	$10 \log \frac{P_{output}}{P_{input}}$
1000	30 dB	0.1	-10 dB
100	20 dB	0.01	-20 dB
10	10 dB	0.001	-30 dB
1 (no loss or gain)	0 dB	0.0001	-40 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 (Figure 1.6) 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:

```
Overall gain = (3)(5) = 15
```

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. (Figure 1.7)

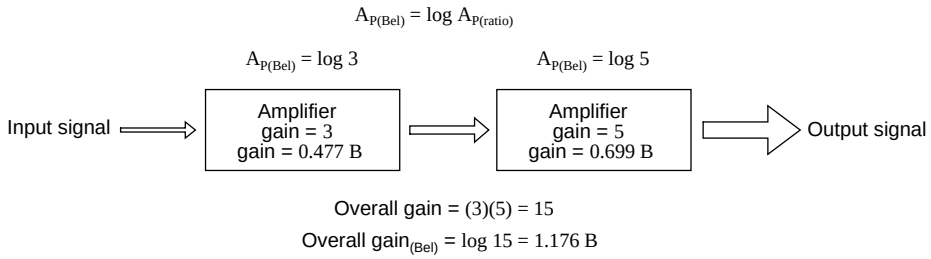


Figure 1.7: *Power gain in bels is additive: 0.477 B + 0.699 B = 1.176 B.*

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. (Figure 1.8)

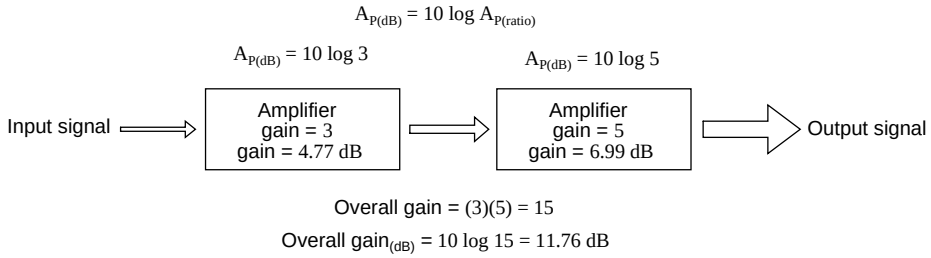


Figure 1.8: *Gain of amplifier stages in decibels is additive: 4.77 dB + 6.99 dB = 11.76 dB.*

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:

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

Example: Power into an amplifier is 1 Watt, the power out is 10 Watts. Find the power gain in dB.

$$A_{P(dB)} = 10 \log_{10}(P_O / P_I) = 10 \log_{10} (10 /1) = 10 \log_{10} (10) = 10 (1) = 10 \text{ dB}$$

Example: Find the power gain ratio $A_{P(ratio)} = (P_O / P_I)$ for a 20 dB Power gain.

$$A_{P(dB)} = 20 = 10 \log_{10} A_{P(ratio)}$$

$$20/10 = \log_{10} A_{P(ratio)}$$

$$10^{20/10} = 10^{log_{10}(A_{P(ratio)})}$$

$$100 = A_{P(ratio)} = (P_O / P_I)$$

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:

$A_{V(Bel)} = \log A_{V(ratio)}^{2}$	$A_{I(Bel)} = \log A_{I(ratio)}^{2}$
. . . is the same as . . .	*. . . is the same as . . .*
$A_{V(Bel)} = 2 \log A_{V(ratio)}$	$A_{I(Bel)} = 2 \log A_{I(ratio)}$

For decibels:

$A_{V(dB)} = 10 \log A_{V(ratio)}^{2}$	$A_{I(dB)} = 10 \log A_{I(ratio)}^{2}$
. . . is the same as . . .	*. . . is the same as . . .*
$A_{V(dB)} = 20 \log A_{V(ratio)}$	$A_{I(dB)} = 20 \log A_{I(ratio)}$

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 "N_p; though, lower-case "n" may be encountered.

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

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

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

Example: The voltage into a 600 Ω audio line amplifier is 10 mV, the voltage across a 600 Ω load is 1 V. Find the power gain in dB.

$$A_{(dB)} = 20 \log_{10}(V_O / V_I) = 20 \log_{10} (1 / 0.01) = 20 \log_{10} (100) = 20 (2) = 40 \text{ dB}$$

Example: Find the voltage gain ratio $A_{V(ratio)} = (V_O / V_I)$ for a 20 dB gain amplifier having a 50 Ω input and out impedance.

$$A_{V(dB)} = 20 \log_{10} A_{V(ratio)}$$

$$20 = 20 \log_{10} A_{V(ratio)}$$

$$20/20 = \log_{10} A_{P(ratio)}$$

$$10^{20/20} = 10^{log_{10}(A_{V(ratio)})}$$

$$10 = A_{V(ratio)} = (V_O / V_I)$$

- **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 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. (Figure 1.9) 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.

Radio frequency power measurements for low level signals encountered in radio receivers use dBm measurements referenced to a 50 Ω load. Signal generators for the evaluation of radio receivers may output an adjustable dBm rated signal. The signal level is selected by a device called an attenuator, described in the next section.

Table: Absolute power levels in dBm (decibel milliwatt)

Power in watts	Power in milliwatts	Power in dBm	Power in milliwatts	Power in dBm
1	1000	30 dB	1	0 dB
0.1	100	20 dB	0.1	-10 dB
0.01	10	10 dB	0.01	-20 dB
0.004	4	6 dB	0.001	-30 dB
0.002	2	3 dB	0.0001	-40 dB

Figure 1.9: *Absolute power levels in dBm (decibels referenced to 1 milliwatt).*

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 1 mW 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 Attenuators

Attenuators are passive devices. It is convenient to discuss them along with decibels. Attenuators weaken or *attenuate* the high level output of a signal generator, for example, to provide a lower level signal for something like the antenna input of a sensitive radio receiver. (Figure 1.10) The attenuator could be built into the signal generator, or be a stand-alone device. It could provide a fixed or adjustable amount of attenuation. An attenuator section can also provide isolation between a source and a troublesome load.

In the case of a stand-alone attenuator, it must be placed in series between the signal source and the load by breaking open the signal path as shown in Figure 1.10. In addition, it must match both the source impedance $\mathbf{Z}_I$ and the load impedance $\mathbf{Z}_O$, while providing a specified amount of attenuation. In this section we will only consider the special, and most

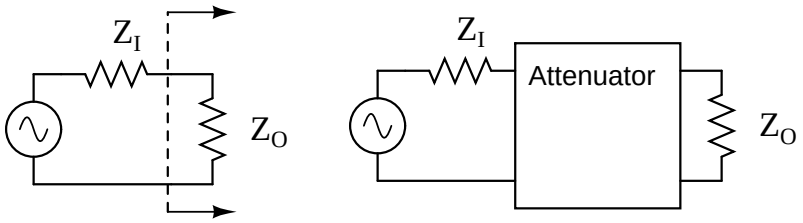


Figure 1.10: *Constant impedance attenuator is matched to source impedance Z_I and load impedance Z_O. For radio frequency equipment Z is 50 Ω.*

common, case where the source and load impedances are equal. Not considered in this section, unequal source and load impedances may be matched by an attenuator section. However, the formulation is more complex.

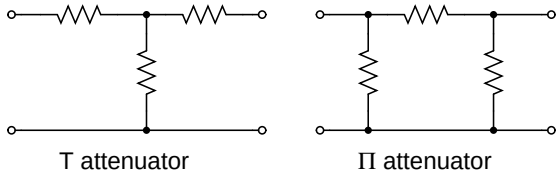


Figure 1.11: *T section and* Π *section attenuators are common forms.*

Common configurations are the **T** and Π networks shown in Figure 1.11 Multiple attenuator sections may be cascaded when even weaker signals are needed as in Figure 1.19.

1.7.1 Decibels

Voltage ratios, as used in the design of attenuators are often expressed in terms of decibels. The voltage ratio (K below) must be derived from the attenuation in decibels. Power ratios expressed as decibels are additive. For example, a 10 dB attenuator followed by a 6 dB attenuator provides 16dB of attenuation overall.

10 dB + 6 db = 16 dB

Changing sound levels are perceptible roughly proportional to the logarithm of the power ratio (P_I / P_O).

sound level = $\log_{10}(P_I / P_O)$

A change of 1 dB in sound level is barely perceptible to a listener, while 2 db is readily perceptible. An attenuation of 3 dB corresponds to cutting power in half, while a gain of 3 db corresponds to a doubling of the power level. A gain of -3 dB is the same as an attenuation of +3 dB, corresponding to half the original power level.

The power change in decibels in terms of power ratio is:

$$dB = 10 \log_{10}(P_I / P_O)$$

Assuming that the load R_I at P_I is the same as the load resistor R_O at P_O ($R_I = R_O$), the decibels may be derived from the voltage ratio (V_I / V_O) or current ratio (I_I / I_O):

$$P_O = V_O I_O = V_O^2 / R = I_O^2 R$$

$$P_I = V_I I_I = V_I^2 / R = I_I^2 R$$

$$dB = 10 \log_{10}(P_I / P_O) = 10 \log_{10}(V_I^2 / V_O^2) = 20 \log_{10}(V_I/V_O)$$

$$dB = 10 \log_{10}(P_I / P_O) = 10 \log_{10}(I_I^2 / I_O^2) = 20 \log_{10}(I_I/I_O)$$

The two most often used forms of the decibel equation are:

$$dB = 10 \log_{10}(P_I / P_O) \quad \text{or} \quad dB = 20 \log_{10}(V_I / V_O)$$

We will use the latter form, since we need the voltage ratio. Once again, the voltage ratio form of equation is only applicable where the two corresponding resistors are equal. That is, the source and load resistance need to be equal.

Example: Power into an attenuator is 10 Watts, the power out is 1 Watt. Find the attenuation in dB.

$$dB = 10 \log_{10}(P_I / P_O) = 10 \log_{10}(10 /1) = 10 \log_{10}(10) = 10 (1) = 10 \text{ dB}$$

Example: Find the voltage attenuation ratio ($K = (V_I / V_O)$) for a 10 dB attenuator.

$$dB = 10 = 20 \log_{10}(V_I / V_O)$$

$$10/20 = \log_{10}(V_I / V_O)$$

$$10^{10/20} = 10^{log_{10}(V_I/V_O)}$$

$$3.16 = (V_I / V_O) = A_{P(ratio)}$$

Example: Power into an attenuator is 100 milliwatts, the power out is 1 milliwatt. Find the attenuation in dB.

$$dB = 10 \log_{10}(P_I / P_O) = 10 \log_{10}(100 /1) = 10 \log_{10}(100) = 10 (2) = 20 \text{ dB}$$

Example: Find the voltage attenuation ratio ($K = (V_I / V_O)$) for a 20 dB attenuator.

$$dB = 20 = 20 \log_{10}(V_I / V_O)$$

$$10^{20/20} = 10^{log_{10}(V_I/V_O)}$$

$$10 = (V_I / V_O) = K$$

1.7.2 T-section attenuator

The T and Π attenuators must be connected to a **Z** source and **Z** load impedance. The **Z**-(arrows) pointing away from the attenuator in the figure below indicate this. The **Z**-(arrows) pointing toward the attenuator indicates that the impedance seen looking into the attenuator with a load Z on the opposite end is Z, Z=50 Ω for our case. This impedance is a constant (50 Ω) with respect to attenuation– impedance does not change when attenuation is changed.

The table in Figure 1.12 lists resistor values for the **T** and Π attenuators to match a 50 Ω source/ load, as is the usual requirement in radio frequency work.

Telephone utility and other audio work often requires matching to 600 Ω. Multiply all **R** values by the ratio (600/50) to correct for 600 Ω matching. Multiplying by 75/50 would convert table values to match a 75 Ω source and load.

dB = attenuation in decibels

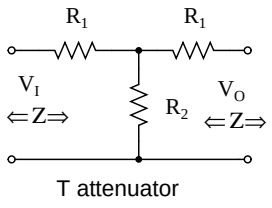
Z = source/load impedance (resistive)

K > 1

$$K = \frac{V_I}{V_O} = 10^{dB/20}$$

$$R_1 = Z\left(\frac{K-1}{K+1}\right)$$

$$R_2 = Z\left(\frac{2K}{K^2-1}\right)$$



Resistors for T-section			
Z = 50			
Attenuation			
dB	K=Vi/Vo	R1	R2
1.0	1.12	2.88	433.34
2.0	1.26	5.73	215.24
3.0	1.41	8.55	141.93
4.0	1.58	11.31	104.83
6.0	2.00	16.61	66.93
10.0	3.16	25.97	35.14
20.0	10.00	40.91	10.10

Figure 1.12: *Formulas for T-section attenuator resistors, given K, the voltage attenuation ratio, and $Z_I = Z_O = 50\ \Omega$.*

The amount of attenuation is customarily specified in dB (decibels). Though, we need the voltage (or current) ratio **K** to find the resistor values from equations. See the **dB/20** term in the power of **10** term for computing the voltage ratio **K** from dB, above.

The **T** (and below Π) configurations are most commonly used as they provide bidirectional matching. That is, the attenuator input and output may be swapped end for end and still match the source and load impedances while supplying the same attenuation.

Disconnecting the source and looking in to the right at $\mathbf{V}_I$, we need to see a series parallel combination of $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_1$, and **Z** looking like an equivalent resistance of $\mathbf{Z}_{IN}$, the same as the source/load impedance Z: (a load of Z is connected to the output.)

$$Z_{IN} = R_1 + (R_2 \,||(R_1 + Z))$$

For example, substitute the 10 dB values from the 50 Ω attenuator table for $\mathbf{R}_1$ and $\mathbf{R}_2$ as shown in Figure 1.13.

$$Z_{IN} = 25.97 + (35.14 \,||(25.97 + 50))$$

$$Z_{IN} = 25.97 + (35.14 \,||\, 75.97\ )$$

$$Z_{IN} = 25.97 + 24.03 = 50$$

This shows us that we see 50 Ω looking right into the example attenuator (Figure 1.13) with a 50 Ω load.

Replacing the source generator, disconnecting load **Z** at $\mathbf{V}_O$, and looking in to the left, should give us the same equation as above for the impedance at $\mathbf{V}_O$, due to symmetry. Moreover, the three resistors must be values which supply the required attenuation from input to output. This is accomplished by the equations for $\mathbf{R}_1$ and $\mathbf{R}_2$ above as applied to the **T**-attenuator below.

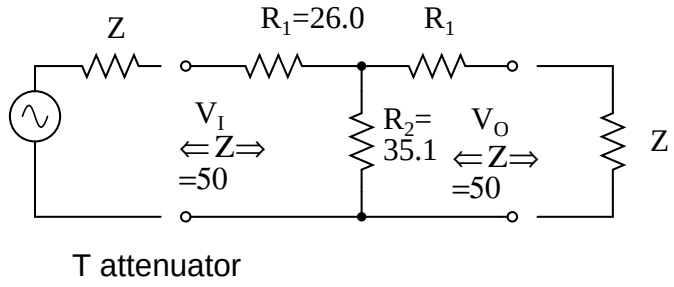


Figure 1.13: *10 dB T-section attenuator for insertion between a 50 Ω source and load.*

1.7.3 PI-section attenuator

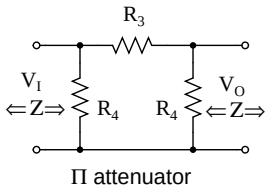
The table in Figure 1.14 lists resistor values for the Π attenuator matching a 50 Ω source/ load at some common attenuation levels. The resistors corresponding to other attenuation levels may be calculated from the equations.

dB = attenuation in decibels
Z = source/load impedance (resistive)
K > 1

$$K = \frac{V_I}{V_O} = 10^{dB/20}$$

$$R_3 = Z\left(\frac{K^2-1}{2K}\right)$$

$$R_4 = Z\left(\frac{K+1}{K-1}\right)$$



Resistors for Π-section			
Z=50.00			
Attenuation			
dB	K=Vi/Vo	R3	R4
1.0	1.12	5.77	869.55
2.0	1.26	11.61	436.21
3.0	1.41	17.61	292.40
4.0	1.58	23.85	220.97
6.0	2.00	37.35	150.48
10.0	3.16	71.15	96.25
20.0	10.00	247.50	61.11

Figure 1.14: *Formulas for* Π*-section attenuator resistors, given K, the voltage attenuation ratio, and* $Z_I = Z_O = 50\ \Omega$.

The above apply to the π-attenuator below.

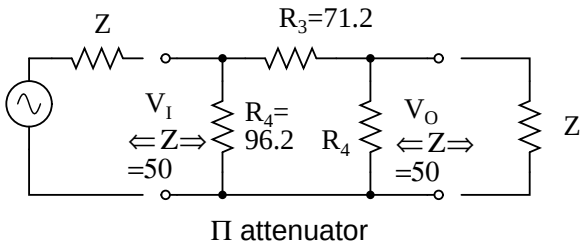


Figure 1.15: *10 dB Π-section attenuator example for matching a 50 Ω source and load.*

What resistor values would be required for both the Π attenuators for 10 dB of attenuation matching a 50 Ω source and load?

The **10 dB** corresponds to a voltage attenuation ratio of **K=3.16** in the next to last line of the above table. Transfer the resistor values in that line to the resistors on the schematic diagram in Figure 1.15.

1.7.4 L-section attenuator

The table in Figure 1.16 lists resistor values for the **L** attenuators to match a 50 Ω source/load. The table in Figure 1.17 lists resistor values for an alternate form. Note that the resistor values are not the same.

dB = attenuation in decibels

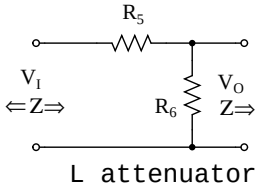
Z = source/load impedance (resistive)

K > 1

$$K = \frac{V_I}{V_O} = 10^{dB/20}$$

$$R_5 = Z\left(\frac{K-1}{K}\right)$$

$$R_6 = \frac{Z}{(K-1)}$$



Resistors for L-section			
Z=,50.00			
Attenuation L			
dB	K=Vi/Vo	R5	R6
1.0	1.12	5.44	409.77
2.0	1.26	10.28	193.11
3.0	1.41	14.60	121.20
4.0	1.58	18.45	85.49
6.0	2.00	24.94	50.24
10.0	3.16	34.19	23.12
20.0	10.00	45.00	5.56

Figure 1.16: *L-section attenuator table for 50 Ω source and load impedance.*

The above apply to the **L** attenuator below.

1.7.5 Bridged T attenuator

The table in Figure 1.18 lists resistor values for the bridged **T** attenuators to match a 50 Ω source and load. The bridged-T attenuator is not often used. Why not?

dB = attenuation in decibels

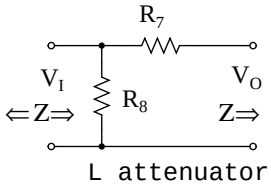
Z = source/load impedance (resistive)

K > 1

$$K = \frac{V_I}{V_O} = 10^{dB/20}$$

$$R_7 = Z(K-1)$$

$$R_8 = Z\left(\frac{K}{K-1}\right)$$



Resistors for L-section			
Z=50.00			
Attenuation			
dB	K=Vi/Vo	R7	R8
1.0	1.12	6.10	459.77
2.0	1.26	12.95	243.11
3.0	1.41	20.63	171.20
4.0	1.58	29.24	135.49
6.0	2.00	49.76	100.24
10.0	3.16	108.11	73.12
20.0	10.00	450.00	55.56

Figure 1.17: *Alternate form L-section attenuator table for 50 Ω source and load impedance.*

dB = attenuation in decibels

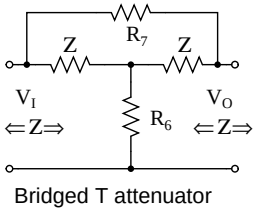
Z = source/load impedance (resistive)

K > 1

$$K = \frac{V_I}{V_O} = 10^{dB/20}$$

$$R_6 = \frac{Z}{(K-1)}$$

$$R_7 = Z(K-1)$$



Resistors for bridged T			
Z=50.00			
Attenuation			
dB	K=Vi/Vo	R7	R6
1.0	1.12	6.10	409.77
2.0	1.26	12.95	193.11
3.0	1.41	20.63	121.20
4.0	1.58	29.24	85.49
6.0	2.00	49.76	50.24
10.0	3.16	108.11	23.12
20.0	10.00	450.00	5.56

Figure 1.18: *Formulas and abbreviated table for bridged-T attenuator section, Z = 50 Ω.*

1.7.6 Cascaded sections

Attenuator sections can be cascaded as in Figure 1.19 for more attenuation than may be available from a single section. For example two 10 db attenuators may be cascaded to provide 20 dB of attenuation, the dB values being additive. The voltage attenuation ratio **K** or $\mathbf{V}_I/\mathbf{V}_O$ for a 10 dB attenuator section is 3.16. The voltage attenuation ratio for the two cascaded sections is the product of the two **K**s or 3.16x3.16=10 for the two cascaded sections.

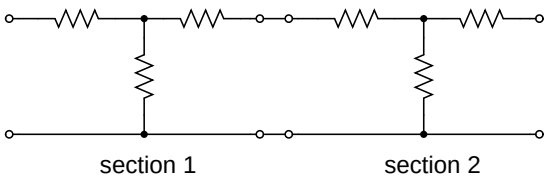


Figure 1.19: *Cascaded attenuator sections: dB attenuation is additive.*

Variable attenuation can be provided in discrete steps by a switched attenuator. The example Figure 1.20, shown in the 0 dB position, is capable of 0 through 7 dB of attenuation by additive switching of none, one or more sections.

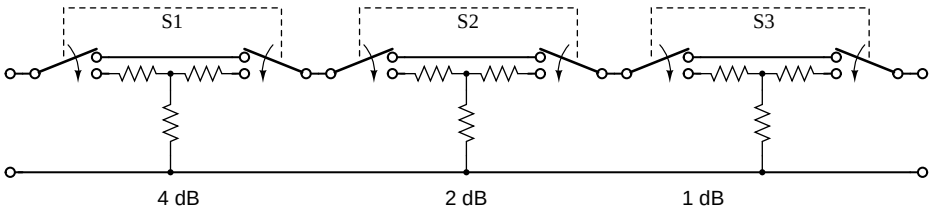


Figure 1.20: *Switched attenuator: attenuation is variable in discrete steps.*

The typical multi section attenuator has more sections than the above figure shows. The addition of a 3 or 8 dB section above enables the unit to cover to 10 dB and beyond. Lower signal levels are achieved by the addition of 10 dB and 20 dB sections, or a binary multiple 16 dB section.

1.7.7 RF attenuators

For radio frequency (RF) work (<1000 Mhz), the individual sections must be mounted in shielded compartments to thwart capacitive coupling if lower signal levels are to be achieved at the highest frequencies. The individual sections of the switched attenuators in the previous section are mounted in shielded sections. Additional measures may be taken to extend the frequency range to beyond 1000 Mhz. This involves construction from special shaped lead-less resistive elements.

A coaxial T-section attenuator consisting of resistive rods and a resistive disk is shown in Figure 1.21. This construction is usable to a few gigahertz. The coaxial Π version would have one resistive rod between two resistive disks in the coaxial line as in Figure 1.22.

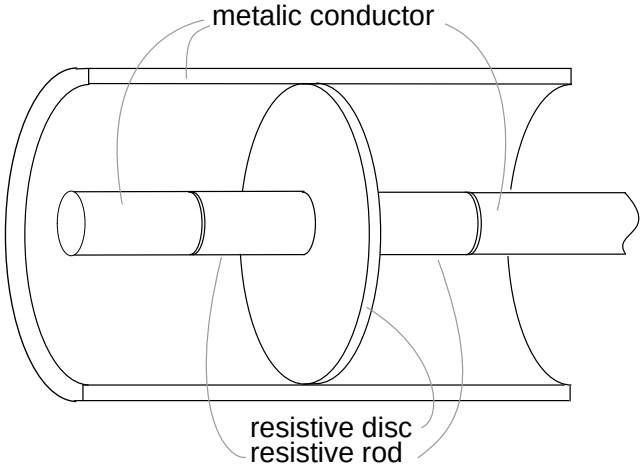


Figure 1.21: *Coaxial T-attenuator for radio frequency work.*

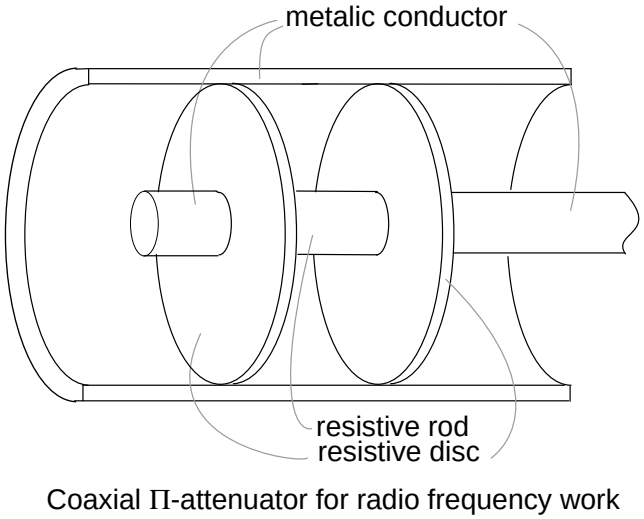


Figure 1.22: *Coaxial* Π*-attenuator for radio frequency work.*

RF connectors, not shown, are attached to the ends of the above T and Π attenuators. The connectors allow individual attenuators to be cascaded, in addition to connecting between a source and load. For example, a 10 dB attenuator may be placed between a troublesome signal source and an expensive spectrum analyzer input. Even though we may not need the attenuation, the expensive test equipment is protected from the source by attenuating any overvoltage.

Summary: Attenuators

- An *attenuator* reduces an input signal to a lower level.
- The amount of attenuation is specified in *decibels* (dB). Decibel values are additive for cascaded attenuator sections.
- dB from power ratio: $dB = 10 \log_{10}(P_I / P_O)$
- dB from voltage ratio: $dB = 20 \log_{10}(V_I / V_O)$
- *T* and Π section attenuators are the most common circuit configurations.

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

2.1 Introduction 27
2.2 Quantum physics 28
2.3 Valence and Crystal structure 41
2.4 Band theory of solids 47
2.5 Electrons and "holes" 50
2.6 The P-N junction 55
2.7 Junction diodes 58
2.8 Bipolar junction transistors 60
2.9 Junction field-effect transistors 65
2.10 Insulated-gate field-effect transistors (MOSFET) 70
2.11 Thyristors 73
2.12 Semiconductor manufacturing techniques 75
2.13 Superconducting devices 80
2.14 Quantum devices 83
2.15 Semiconductor devices in SPICE 91
Bibliography 93

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. This chapter is but 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 try 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 Figure 2.1.

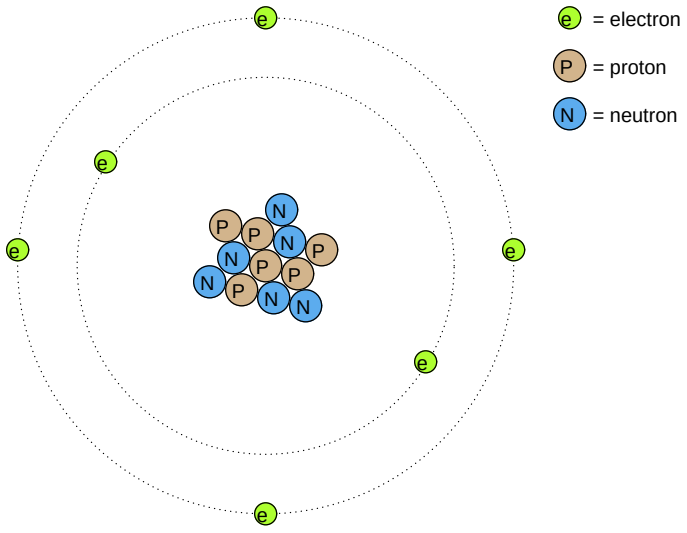


Figure 2.1: *Rutherford atom: negative electrons orbit a small positive nucleus.*

Tiny particles of matter called *protons* and *neutrons* make up the center of the atom; *electrons* orbit like 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 are attracted to the positive protons just as planets are gravitationally attracted by the Sun, yet the orbits are stable because of 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.

Rutherford's scattering experiment involved bombarding a thin gold foil with positively charged alpha particles as in Figure 2.2. Young graduate students H. Geiger and E. Marsden experienced unexpected results. A few Alpha particles were deflected at large angles. A few Alpha particles were back-scattering, recoiling at nearly 180^o. Most of the particles passed through the gold foil undeflected, indicating that the foil was mostly empty space. The fact that a few alpha particles experienced large deflections indicated the presence of a minuscule positively charged nucleus.

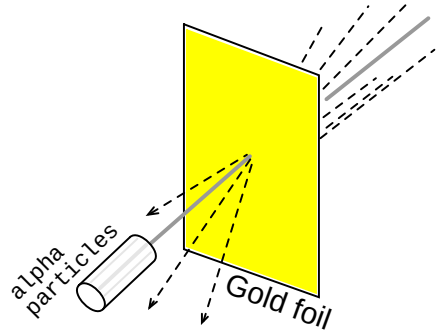


Figure 2.2: *Rutherford scattering: a beam of alpha particles is scattered by a thin gold foil.*

Although 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. Nevertheless, despite the revolution of quantum physics and its contribution to our understanding of atomic structure, Rutherford's solar-system picture of the atom embedded itself in the popular consciousness 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 because of two equal but opposite forces. The centrifugal outward force exerted on the electrons because of the orbit counteracts the attractive inward force (centripetal) trying to pull the electrons toward the nucleus because of 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 because of its popular acceptance: in reality, there is no such thing as a force pushing *any* orbiting body *away* from its center of orbit. It 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. Nevertheless, 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 only occurs 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 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. This phenomenon 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 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 had a certain, specific amount of energy, and that their orbits were *quantized* such that each may occupy certain places around the nucleus, as marbles fixed in circular tracks around the nucleus rather than the free-ranging satellites each were formerly imagined to be. (Figure 2.3) In deference to the laws of electromagnetics and accelerating charges, Bohr alluded to these "orbits" as *stationary states* to escape the implication that they were in motion.

Although 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 not complete. His mathematical analysis produced better predictions of experimental events than analyses belonging to previous models, but there were still some unanswered questions about *why* electrons should behave in such strange ways. The assertion that electrons existed in stationary,

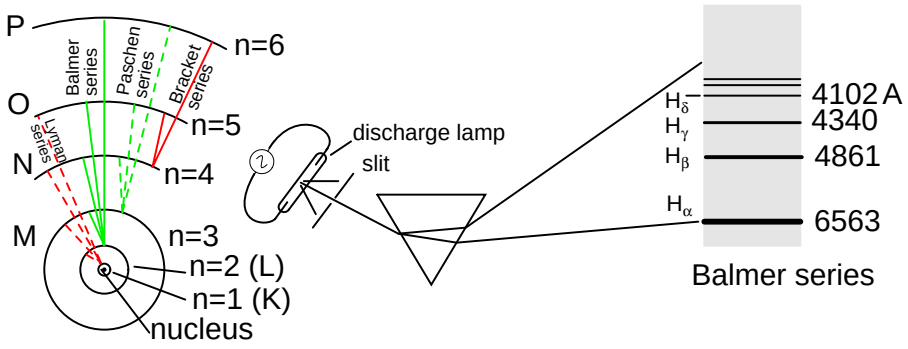


Figure 2.3: *Bohr hydrogen atom (with orbits drawn to scale) only allows electrons to inhabit discrete orbitals. Electrons falling from n=3,4,5, or 6 to n=2 accounts for Balmer series of spectral lines.*

quantized states around the nucleus 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, as 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, another breakthrough in understanding was reached.

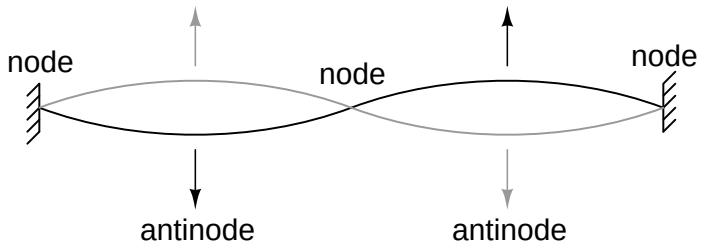


Figure 2.4: *String vibrating at resonant frequency between two fixed points forms* ***standing wave****.*

The atom according to de Broglie consisted of electrons existing as *standing waves*, a phenomenon well known to physicists in a variety of forms. As the plucked string of a musical instrument (Figure 2.4) 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 as in Figure 2.5.

Electrons only could 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 should destructively interfere with itself and thus cease to exist.

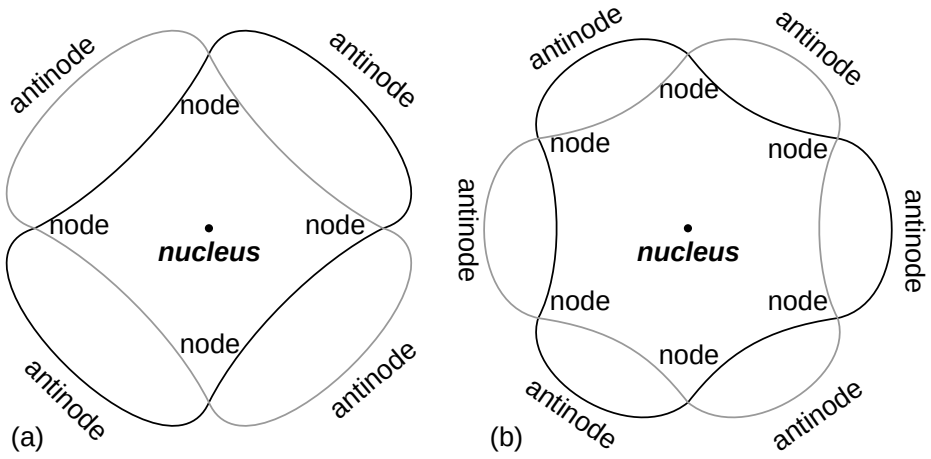


Figure 2.5: *"Orbiting" electron as standing wave around the nucleus, (a) two cycles per orbit, (b) three cycles per orbit.*

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. The popular explanation of this "uncertainty principle" was that it was a measurement error (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 vice versa). The startling implication of quantum mechanics is that particles do not actually have 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.

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 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. 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 have both a precise, instantaneous amplitude, and a precise frequency at the same time.

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" have other characteristics that *are* discrete. Any electron in an atom can be described by 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 principal quantum number must be a positive integer (a whole number, greater than or equal to 1). In other words, principal quantum number for an electron cannot be 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 as illustrated in Figure 2.3.

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 (one cannot sit *between* rows), electrons must "choose" a particular shell to "sit" in. As in amphitheater rows, the outermost shells hold more electrons than the inner shells. Also, electrons tend to seek the lowest available shell, as people in an amphitheater seek 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 may hold is described by the equation $2n^2$, where "n" is the principal 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. (Figure 2.6)

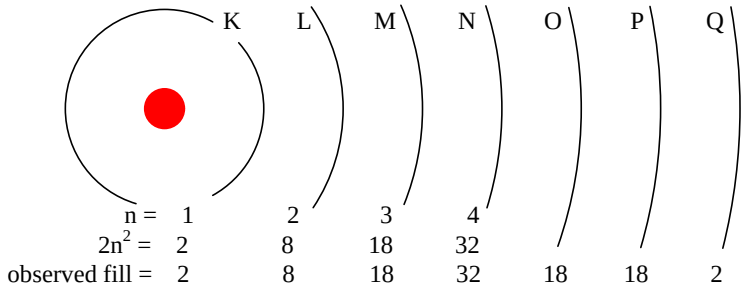


Figure 2.6: *Principal quantum number n and maximum number of electrons per shell both predicted by 2(n^2), and observed. Orbitals not to scale.*

Electron shells in an atom were formerly designated by letter rather than by number. The first shell (n=1) was 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: A shell, is composed of *subshells*. One might be inclined to think of subshells as simple subdivisions of shells, as lanes dividing a road. The subshells are much stranger. 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, (Figure 2.7(s)) which makes sense when visualized as 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. (Figure 2.7(p)) 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. (Figure 2.7(d))

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 notation, 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

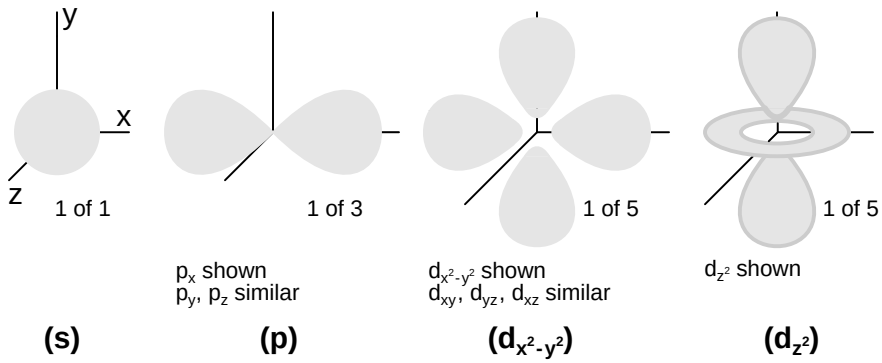


Figure 2.7: *Orbitals: (s) Three fold symmetry. (p) Shown: p_x, one of three possible orientations (p_x, p_y, p_z), about their respective axes. (d) Shown: $d_x{}^2$-$_y{}^2$ similar to d_{xy}, d_{yz}, d_{xz}. Shown: $d_z{}^2$. Possible d-orbital orientations: five.*

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. (Figure 2.8)

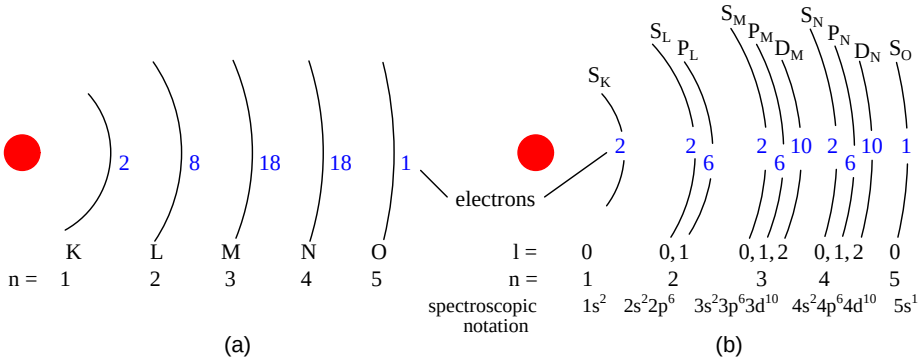


Figure 2.8: *(a) Bohr representation of Silver atom, (b) Subshell representation of Ag with division of shells into subshells (angular quantum number l). This diagram implies nothing about the actual position of electrons, but represents energy levels.*

Magnetic Quantum Number: The magnetic quantum number for an electron classifies which orientation its subshell shape is pointed. The "lobes" for subshells point in multiple directions. These different orientations are called *orbitals*. For the first subshell (s; l=0), which resembles a sphere pointing in no "direction", so there is only one orbital. For the second (p; l=1) subshell in each shell, which resembles dumbbells point in three possible directions. Think of three dumbbells intersecting at the origin, each oriented along a different axis in a three-axis coordinate space.

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, $(2 \cdot l + 1)$. 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.

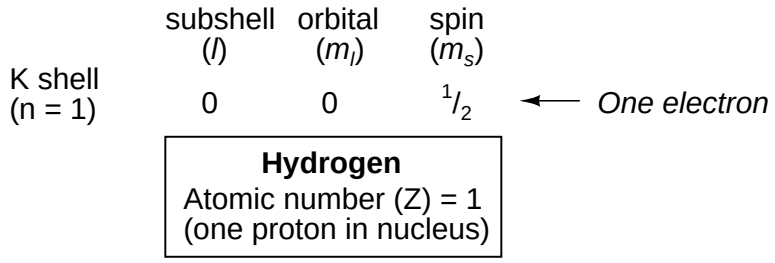
Like principal quantum numbers, the magnetic quantum number arose directly from experimental evidence: The Zeeman effect, the division of spectral lines by exposing an ionized gas to a magnetic field, hence the name "magnetic" quantum number.

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 result from each electron "spinning" on an axis as if a planet. Electrons with different "spins" would give off slightly different frequencies of light when excited. The name "spin" was assigned to this quantum number. 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 "clouds"; 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 may 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.

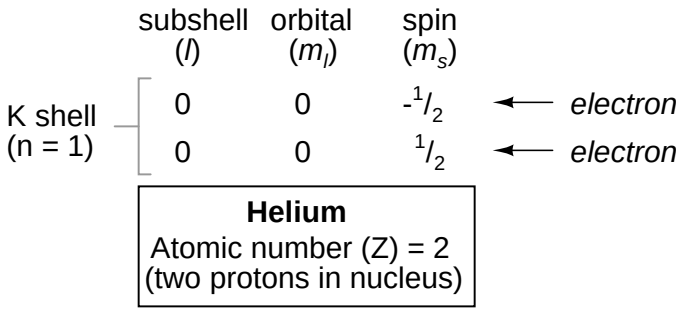
Shown here is the electron arrangement for a hydrogen atom:



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 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, is described as $1s^1$.

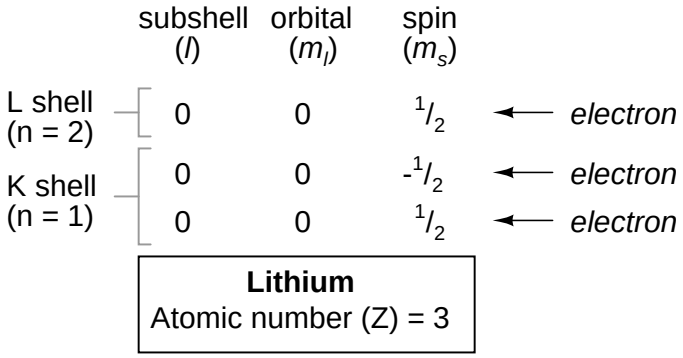
Proceeding to the next atom (in order of atomic number), we have the element helium:



Spectroscopic notation: $1s^2$

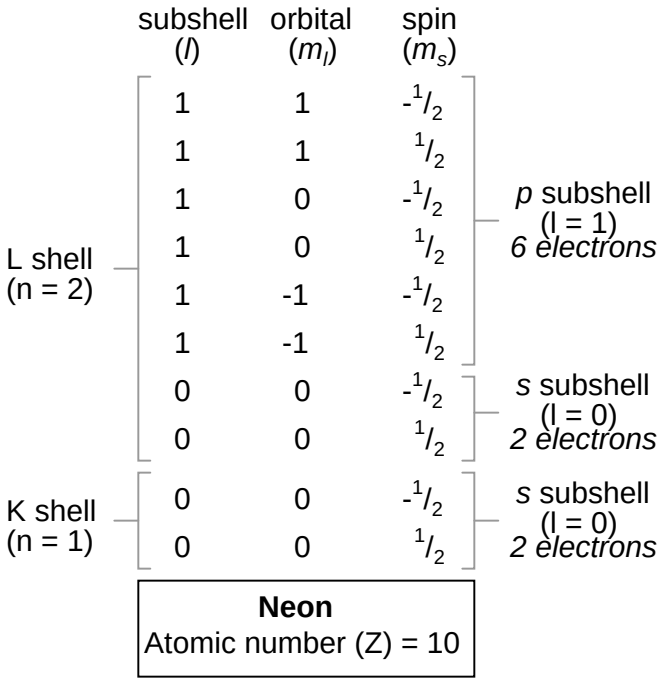
A helium atom has two protons in the nucleus, and this necessitates two electrons to balance the double-positive electric charge. Since two electrons – one with spin=1/2 and the other with spin=-1/2 – fit into one orbital, the electron configuration of helium requires no additional subshells or shells 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 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 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), giving 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 almost complete non-reactivity with other elements. These elements used to be classified as *inert*, when it was thought that these were completely unreactive, but are now known to form compounds with other elements under specific conditions.

Since elements with identical electron configurations in their outermost shell(s) exhibit similar chemical properties, Dmitri Mendeleev organized 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 in Figure 2.9.

Figure 2.9: *Periodic table of chemical elements.*

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 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 should 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, furthering future discoveries, and 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 may 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" due to the 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, its falsehood is disproven.

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 at micro 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 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*, known as the *shell*; the *Angular Momentum Quantum Number*, 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 no "in-between" conditions exist 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 that 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 ending 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. Subshells may assume as many different orientations as 2-times the subshell number (l) plus 1, (2l+1) (E.g. for l=1, ml= -1, 0, 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 may 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, no more than two electrons may occupy 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 Valence and Crystal structure

Valence: The electrons in the outer most shell, or valence shell, are known as *valence* electrons. These valence electrons are responsible for the chemical properties of the chemical elements. It is these electrons which participate in chemical reactions with other elements. An over simplified chemistry rule applicable to simple reactions is that atoms try to form a complete outer shell of 8 electrons (two for the L shell). Atoms may give away a few electrons to expose an underlying complete shell. Atoms may accept a few electrons to complete the shell. These two processes form ions from atoms. Atoms may even share electrons among atoms in an attempt to complete the outer shell. This process forms molecular bonds. That is, atoms associate to form a molecule.

For example group I elements: Li, Na, K, Cu, Ag, and Au have a single valence electron. (Figure 2.10) These elements all have similar chemical properties. These atoms readily give away one electron to react with other elements. The ability to easily give away an electron makes these elements excellent conductors.

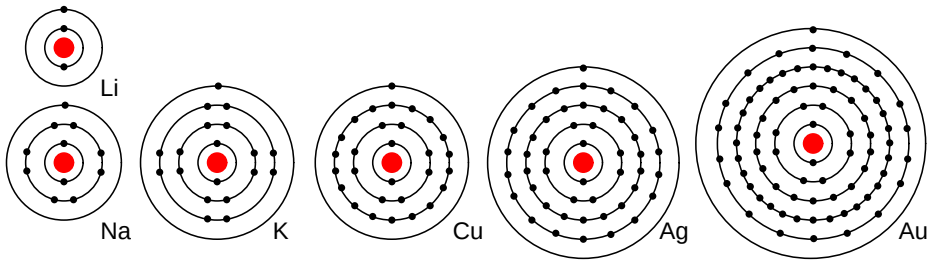


Figure 2.10: *Periodic table group IA elements: Li, Na, and K, and group IB elements: Cu, Ag, and Au have one electron in the outer, or valence, shell, which is readily donated. Inner shell electrons: For n= 1, 2, 3, 4; $2n^2$ = 2, 8, 18, 32.*

Group VIIA elements: Fl, Cl, Br, and I all have 7 electrons in the outer shell. These elements readily accept an electron to fill up the outer shell with a full 8 electrons. (Figure 2.11) If these elements do accept an electron, a negative ion is formed from the neutral atom. These elements which do not give up electrons are insulators.

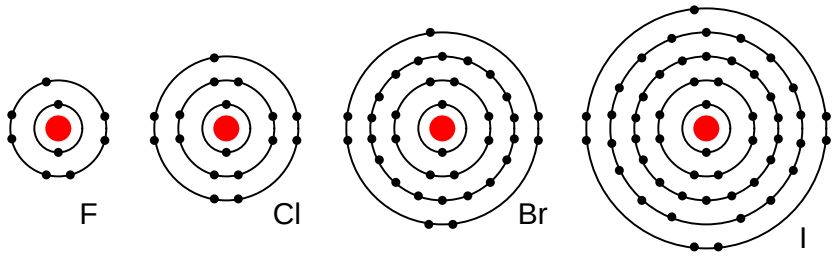


Figure 2.11: *Periodic table group VIIA elements: F, Cl, Br, and I with 7 valence electrons readily accept an electron in reactions with other elements.*

For example, a Cl atom accepts an electron from an Na atom to become a Cl^- ion as shown in Figure 2.12. An *ion* is a charged particle formed from an atom by either donating or accepting an electron. As the Na atom donates an electron, it becomes a Na^+ ion. This is how Na and Cl atoms combine to form NaCl, table salt, which is actually Na^+Cl^-, a pair of ions. The Na^+ and Cl^- carrying opposite charges, attract one other.

Sodium chloride crystallizes in the cubic structure shown in Figure 2.16. This model is not to scale to show the three dimensional structure. The Na^+Cl^- ions are actually packed similar to layers of stacked marbles. The easily drawn cubic crystal structure illustrates that a solid crystal may contain charged particles.

Group VIIIA elements: He, Ne, Ar, Kr, Xe all have 8 electrons in the valence shell. (Figure

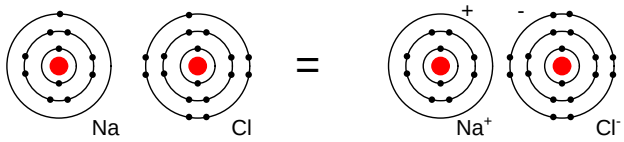


Figure 2.12: *Neutral Sodium atom donates an electron to neutral Chlorine atom forming Na^+ and Cl^- ions.*

below) That is, the valence shell is complete meaning these elements neither donate nor accept electrons. Nor do they readily participate in chemical reactions since group VIIIA elements do not easily combine with other elements. In recent years chemists have forced Xe and Kr to form a few compounds, however for the purposes of our discussion this is not applicable. These elements are good electrical insulators and are gases at room temperature.

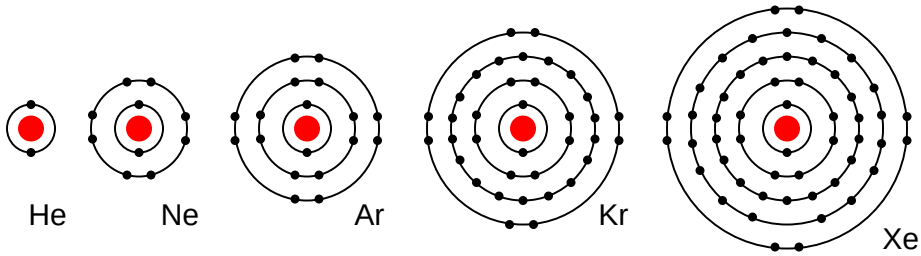


Figure 2.13: *Group VIIIA elements: He, Ne, Ar, Kr, Xe are largely unreactive since the valence shell is complete..*

Group IVA elements: C, Si, Ge, having 4 electrons in the valence shell as shown in Figure 2.14 form compounds by sharing electrons with other elements without forming ions. This shared electron bonding is known as *covalent bonding*. Note that the center atom (and the others by extension) has completed its valence shell by sharing electrons. Note that the figure is a 2-d representation of bonding, which is actually 3-d. It is this group, IVA, that we are interested in for its semiconducting properties.

Crystal structure: Most inorganic substances form their atoms (or ions) into an ordered array known as a *crystal*. The outer electron clouds of atoms interact in an orderly manner. Even metals are composed of crystals at the microscopic level. If a metal sample is given an optical polish, then acid etched, the microscopic *microcrystalline* structure shows as in Figure 2.15. It is also possible to purchase, at considerable expense, metallic single crystal specimens from specialized suppliers. Polishing and etching such a specimen discloses no microcrystalline structure. Practically all industrial metals are polycrystalline. Most modern semiconductors, on the other hand, are single crystal devices. We are primarily interested in monocrystalline structures.

Many metals are soft and easily deformed by the various metal working techniques. The microcrystals are deformed in metal working. Also, the valence electrons are free to move about the crystal lattice, and from crystal to crystal. The valence electrons do not belong to

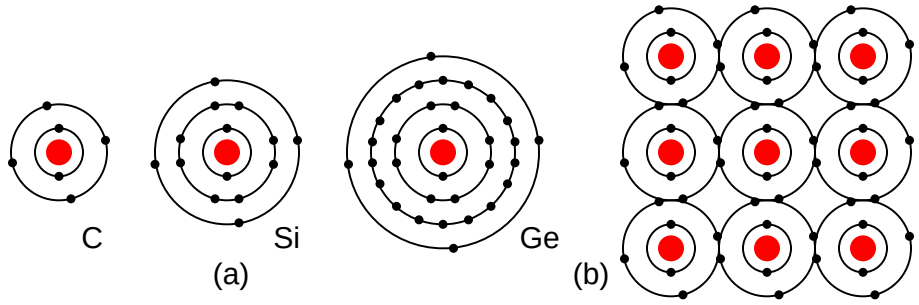


Figure 2.14: *(a) Group IVA elements: C, Si, Ge having 4 electrons in the valence shell, (b) complete the valence shell by sharing electrons with other elements.*

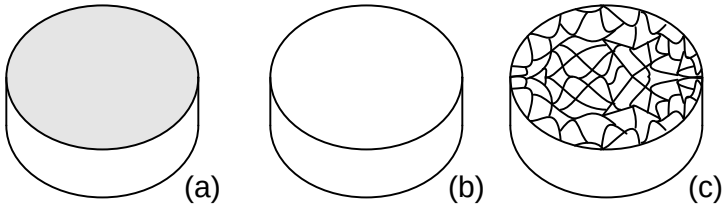


Figure 2.15: *(a) Metal sample, (b) polished, (c) acid etched to show microcrystalline structure.*

any particular atom, but to all atoms.

The rigid crystal structure in Figure 2.16 is composed of a regular repeating pattern of positive Na ions and negative Cl ions. The Na and Cl atoms form Na^+ and Cl^- ions by transferring an electron from Na to Cl, with no free electrons. Electrons are not free to move about the crystal lattice, a difference compared with a metal. Nor are the ions free. Ions are fixed in place within the crystal structure. Though, the ions are free to move about if the NaCl crystal is dissolved in water. However, the crystal no longer exists. The regular, repeating structure is gone. Evaporation of the water deposits the Na^+ and Cl^- ions in the form of new crystals as the oppositely charged ions attract each other. Ionic materials form crystal structures due to the strong electrostatic attraction of the oppositely charged ions.

Semiconductors in Group 14 (formerly part of Group IV) form a tetrahedral bonding pattern utilizing the s and p orbital electrons about the atom, sharing electron-pair bonds to four adjacent atoms. (Figure 2.18(a)). Group 14 elements have four outer electrons: two in a spherical s-orbital and two in p-orbitals. One of the p-orbitals is unoccupied. The three p-orbitals hybridize with the s-orbital to form four sp^3 molecular orbitals. These four electron clouds repel one another to equidistant tetrahedral spacing about the Si atom, attracted by the positive nucleus as shown in Figure 2.17.

Every semiconductor atom, Si, Ge, or C (diamond) is chemically bonded to four other atoms by *covalent bonds*, shared electron bonds. Two electrons may share an orbital if each have opposite spin quantum numbers. Thus, an unpaired electron may share an orbital with an electron from another atom. This corresponds to overlapping Figure 2.18(a) of the electron

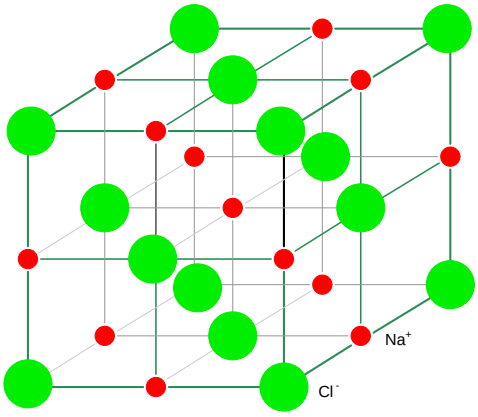


Figure 2.16: *NaCl crystal having a cubic structure.*

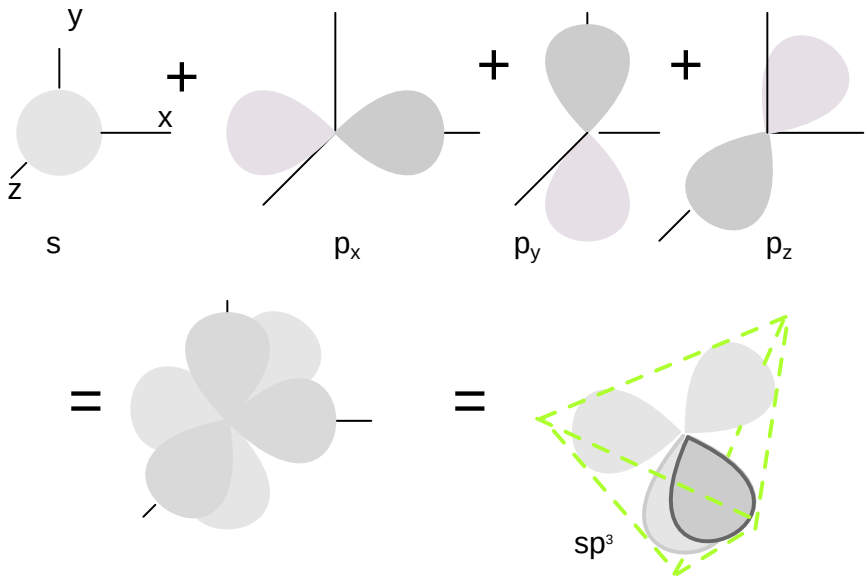


Figure 2.17: *One s-orbital and three p-orbital electrons hybridize, forming four sp^3 molecular orbitals.*

clouds, or bonding. Figure 2.18 (b) is one fourth of the volume of the diamond crystal structure unit cell shown in Figure 2.19 at the origin. The bonds are particularly strong in diamond, decreasing in strength going down group IV to silicon, and germanium. Silicon and germanium both form crystals with a diamond structure.

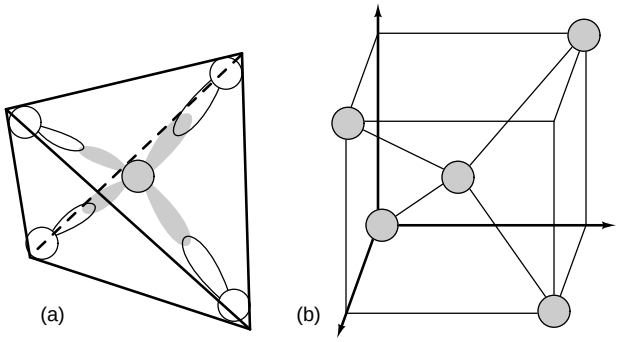


Figure 2.18: *(a) Tetrahedral bonding of Si atom. (b) leads to 1/4 of the cubic unit cell*

The diamond *unit cell* is the basic crystal building block. Figure 2.19 shows four atoms (dark) bonded to four others within the volume of the cell. This is equivalent to placing one of Figure 2.18(b) at the origin in Figure 2.19, then placing three more on adjacent faces to fill the full cube. Six atoms fall on the middle of each of the six cube faces, showing two bonds. The other two bonds to adjacent cubes were omitted for clarity. Out of eight cube corners, four atoms bond to an atom within the cube. Where are the other four atoms bonded? The other four bond to adjacent cubes of the crystal. Keep in mind that even though four corner atoms show no bonds in the cube, all atoms within the crystal are bonded in one giant molecule. A semiconductor crystal is built up from copies of this unit cell.

The crystal is effectively one molecule. An atom covalent bonds to four others, which in turn bond to four others, and so on. The crystal lattice is relatively stiff resisting deformation. Few electrons free themselves for conduction about the crystal. A property of semiconductors is that once an electron is freed, a positively charged empty space develops which also contributes to conduction.

- **REVIEW**
- Atoms try to form a complete outer, valence, shell of 8-electrons (2-electrons for the innermost shell). Atoms may donate a few electrons to expose an underlying shell of 8, accept a few electrons to complete a shell, or share electrons to complete a shell.
- Atoms often form ordered arrays of ions or atoms in a rigid structure known as a crystal.
- A neutral atom may form a positive ion by donating an electron.
- A neutral atom may form a negative ion by accepting an electron
- The group IVA semiconductors: C, Si, Ge crystallize into a diamond structure. Each atom in the crystal is part of a giant molecule, bonding to four other atoms.

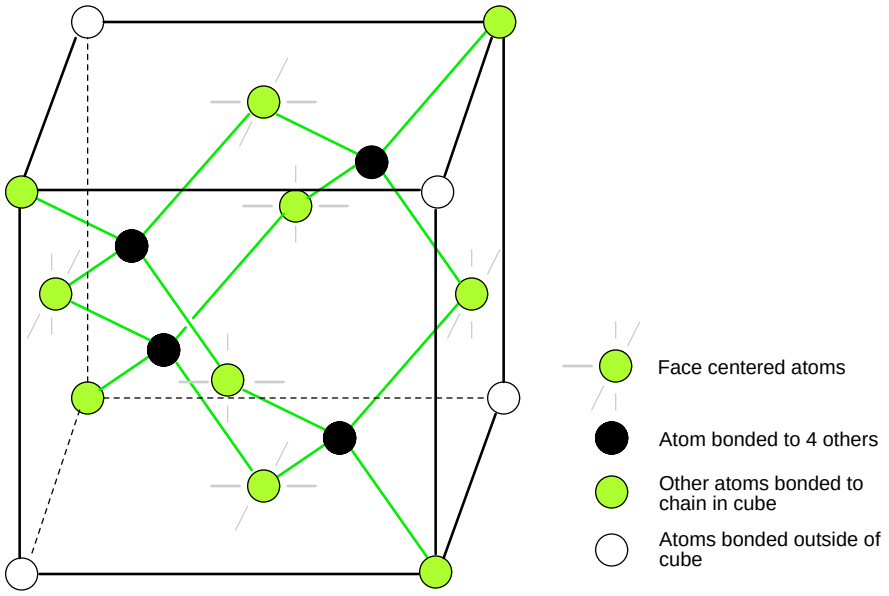


Figure 2.19: *Si, Ge, and C (diamond) form interleaved face centered cube.*

- Most semiconductor devices are manufactured from single crystals.

2.4 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 numbers describe the *allowable states* electrons may assume in an atom. To use the analogy of an amphitheater, quantum numbers describe how many rows and seats are available. Individual electrons may be described by the combination of quantum numbers, like a spectator in an amphitheater assigned to a particular row and seat.

Like spectators in an amphitheater moving between seats and 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.

Not all "leaps" are equal. Leaps between different shells require a substantial exchange of energy, but 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 are close to each other, these available energy levels form a nearly continuous *band* wherein electrons may move as illustrated in Figure 2.20

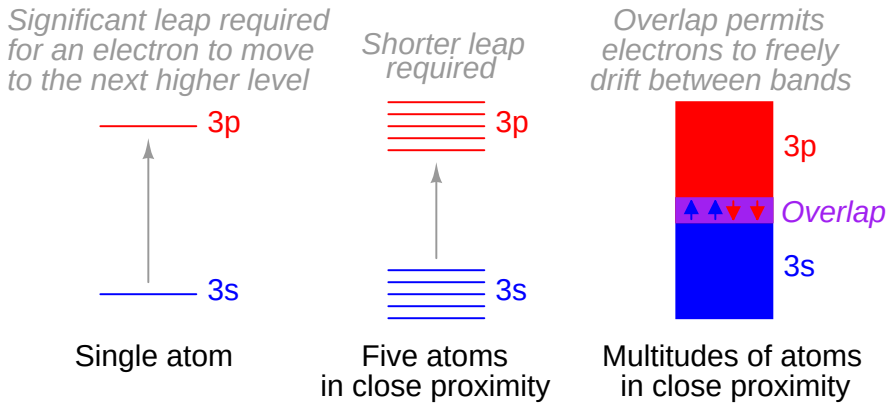


Figure 2.20: *Electron band overlap in metallic elements.*

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 of a single atom may move to what would normally be 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 close 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*). See Figure 2.21. 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.

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. (Figure 2.22)

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

It is difficult to predict the conductive properties of a substance by examining the electron configurations of its constituent atoms. Although 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:

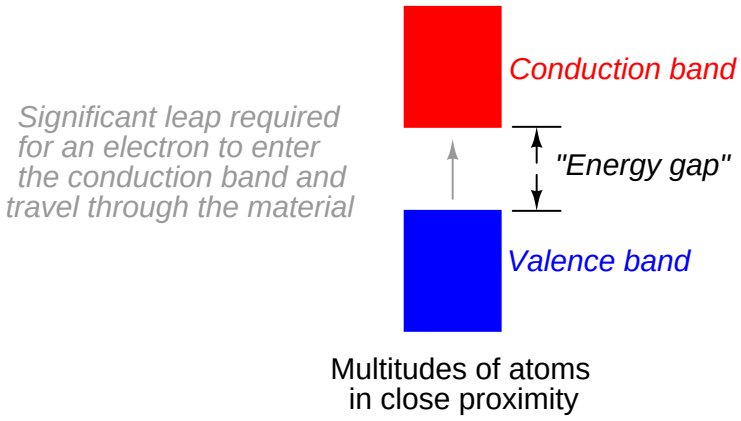


Figure 2.21: *Electron band separation in insulating substances.*

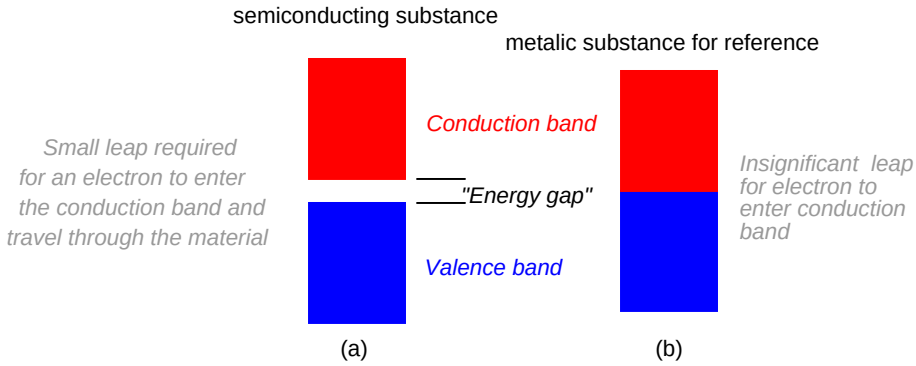


Figure 2.22: *Electron band separation in semiconducting substances, (a) multitudes of semiconducting close atoms still results in a significant band gap, (b) multitudes of close metal atoms for reference.*

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

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

- **REVIEW:**
- Energy is required to remove an electron from the valence band to a higher unoccupied band, a conduction band. More energy is required to move between shells, less between subshells.
- Since the valence and conduction bands overlap in metals, little energy removes an electron. Metals are excellent conductors.
- The large gap between the valence and conduction bands of an insulator requires high energy to remove an electron. Thus, insulators do not conduct.
- Semiconductors have a small non-overlapping gap between the valence and conduction bands. Pure semiconductors are neither good insulators nor conductors. Semiconductors are semi-conductive.

2.5 Electrons and "holes"

Pure semiconductors are relatively good insulators as compared with metals, though not nearly as good as a true insulator like glass. To be useful in semiconductor applications, the *intrinsic semiconductor* (pure undoped semiconductor) must have no more than one impurity atom in 10 billion semiconductor atoms. This is analogous to a grain of salt impurity in a railroad boxcar of sugar. Impure, or dirty semiconductors are considerably more conductive, though not as good as metals. Why might this be? To answer that question, we must look at the electron structure of such materials in Figure 2.23.

Figure 2.23 (a) shows four electrons in the valence shell of a semiconductor forming covalent bonds to four other atoms. This is a flattened, easier to draw, version of Figure 2.19. All electrons of an atom are tied up in four covalent bonds, pairs of shared electrons. Electrons are not free to move about the crystal lattice. Thus, intrinsic, pure, semiconductors are relatively good insulators as compared to metals.

Thermal energy may occasionally free an electron from the crystal lattice as in Figure 2.23 (b). This electron is free for conduction about the crystal lattice. When the electron was freed, it left an empty spot with a positive charge in the crystal lattice known as a *hole*. This hole is not fixed to the lattice; but, is free to move about. The free electron and hole both contribute

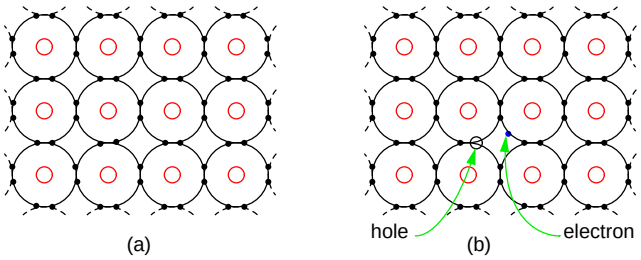


Figure 2.23: *(a) Intrinsic semiconductor is an insulator having a complete electron shell. (b) However, thermal energy can create few electron hole pairs resulting in weak conduction.*

to conduction about the crystal lattice. That is, the electron is free until it falls into a hole. This is called *recombination*. If an external electric field is applied to the semiconductor, the electrons and holes will conduct in opposite directions. Increasing temperature will increase the number of electrons and holes, decreasing the resistance. This is opposite of metals, where resistance increases with temperature by increasing the collisions of electrons with the crystal lattice. The number of electrons and holes in an intrinsic semiconductor are equal. However, both carriers do not necessarily move with the same velocity with the application of an external field. Another way of stating this is that the *mobility* is not the same for electrons and holes.

Pure semiconductors, by themselves, are not particularly useful. Though, semiconductors must be refined to a high level of purity as a starting point prior the addition of specific impurities.

Semiconductor material pure to 1 part in 10 billion, may have specific impurities added at approximately 1 part per 10 million to increase the number of carriers. The addition of a desired impurity to a semiconductor is known as *doping*. Doping increases the conductivity of a semiconductor so that it is more comparable to a metal than an insulator.

It is possible to increase the number of negative charge carriers within the semiconductor crystal lattice by doping with an electron *donor* like Phosphorus. Electron donors, also known as *N-type* dopants include elements from group VA of the periodic table: nitrogen, phosphorus, arsenic, and antimony. Nitrogen and phosphorus are N-type dopants for diamond. Phosphorus, arsenic, and antimony are used with silicon.

The crystal lattice in Figure 2.24 (b) contains atoms having four electrons in the outer shell, forming four covalent bonds to adjacent atoms. This is the anticipated crystal lattice. The addition of a phosphorus atom with five electrons in the outer shell introduces an extra electron into the lattice as compared with the silicon atom. The pentavalent impurity forms four covalent bonds to four silicon atoms with four of the five electrons, fitting into the lattice with one electron left over. Note that this spare electron is not strongly bonded to the lattice as the electrons of normal Si atoms are. It is free to move about the crystal lattice, not being bound to the Phosphorus lattice site. Since we have doped at one part phosphorus in 10 million silicon atoms, few free electrons were created compared with the numerous silicon atoms. However, many electrons were created compared with the fewer electron-hole pairs in intrinsic silicon. Application of an external electric field produces strong conduction in the doped semiconductor in the conduction band (above the valence band). A heavier doping level produces stronger

conduction. Thus, a poorly conducting intrinsic semiconductor has been converted into a good electrical conductor.

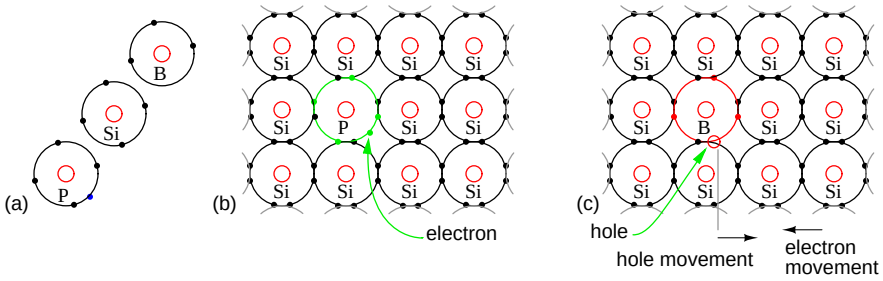


Figure 2.24: *(a) Outer shell electron configuration of donor N-type Phosphorus, Silicon (for reference), and acceptor P-type Boron. (b) N-type donor impurity creates free electron (c) P-type acceptor impurity creates hole, a positive charge carrier.*

It is also possible to introduce an impurity lacking an electron as compared with silicon, having three electrons in the valence shell as compared with four for silicon. In Figure 2.24 (c), this leaves an empty spot known as a *hole*, a positive charge carrier. The boron atom tries to bond to four silicon atoms, but only has three electrons in the valence band. In attempting to form four covalent bonds the three electrons move around trying to form four bonds. This makes the hole appear to move. Furthermore, the trivalent atom may borrow an electron from an adjacent (or more distant) silicon atom to form four covalent bonds. However, this leaves the silicon atom deficient by one electron. In other words, the hole has moved to an adjacent (or more distant) silicon atom. Holes reside in the valence band, a level below the conduction band. Doping with an electron *acceptor*, an atom which may accept an electron, creates a deficiency of electrons, the same as an excess of holes. Since holes are positive charge carriers, an electron acceptor dopant is also known as a *P-type* dopant. The P-type dopant leaves the semiconductor with an excess of holes, positive charge carriers. The P-type elements from group IIIA of the periodic table include: boron, aluminum, gallium, and indium. Boron is used as a P-type dopant for silicon and diamond semiconductors, while indium is used with germanium.

The “marble in a tube” analogy to electron conduction in Figure 2.25 relates the movement of holes with the movement of electrons. The marble represent electrons in a conductor, the tube. The movement of electrons from left to right as in a wire or N-type semiconductor is explained by an electron entering the tube at the left forcing the exit of an electron at the right. Conduction of N-type electrons occurs in the conduction band. Compare that with the movement of a hole in the valence band.

For a hole to enter at the left of Figure 2.25 (b), an electron must be removed. When moving a hole left to right, the electron must be moved right to left. The first electron is ejected from the left end of the tube so that the hole may move to the right into the tube. The electron is moving in the opposite direction of the positive hole. As the hole moves farther to the right, electrons must move left to accommodate the hole. The hole is the absence of an electron in the valence band due to P-type doping. It has a localized positive charge. To move the hole in a given direction, the valence electrons move in the opposite direction.

Electron flow in an N-type semiconductor is similar to electrons moving in a metallic wire.

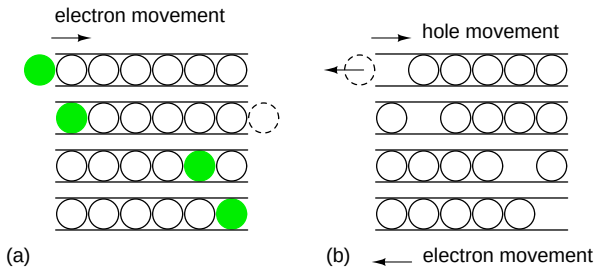


Figure 2.25: *Marble in a tube analogy: (a) Electrons move right in the conduction band as electrons enter tube. (b) Hole moves right in the valence band as electrons move left.*

The N-type dopant atoms will yield electrons available for conduction. These electrons, due to the dopant are known as *majority carriers*, for they are in the majority as compared to the very few thermal holes. If an electric field is applied across the N-type semiconductor bar in Figure 2.26 (a), electrons enter the negative (left) end of the bar, traverse the crystal lattice, and exit at right to the (+) battery terminal.

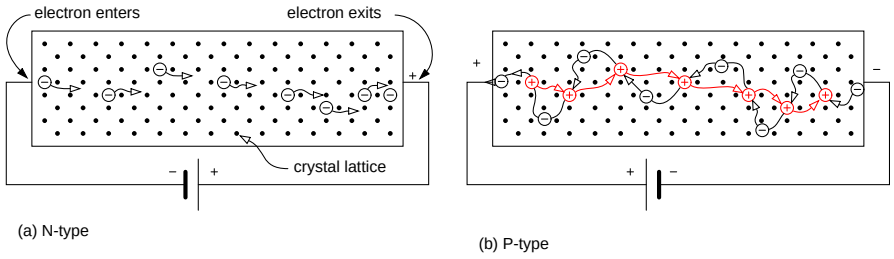


Figure 2.26: *(a) N-type semiconductor with electrons moving left to right through the crystal lattice. (b) P-type semiconductor with holes moving left to right, which corresponds to electrons moving in the opposite direction.*

Current flow in a P-type semiconductor is a little more difficult to explain. The P-type dopant, an electron acceptor, yields localized regions of positive charge known as holes. The majority carrier in a P-type semiconductor is the hole. While holes form at the trivalent dopant atom sites, they may move about the semiconductor bar. Note that the battery in Figure 2.26 (b) is reversed from (a). The positive battery terminal is connected to the left end of the P-type bar. Electron flow is out of the negative battery terminal, through the P-type bar, returning to the positive battery terminal. An electron leaving the positive (left) end of the semiconductor bar for the positive battery terminal leaves a hole in the semiconductor, that may move to the right. Holes traverse the crystal lattice from left to right. At the negative end of the bar an electron from the battery combines with a hole, neutralizing it. This makes room for another hole to move in at the positive end of the bar toward the right. Keep in mind that as holes move left to right, that it is actually electrons moving in the opposite direction that is responsible for the apparant hole movement.

The elements used to produce semiconductors are summarized in Figure 2.27. The oldest group IVA bulk semiconductor material germanium is only used to a limited extent today. Silicon based semiconductors account for about 90% of commercial production of all semiconductors. Diamond based semiconductors are a research and development activity with considerable potential at this time. Compound semiconductors not listed include silicon germanium (thin layers on Si wafers), silicon carbide and III-V compounds such as gallium arsenide. III-VI compound semiconductors include: AlN, GaN, InN, AlP, AlAs, AlSb, GaP, GaAs, GaSb, InP, InAs, InSb, $Al_xGa_{1-x}As$ and $In_xGa_{1-x}As$. Columns II and VI of periodic table, not shown in the figure, also form compound semiconductors.

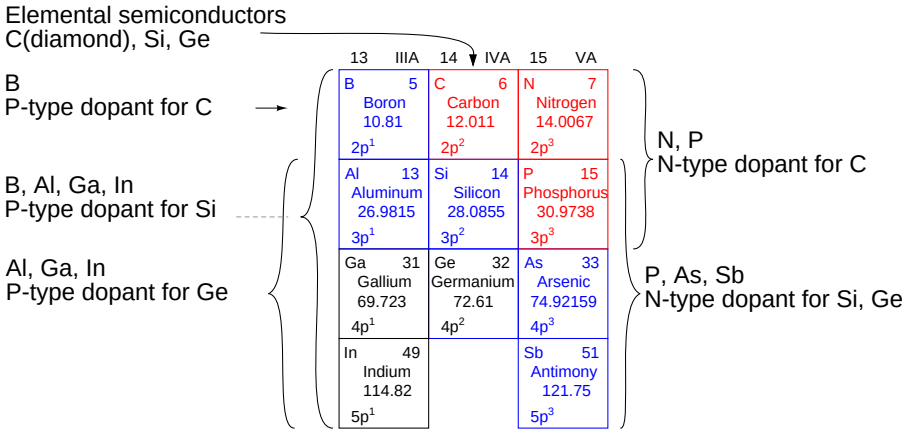


Figure 2.27: *Group IIIA P-type dopants, group IV basic semiconductor materials, and group VA N-type dopants.*

The main reason for the inclusion of the IIIA and VA groups in Figure 2.27 is to show the dopants used with the group IVA semiconductors. Group IIIA elements are acceptors, P-type dopants, which accept electrons leaving a hole in the crystal lattice, a positive carrier. Boron is the P-type dopant for diamond, and the most common dopant for silicon semiconductors. Indium is the P-type dopant for germanium.

Group VA elements are donors, N-type dopants, yielding a free electron. Nitrogen and Phosphorus are suitable N-type dopants for diamond. Phosphorus and arsenic are the most commonly used N-type dopants for silicon; though, antimony can be used.

- **REVIEW:**
- Intrinsic semiconductor materials, pure to 1 part in 10 billion, are poor conductors.
- N-type semiconductor is doped with a pentavalent impurity to create free electrons. Such a material is conductive. The electron is the majority carrier.
- P-type semiconductor, doped with a trivalent impurity, has an abundance of free holes. These are positive charge carriers. The P-type material is conductive. The hole is the majority carrier.

- Most semiconductors are based on elements from group IVA of the periodic table, silicon being the most prevalent. Germanium is all but obsolete. Carbon (diamond) is being developed.
- Compound semiconductors such as silicon carbide (group IVA) and gallium arsenide (group III-V) are widely used.

2.6 The P-N junction

If a block of P-type semiconductor is placed in contact with a block of N-type semiconductor in Figure 2.28(a), the result is of no value. We have two conductive blocks in contact with each other, showing no unique properties. The problem is two separate and distinct crystal bodies. The number of electrons is balanced by the number of protons in both blocks. Thus, neither block has any net charge.

However, a single semiconductor crystal manufactured with P-type material at one end and N-type material at the other in Figure 2.28 (b) has some unique properties. The P-type material has positive majority charge carriers, holes, which are free to move about the crystal lattice. The N-type material has mobile negative majority carriers, electrons. Near the junction, the N-type material electrons diffuse across the junction, combining with holes in P-type material. The region of the P-type material near the junction takes on a net negative charge because of the electrons attracted. Since electrons departed the N-type region, it takes on a localized positive charge. The thin layer of the crystal lattice between these charges has been depleted of majority carriers, thus, is known as the *depletion region*. It becomes nonconductive intrinsic semiconductor material. In effect, we have nearly an insulator separating the conductive P and N doped regions.

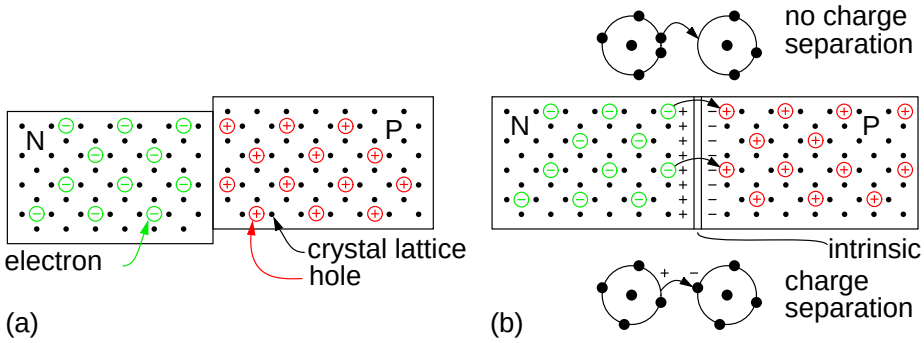


Figure 2.28: *(a) Blocks of P and N semiconductor in contact have no exploitable properties. (b) Single crystal doped with P and N type impurities develops a potential barrier.*

This separation of charges at the PN junction constitutes a potential barrier. This potential barrier must be overcome by an external voltage source to make the junction conduct. The formation of the junction and potential barrier happens during the manufacturing process. The magnitude of the potential barrier is a function of the materials used in manufacturing. Silicon PN junctions have a higher potential barrier than germanium junctions.

In Figure 2.29(a) the battery is arranged so that the negative terminal supplies electrons to the N-type material. These electrons diffuse toward the junction. The positive terminal removes electrons from the P-type semiconductor, creating holes that diffuse toward the junction. If the battery voltage is great enough to overcome the junction potential (0.6V in Si), the N-type electrons and P-holes combine annihilating each other. This frees up space within the lattice for more carriers to flow toward the junction. Thus, currents of N-type and P-type majority carriers flow toward the junction. The recombination at the junction allows a battery current to flow through the PN junction diode. Such a junction is said to be *forward biased*.

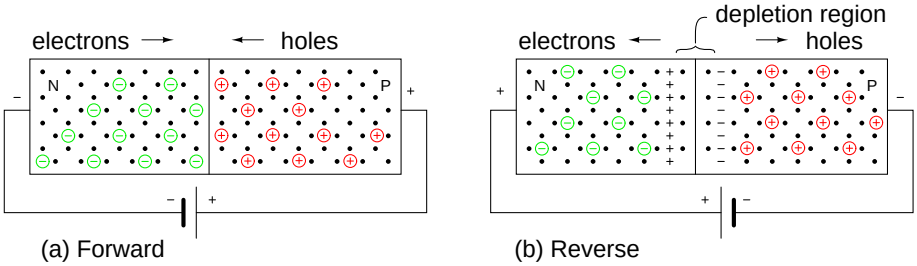


Figure 2.29: *(a) Forward battery bias repels carriers toward junction, where recombination results in battery current. (b) Reverse battery bias attracts carriers toward battery terminals, away from junction. Depletion region thickness increases. No sustained battery current flows.*

If the battery polarity is reversed as in Figure 2.29(b) majority carriers are attracted away from the junction toward the battery terminals. The positive battery terminal attracts N-type majority carriers, electrons, away from the junction. The negative terminal attracts P-type majority carriers, holes, away from the junction. This increases the thickness of the nonconducting depletion region. There is no recombination of majority carriers; thus, no conduction. This arrangement of battery polarity is called *reverse bias*.

The diode schematic symbol is illustrated in Figure 2.30(b) corresponding to the doped semiconductor bar at (a). The diode is a *unidirectional* device. Electron current only flows in one direction, against the arrow, corresponding to forward bias. The cathode, bar, of the diode symbol corresponds to N-type semiconductor. The anode, arrow, corresponds to the P-type semiconductor. To remember this relationship, **N**ot-pointing (bar) on the symbol corresponds to **N**-type semiconductor. **P**ointing (arrow) corresponds to **P**-type.

If a diode is forward biased as in Figure 2.30(a), current will increase slightly as voltage is increased from 0 V. In the case of a silicon diode a measurable current flows when the voltage approaches 0.6 V in Figure 2.30(c). As the voltage increases past 0.6 V, current increases considerably after the knee. Increasing the voltage well beyond 0.7 V may result in high enough current to destroy the diode. The forward voltage, V_F, is a characteristic of the semiconductor: 0.6 to 0.7 V for silicon, 0.2 V for germanium, a few volts for Light Emitting Diodes (LED). The forward current ranges from a few mA for point contact diodes to 100 mA for small signal diodes to tens or thousands of amperes for power diodes.

If the diode is reverse biased, only the leakage current of the intrinsic semiconductor flows. This is plotted to the left of the origin in Figure 2.30(c). This current will only be as high as 1 μA for the most extreme conditions for silicon small signal diodes. This current does not

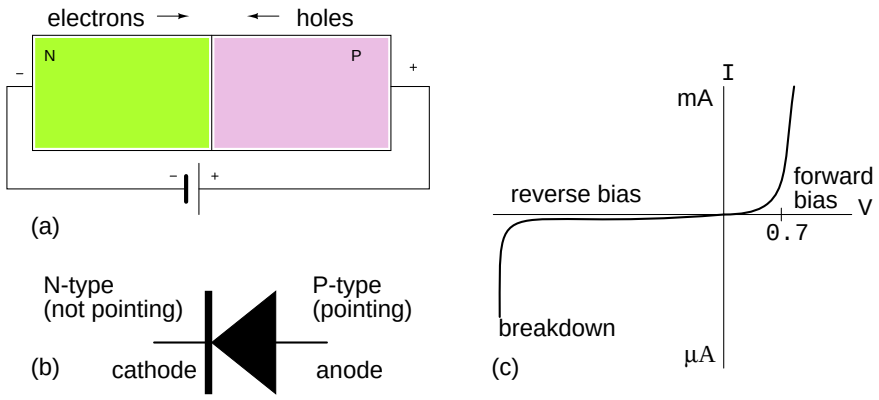


Figure 2.30: *(a) Forward biased PN junction, (b) Corresponding diode schematic symbol (c) Silicon Diode I vs V characteristic curve.*

increase appreciably with increasing reverse bias until the diode breaks down. At breakdown, the current increases so greatly that the diode will be destroyed unless a high series resistance limits current. We normally select a diode with a higher reverse voltage rating than any applied voltage to prevent this. Silicon diodes are typically available with reverse break down ratings of 50, 100, 200, 400, 800 V and higher. It is possible to fabricate diodes with a lower rating of a few volts for use as voltage standards.

We previously mentioned that the reverse leakage current of under a μA for silicon diodes was due to conduction of the intrinsic semiconductor. This is the leakage that can be explained by theory. Thermal energy produces few electron hole pairs, which conduct leakage current until recombination. In actual practice this predictable current is only part of the leakage current. Much of the leakage current is due to surface conduction, related to the lack of cleanliness of the semiconductor surface. Both leakage currents increase with increasing temperature, approaching a μA for small silicon diodes.

For germanium, the leakage current is orders of magnitude higher. Since germanium semiconductors are rarely used today, this is not a problem in practice.

- **REVIEW:**

- PN junctions are fabricated from a monocrystalline piece of semiconductor with both a P-type and N-type region in proximity at a junction.

- The transfer of electrons from the N side of the junction to holes annihilated on the P side of the junction produces a barrier voltage. This is 0.6 to 0.7 V in silicon, and varies with other semiconductors.

- A forward biased PN junction conducts a current once the barrier voltage is overcome. The external applied potential forces majority carriers toward the junction where recombination takes place, allowing current flow.

- A reverse biased PN junction conducts almost no current. The applied reverse bias attracts majority carriers away from the junction. This increases the thickness of the nonconducting depletion region.
- Reverse biased PN junctions show a temperature dependent reverse leakage current. This is less than a μA in small silicon diodes.

2.7 Junction diodes

There were some historic crude, but usable semiconductor rectifiers before high purity materials were available. Ferdinand Braun invented a lead sulfide, PbS, based point contact rectifier in 1874. Cuprous oxide rectifiers were used as power rectifiers in 1924. The forward voltage drop is 0.2 V. The linear characteristic curve perhaps is why Cu_2O was used as a rectifier for the AC scale on D'Arsonval based multimeters. This diode is also photosensitive.

Selenium oxide rectifiers were used before modern power diode rectifiers became available. These and the Cu_2O rectifiers were polycrystalline devices. Photoelectric cells were once made from Selenium.

Before the modern semiconductor era, an early diode application was as a radio frequency *detector*, which recovered audio from a radio signal. The "semiconductor" was a polycrystalline piece of the mineral galena, lead sulfide, PbS. A pointed metallic wire known as a *cat whisker* was brought in contact with a spot on a crystal within the polycrystalline mineral. (Figure 2.31) The operator labored to find a "sensitive" spot on the galena by moving the cat whisker about. Presumably there were P and N-type spots randomly distributed throughout the crystal due to the variability of uncontrolled impurities. Less often the mineral iron pyrites, fools gold, was used, as was the mineral carborundum, silicon carbide, SiC, another detector, part of a *foxhole radio*, consisted of a sharpened pencil lead bound to a bent safety pin, touching a rusty blue-blade disposable razor blade. These all required searching for a sensitive spot, easily lost because of vibration.

Replacing the mineral with an N-doped semiconductor (Figure 2.32(a)) makes the whole surface sensitive, so that searching for a sensitive spot was no longer required. This device was perfected by G.W.Pickard in 1906. The pointed metal contact produced a localized P-type region within the semiconductor. The metal point was fixed in place, and the whole *point contact diode* encapsulated in a cylindrical body for mechanical and electrical stability. (Figure 2.32(d)) Note that the cathode bar on the schematic corresponds to the bar on the physical package.

Silicon point contact diodes made an important contribution to radar in World War II, detecting giga-hertz radio frequency echo signals in the radar receiver. The concept to be made clear is that the point contact diode preceded the junction diode and modern semiconductors by several decades. Even to this day, the point contact diode is a practical means of microwave frequency detection because of its low capacitance. Germanium point contact diodes were once more readily available than they are today, being preferred for the lower 0.2 V forward voltage in some applications like self-powered crystal radios. Point contact diodes, though sensitive to a wide bandwidth, have a low current capability compared with junction diodes.

Most diodes today are silicon junction diodes. The cross-section in Figure 2.32(b) looks a bit more complex than a simple PN junction; though, it is still a PN junction. Starting at the cathode connection, the N^+ indicates this region is heavily doped, having nothing to do with

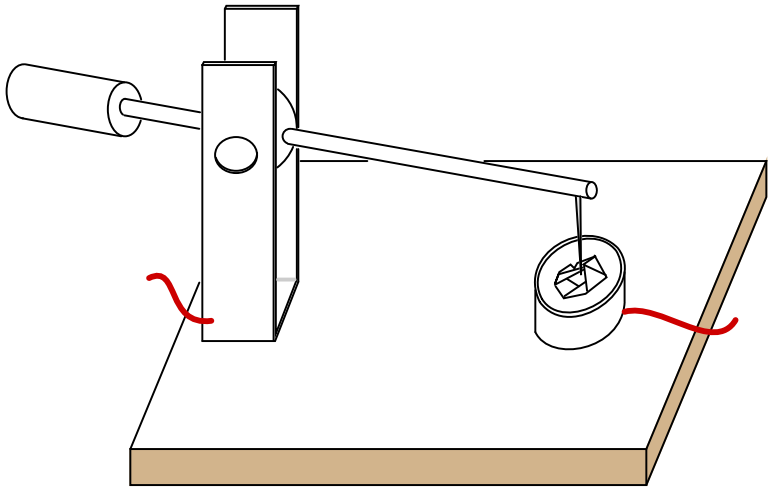


Figure 2.31: *Crystal detector*

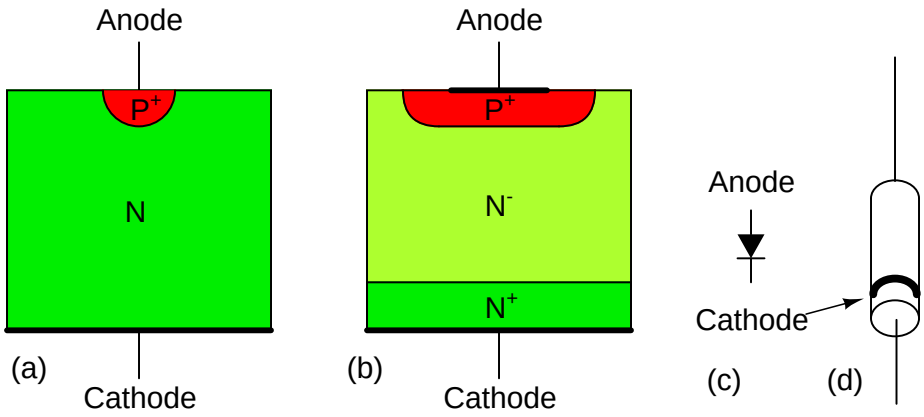


Figure 2.32: *Silicon diode cross-section: (a) point contact diode, (b) junction diode, (c) schematic symbol, (d) small signal diode package.*

polarity. This reduces the series resistance of the diode. The N^- region is lightly doped as indicated by the (-). Light doping produces a diode with a higher reverse breakdown voltage, important for high voltage power rectifier diodes. Lower voltage diodes, even low voltage power rectifiers, would have lower forward losses with heavier doping. The heaviest level of doping produce zener diodes designed for a low reverse breakdown voltage. However, heavy doping increases the reverse leakage current. The P^+ region at the anode contact is heavily doped P-type semiconductor, a good contact strategy. Glass encapsulated small signal junction diodes are capable of 10's to 100's of mA of current. Plastic or ceramic encapsulated power rectifier diodes handle to 1000's of amperes of current.

- **REVIEW:**
- Point contact diodes have superb high frequency characteristics, usable well into the microwave frequencies.
- Junction diodes range in size from small signal diodes to power rectifiers capable of 1000's of amperes.
- The level of doping near the junction determines the reverse breakdown voltage. Light doping produces a high voltage diode. Heavy doping produces a lower breakdown voltage, and increases reverse leakage current. Zener diodes have a lower breakdown voltage because of heavy doping.

2.8 Bipolar junction transistors

The *bipolar junction transistor* (BJT) was named because its operation involves conduction by two carriers: electrons and holes in the same crystal. The first bipolar transistor was invented at Bell Labs by William Shockley, Walter Brattain, and John Bardeen so late in 1947 that it was not published until 1948. Thus, many texts differ as to the date of invention. Brattain fabricated a germanium *point contact transistor*, bearing some resemblance to a point contact diode. Within a month, Shockley had a more practical *junction transistor*, which we describe in following paragraphs. They were awarded the Nobel Prize in Physics in 1956 for the transistor.

The bipolar junction transistor shown in Figure 2.33(a) is an NPN three layer semiconductor sandwich with an *emitter* and *collector* at the ends, and a *base* in between. It is as if a third layer were added to a two layer diode. If this were the only requirement, we would have no more than a pair of back-to-back diodes. In fact, it is far easier to build a pair of back-to-back diodes. The key to the fabrication of a bipolar junction transistor is to make the middle layer, the base, as thin as possible without shorting the outside layers, the emitter and collector. We cannot over emphasize the importance of the thin base region.

The device in Figure 2.33(a) has a pair of junctions, emitter to base and base to collector, and two depletion regions.

It is customary to reverse bias the base-collector junction of a bipolar junction transistor as shown in (Figure 2.33(b). Note that this increases the width of the depletion region. The reverse bias voltage could be a few volts to tens of volts for most transistors. There is no current flow, except leakage current, in the collector circuit.

In Figure 2.34(a), a voltage source has been added to the emitter base circuit. Normally we forward bias the emitter-base junction, overcoming the 0.6 V potential barrier. This is similar

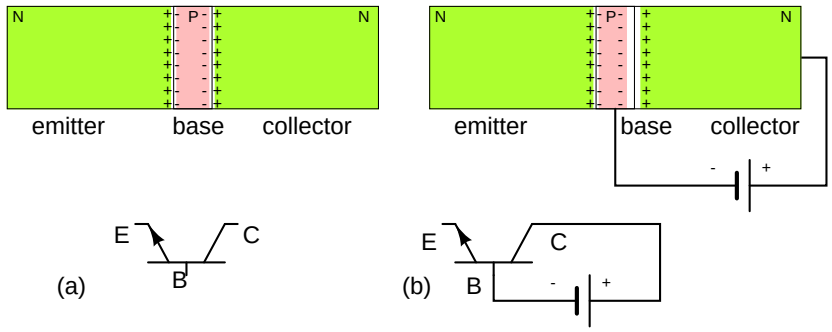


Figure 2.33: *(a) NPN junction bipolar transistor. (b) Apply reverse bias to collector base junction.*

to forward biasing a junction diode. This voltage source needs to exceed 0.6 V for majority carriers (electrons for NPN) to flow from the emitter into the base becoming minority carriers in the P-type semiconductor.

If the base region were thick, as in a pair of back-to-back diodes, all the current entering the base would flow out the base lead. In our NPN transistor example, electrons leaving the emitter for the base would combine with holes in the base, making room for more holes to be created at the (+) battery terminal on the base as electrons exit.

However, the base is manufactured thin. A few majority carriers in the emitter, injected as minority carriers into the base, actually recombine. See Figure 2.34(b). Few electrons injected by the emitter into the base of an NPN transistor fall into holes. Also, few electrons entering the base flow directly through the base to the positive battery terminal. Most of the emitter current of electrons diffuses through the thin base into the collector. Moreover, modulating the small base current produces a larger change in collector current. If the base voltage falls below approximately 0.6 V for a silicon transistor, the large emitter-collector current ceases to flow.

In Figure 2.35 we take a closer look at the current amplification mechanism. We have an enlarged view of an NPN junction transistor with emphasis on the thin base region. Though not shown, we assume that external voltage sources 1) forward bias the emitter-base junction, 2) reverse bias the base-collector junction. Electrons, majority carriers, enter the emitter from the (-) battery terminal. The base current flow corresponds to electrons leaving the base terminal for the (+) battery terminal. This is but a small current compared to the emitter current.

Majority carriers within the N-type emitter are electrons, becoming minority carriers when entering the P-type base. These electrons face four possible fates entering the thin P-type base. A few at Figure 2.35(a) fall into holes in the base that contributes to base current flow to the (+) battery terminal. Not shown, holes in the base may diffuse into the emitter and combine with electrons, contributing to base terminal current. Few at (b) flow on through the base to the (+) battery terminal as if the base were a resistor. Both (a) and (b) contribute to the very small base current flow. Base current is typically 1% of emitter or collector current for small signal transistors. Most of the emitter electrons diffuse right through the thin base (c) into the base-collector depletion region. Note the polarity of the depletion region surrounding the electron at (d). The strong electric field sweeps the electron rapidly into the collector. The

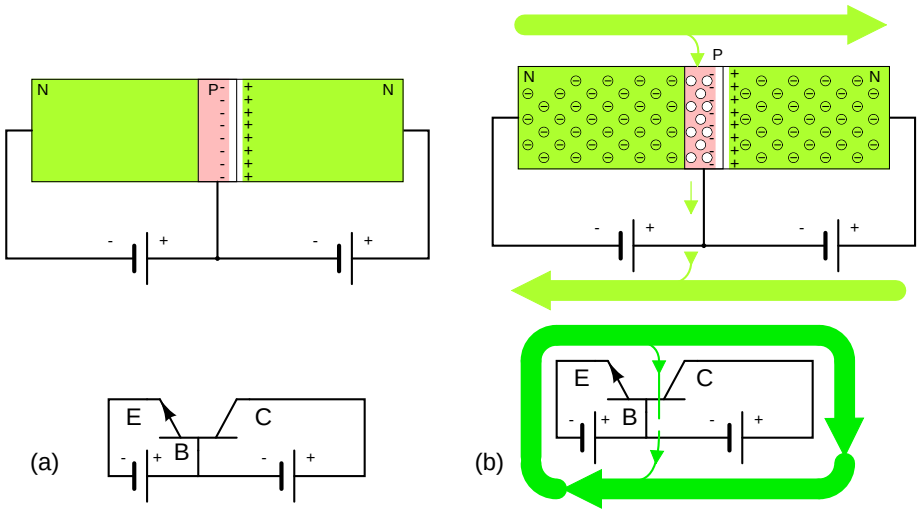


Figure 2.34: *NPN junction bipolar transistor with reverse biased collector-base: (a) Adding forward bias to base-emitter junction, results in (b) a small base current and large emitter and collector currents.*

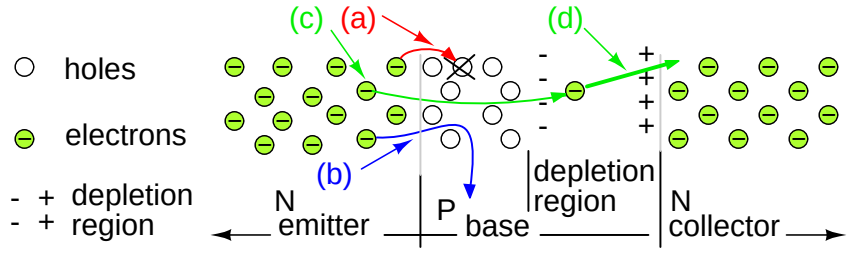


Figure 2.35: *Disposition of electrons entering base: (a) Lost due to recombination with base holes. (b) Flows out base lead. (c) Most diffuse from emitter through thin base into base-collector depletion region, and (d) are rapidly swept by the strong depletion region electric field into the collector.*

strength of the field is proportional to the collector battery voltage. Thus 99% of the emitter current flows into the collector. It is controlled by the base current, which is 1% of the emitter current. This is a potential current gain of 99, the ratio of I_C/I_B , also known as beta, β.

This magic, the diffusion of 99% of the emitter carriers through the base, is only possible if the base is very thin. What would be the fate of the base minority carriers in a base 100 times thicker? One would expect the recombination rate, electrons falling into holes, to be much higher. Perhaps 99%, instead of 1%, would fall into holes, never getting to the collector. The second point to make is that the base current may control 99% of the emitter current, only if 99% of the emitter current diffuses into the collector. If it all flows out the base, no control is possible.

Another feature accounting for passing 99% of the electrons from emitter to collector is that real bipolar junction transistors use a small heavily doped emitter. The high concentration of emitter electrons forces many electrons to diffuse into the base. The lower doping concentration in the base means fewer holes diffuse into the emitter, which would increase the base current. Diffusion of carriers from emitter to base is strongly favored.

The thin base and the heavily doped emitter help keep the *emitter efficiency* high, 99% for example. This corresponds to 100% emitter current splitting between the base as 1% and the collector as 99%. The emitter efficiency is known as $\alpha = I_C/I_E$.

Bipolar junction transistors are available as PNP as well as NPN devices. We present a comparison of these two in Figure 2.36. The difference is the polarity of the base emitter diode junctions, as signified by the direction of the schematic symbol emitter arrow. It points in the same direction as the anode arrow for a junction diode, against electron current flow. See diode junction, Figure 2.30. The point of the arrow and bar correspond to P-type and N-type semiconductors, respectively. For NPN and PNP emitters, the arrow points away and toward the base respectively. There is no schematic arrow on the collector. However, the base-collector junction is the same polarity as the base-emitter junction compared to a diode. Note, we speak of diode, not power supply, polarity.

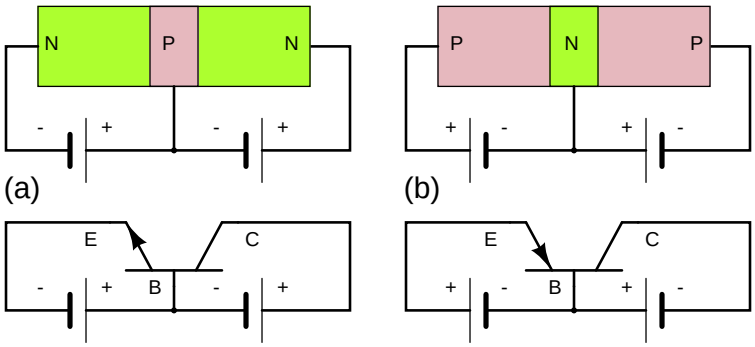


Figure 2.36: *Compare NPN transistor at (a) with the PNP transistor at (b). Note direction of emitter arrow and supply polarity.*

The voltage sources for PNP transistors are reversed compared with an NPN transistors as shown in Figure 2.36. The base-emitter junction must be forward biased in both cases. The base on a PNP transistor is biased negative (b) compared with positive (a) for an NPN. In both

cases the base-collector junction is reverse biased. The PNP collector power supply is negative compared with positive for an NPN transistor.

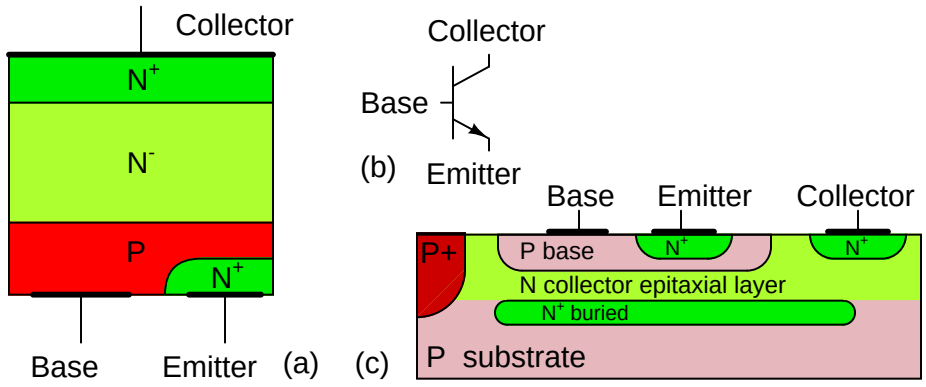


Figure 2.37: *Bipolar junction transistor: (a) discrete device cross-section, (b) schematic symbol, (c) integrated circuit cross-section.*

Note that the BJT in Figure 2.37(a) has heavy doping in the emitter as indicated by the N^+ notation. The base has a normal P-dopant level. The base is much thinner than the not-to-scale cross-section shows. The collector is lightly doped as indicated by the N^- notation. The collector needs to be lightly doped so that the collector-base junction will have a high breakdown voltage. This translates into a high allowable collector power supply voltage. Small signal silicon transistors have a 60-80 V breakdown voltage. Though, it may run to hundreds of volts for high voltage transistors. The collector also needs to be heavily doped to minimize ohmic losses if the transistor must handle high current. These contradicting requirements are met by doping the collector more heavily at the metallic contact area. The collector near the base is lightly doped as compared with the emitter. The heavy doping in the emitter gives the emitter-base a low approximate 7 V breakdown voltage in small signal transistors. The heavily doped emitter makes the emitter-base junction have zener diode like characteristics in reverse bias.

The BJT *die*, a piece of a sliced and diced semiconductor wafer, is mounted collector down to a metal case for power transistors. That is, the metal case is electrically connected to the collector. A small signal die may be encapsulated in epoxy. In power transistors, aluminum bonding wires connect the base and emitter to package leads. Small signal transistor dies may be mounted directly to the lead wires. Multiple transistors may be fabricated on a single die called an *integrated circuit*. Even the collector may be bonded out to a lead instead of the case. The integrated circuit may contain internal wiring of the transistors and other integrated components. The integrated BJT shown in Figure 2.37(c) is much thinner than the "not to scale" drawing. The P^+ region isolates multiple transistors in a single die. An aluminum metalization layer (not shown) interconnects multiple transistors and other components. The emitter region is heavily doped, N^+ compared to the base and collector to improve emitter efficiency.

Discrete PNP transistors are almost as high quality as the NPN counterpart. However, integrated PNP transistors are not nearly a good as the NPN variety within the same integrated

circuit die. Thus, integrated circuits use the NPN variety as much as possible.

- **REVIEW:**
- Bipolar transistors conduct current using both electrons and holes in the same device.
- Operation of a bipolar transistor as a current amplifier requires that the collector-base junction be reverse biased and the emitter-base junction be forward biased.
- A transistor differs from a pair of back to back diodes in that the base, the center layer, is very thin. This allows majority carriers from the emitter to diffuse as minority carriers through the base into the depletion region of the base-collector junction, where the strong electric field collects them.
- Emitter efficiency is improved by heavier doping compared with the collector. Emitter efficiency: α = I_C/I_E, 0.99 for small signal devices
- Current gain is β=I_C/I_B, 100 to 300 for small signal transistors.

2.9 Junction field-effect transistors

The field effect transistor was proposed by Julius Lilienfeld in US patents in 1926 and 1933 (1,900,018). Moreover, Shockley, Brattain, and Bardeen were investigating the field effect transistor in 1947. Though, the extreme difficulties sidetracked them into inventing the bipolar transistor instead. Shockley's field effect transistor theory was published in 1952. However, the materials processing technology was not mature enough until 1960 when John Atalla produced a working device.

A *field effect transistor* (FET) is a *unipolar* device, conducting a current using only one kind of charge carrier. If based on an N-type slab of semiconductor, the carriers are electrons. Conversely, a P-type based device uses only holes.

At the circuit level, field effect transistor operation is simple. A voltage applied to the *gate*, input element, controls the resistance of the *channel*, the unipolar region between the gate regions. (Figure 2.38) In an N-channel device, this is a lightly doped N-type slab of silicon with terminals at the ends. The *source* and *drain* terminals are analogous to the emitter and collector, respectively, of a BJT. In an N-channel device, a heavy P-type region on both sides of the center of the slab serves as a control electrode, the gate. The gate is analogous to the base of a BJT.

"Cleanliness is next to godliness" applies to the manufacture of field effect transistors. Though it is possible to make bipolar transistors outside of a *clean room*, it is a necessity for field effect transistors. Even in such an environment, manufacture is tricky because of contamination control issues. The unipolar field effect transistor is conceptually simple, but difficult to manufacture. Most transistors today are a metal oxide semiconductor variety (later section) of the field effect transistor contained within integrated circuits. However, discrete JFET devices are available.

A properly biased N-channel junction field effect transistor (JFET) is shown in Figure 2.38. The gate constitutes a diode junction to the source to drain semiconductor slab. The gate is reverse biased. If a voltage (or an ohmmeter) were applied between the source and drain, the

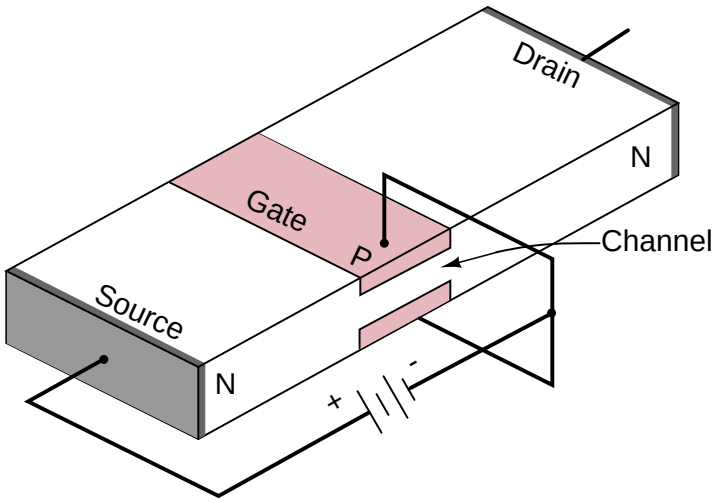


Figure 2.38: *Junction field effect transistor cross-section.*

N-type bar would conduct in either direction because of the doping. Neither gate nor gate bias is required for conduction. If a gate junction is formed as shown, conduction can be controlled by the degree of reverse bias.

Figure 2.39(a) shows the depletion region at the gate junction. This is due to diffusion of holes from the P-type gate region into the N-type channel, giving the charge separation about the junction, with a non-conductive depletion region at the junction. The depletion region extends more deeply into the channel side due to the heavy gate doping and light channel doping.

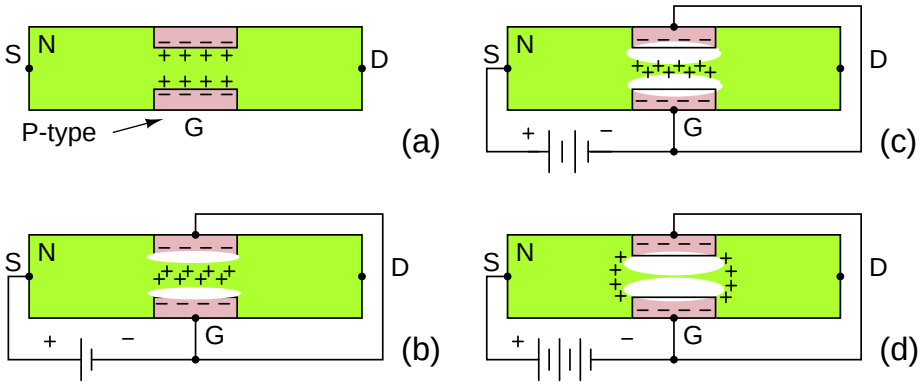


Figure 2.39: *N-channel JFET: (a) Depletion at gate diode. (b) Reverse biased gate diode increases depletion region. (c) Increasing reverse bias enlarges depletion region. (d) Increasing reverse bias pinches-off the S-D channel.*

The thickness of the depletion region can be increased Figure 2.39(b) by applying moderate reverse bias. This increases the resistance of the source to drain channel by narrowing the channel. Increasing the reverse bias at (c) increases the depletion region, decreases the channel width, and increases the channel resistance. Increasing the reverse bias V_{GS} at (d) will *pinch-off* the channel current. The channel resistance will be very high. This V_{GS} at which pinch-off occurs is V_P, the pinch-off voltage. It is typically a few volts. In summation, the channel resistance can be controlled by the degree of reverse biasing on the gate.

The source and drain are interchangeable, and the source to drain current may flow in either direction for low level drain battery voltage (¡ 0.6 V). That is, the drain battery may be replaced by a low voltage AC source. For a high drain power supply voltage, to 10's of volts for small signal devices, the polarity must be as indicated in Figure 2.40(a). This drain power supply, not shown in previous figures, distorts the depletion region, enlarging it on the drain side of the gate. This is a more correct representation for common DC drain supply voltages, from a few to tens of volts. As drain voltage V_{DS} increased,the gate depletion region expands toward the drain. This increases the length of the narrow channel, increasing its resistance a little. We say "a little" because large resistance changes are due to changing gate bias. Figure 2.40(b) shows the schematic symbol for an N-channel field effect transistor compared to the silicon cross-section at (a). The gate arrow points in the same direction as a junction diode. The "pointing" arrow and "non-pointing" bar correspond to P and N-type semiconductors, respectively.

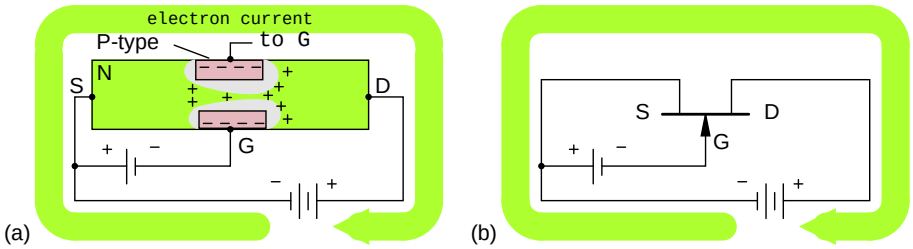


Figure 2.40: *N-channel JFET electron current flow from source to drain in (a) cross-section, (b) schematic symbol.*

Figure 2.40 shows a large electron current flow from (-) battery terminal, to FET source, out the drain, returning to the (+) battery terminal. This current flow may be controlled by varying the gate voltage. A load in series with the battery sees an amplified version of the changing gate voltage.

P-channel field effect transistors are also available. The channel is made of P-type material. The gate is a heavily dopped N-type region. All the voltage sources are reversed in the P-channel circuit (Figure 2.41) as compared with the more popular N-channel device. Also note, the arrow points out of the gate of the schematic symbol (b) of the P-channel field effect transistor.

As the positive gate bias voltage is increased, the resistance of the P-channel increases, decreasing the current flow in the drain circuit.

Discrete devices are manufactured with the cross-section shown in Figure 2.42. The cross-section, oriented so that it corresponds to the schematic symbol, is upside down with respect

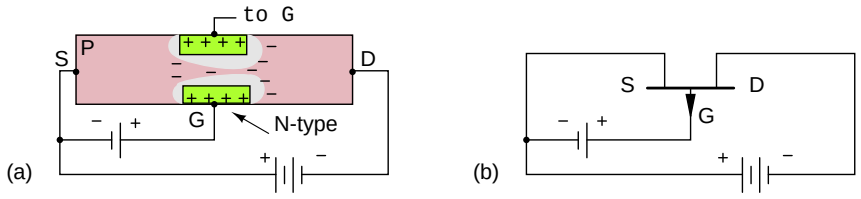


Figure 2.41: *P-channel JFET: (a) N-type gate, P-type channel, reversed voltage sources compared with N-channel device. (b) Note reversed gate arrow and voltage sources on schematic.*

to a semiconductor wafer. That is, the gate connections are on the top of the wafer. The gate is heavily doped, P^+, to diffuse holes well into the channel for a large depletion region. The source and drain connections in this N-channel device are heavily doped, N^+ to lower connection resistance. However, the channel surrounding the gate is lightly doped to allow holes from the gate to diffuse deeply into the channel. That is the N^- region.

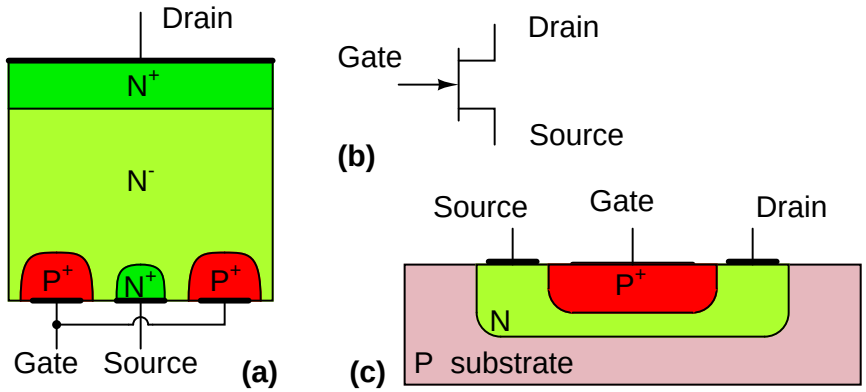


Figure 2.42: *Junction field effect transistor: (a) Discrete device cross-section, (b) schematic symbol, (c) integrated circuit device cross-section.*

All three FET terminals are available on the top of the die for the integrated circuit version so that a metalization layer (not shown) can interconnect multiple components. (Figure 2.42(c)) Integrated circuit FET's are used in analog circuits for the high gate input resistance.. The N-channel region under the gate must be very thin so that the intrinsic region about the gate can control and pinch-off the channel. Thus, gate regions on both sides of the channel are not necessary.

The static induction field effect transistor (SIT) is a short channel device with a buried gate. (Figure 2.43) It is a power device, as opposed to a small signal device. The low gate resistance and low gate to source capacitance make for a fast switching device. The SIT is capable of hundreds of amps and thousands of volts. And, is said to be capable of an incredible frequency of 10 gHz.[25]

The *Metal semiconductor field effect transistor (MESFET)* is similar to a JFET except the

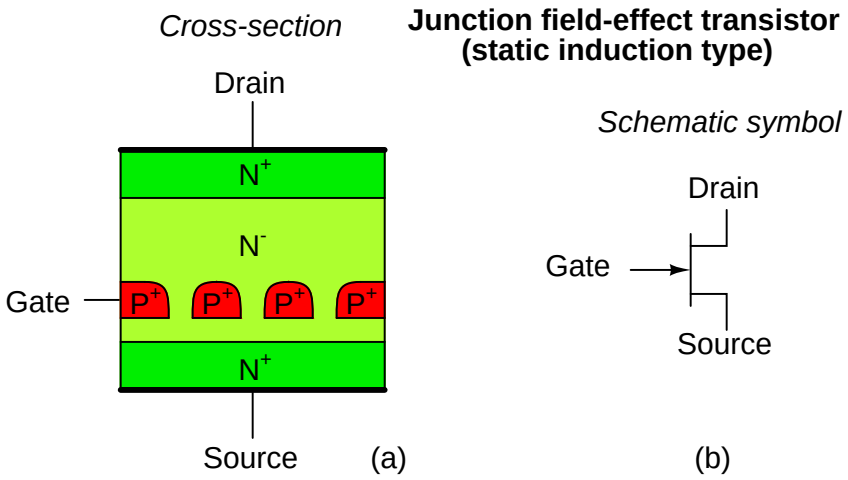


Figure 2.43: *Junction field effect transistor (static induction type): (a) Cross-section, (b) schematic symbol.*

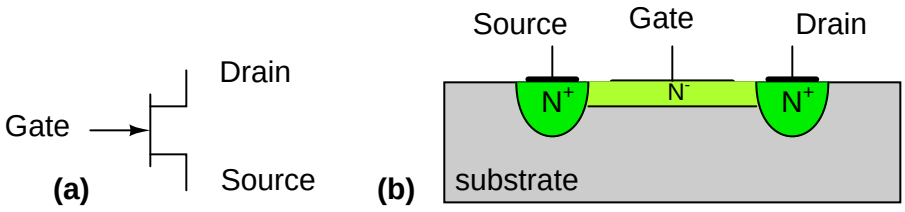


Figure 2.44: *Metal semiconductor field effect transistor (MESFET): (a) schematic symbol, (b) cross-section.*

gate is a schottky diode instead of a junction diode. A *schottky diode* is a metal rectifying contact to a semiconductor compared with a more common ohmic contact. In Figure 2.44 the source and drain are heavily doped (N^+). The channel is lightly doped (N^-). MESFET's are higher speed than JFET's. The MESET is a depletion mode device, normally on, like a JFET. They are used as microwave power amplifiers to 30 gHz. MESFET's can be fabricated from silicon, gallium arsenide, indium phosphide, silicon carbide, and the diamond allotrope of carbon.

- **REVIEW:**
- The unipolar junction field effect transistor (FET or JFET) is so called because conduction in the channel is due to one type of carrier
- The JFET source, gate, and drain correspond to the BJT's emitter, base, and collector, respectively.
- Application of reverse bias to the gate varies the channel resistance by expanding the gate diode depletion region.

2.10 Insulated-gate field-effect transistors (MOSFET)

The *insulated-gate field-effect transistor* (IGFET), also known as the *metal oxide field effect transistor* (MOSFET), is a derivative of the field effect transistor (FET). Today, most transistors are of the MOSFET type as components of digital integrated circuits. Though discrete BJT's are more numerous than discrete MOSFET's. The MOSFET transistor count within an integrated circuit may approach hundreds of a million. The dimensions of individual MOSFET devices are under a micron, decreasing every 18 months. Much larger MOSFET's are capable of switching nearly 100 amperes of current at low voltages; some handle nearly 1000 V at lower currents. These devices occupy a good fraction of a square centimeter of silicon. MOSFET's find much wider application than JFET's. However, MOSFET power devices are not as widely used as bipolar junction transistors at this time.

The MOSFET has source, gate, and drain terminals like the FET. However, the gate lead does not make a direct connection to the silicon compared with the case for the FET. The MOSFET gate is a metallic or polysilicon layer atop a silicon dioxide insulator. The gate bears a resemblance to a *metal oxide semiconductor* (MOS) capacitor in Figure 2.45. When charged, the plates of the capacitor take on the charge polarity of the respective battery terminals. The lower plate is P-type silicon from which electrons are repelled by the negative (-) battery terminal toward the oxide, and attracted by the positive (+) top plate. This excess of electrons near the oxide creates an inverted (excess of electrons) channel under the oxide. This channel is also accompanied by a depletion region isolating the channel from the bulk silicon substrate.

In Figure 2.46 (a) the MOS capacitor is placed between a pair of N-type diffusions in a P-type substrate. With no charge on the capacitor, no bias on the gate, the N-type diffusions, the source and drain, remain electrically isolated.

A positive bias applied to the gate, charges the capacitor (the gate). The gate atop the oxide takes on a positive charge from the gate bias battery. The P-type substrate below the gate takes on a negative charge. An inversion region with an excess of electrons forms below the gate oxide. This region now connects the source and drain N-type regions, forming a continuous

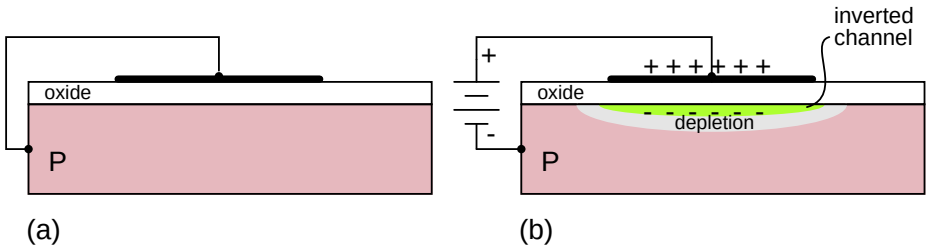


Figure 2.45: *N-channel MOS capacitor: (a) no charge, (b) charged.*

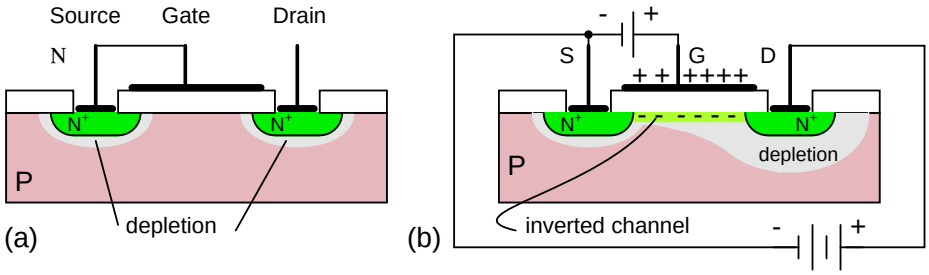


Figure 2.46: *N-channel MOSFET (enhancement type): (a) 0 V gate bias, (b) positive gate bias.*

N-region from source to drain. Thus, the MOSFET, like the FET is a unipolar device. One type of charge carrier is responsible for conduction. This example is an N-channel MOSFET. Conduction of a large current from source to drain is possible with a voltage applied between these connections. A practical circuit would have a load in series with the drain battery in Figure 2.46 (b).

The MOSFET described above in Figure 2.46 is known as an *enhancement mode* MOSFET. The non-conducting, off, channel is turned on by enhancing the channel below the gate by application of a bias. This is the most common kind of device. The other kind of MOSFET will not be described here. See the Insulated-gate field-effect transistor chapter for the *depletion mode* device.

The MOSFET, like the FET, is a voltage controlled device. A voltage input to the gate controls the flow of current from source to drain. The gate does not draw a continuous current. Though, the gate draws a surge of current to charge the gate capacitance.

The cross-section of an N-channel discrete MOSFET is shown in Figure 2.47 (a). Discrete devices are usually optimized for high power switching. The N^+ indicates that the source and drain are heavily N-type doped. This minimizes resistive losses in the high current path from source to drain. The N^- indicates light doping. The P-region under the gate, between source and drain can be inverted by application of a positive bias voltage. The doping profile is a cross-section, which may be laid out in a serpentine pattern on the silicon die. This greatly increases the area, and consequently, the current handling ability.

The MOSFET schematic symbol in Figure 2.47 (b) shows a “floating” gate, indicating no direct connection to the silicon substrate. The broken line from source to drain indicates that

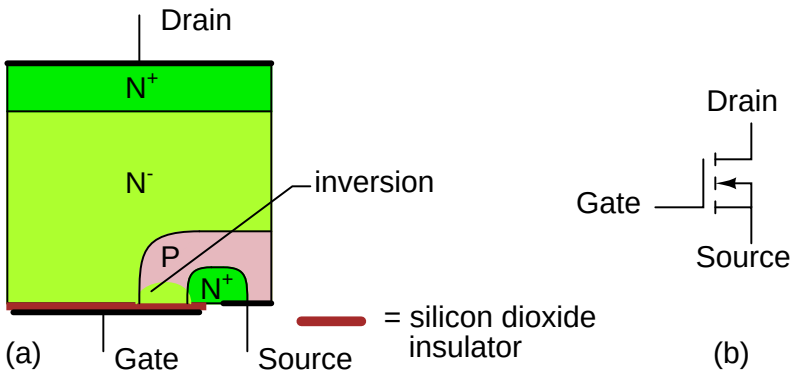


Figure 2.47: *N-channel MOSFET (enhancement type): (a) Cross-section, (b) schematic symbol.*

this device is off, not conducting, with zero bias on the gate. A normally "off" MOSFET is an enhancement mode device. The channel must be enhanced by application of a bias to the gate for conduction. The "pointing" end of the substrate arrow corresponds to P-type material, which points toward an N-type channel, the "non-pointing" end. This is the symbol for an N-channel MOSFET. The arrow points in the opposite direction for a P-channel device (not shown). MOSFET's are four terminal devices: source, gate, drain, and substrate. The substrate is connected to the source in discrete MOSFET's, making the packaged part a three terminal device. MOSFET's, that are part of an integrated circuit, have the substrate common to all devices, unless purposely isolated. This common connection may be bonded out of the die for connection to a ground or power supply bias voltage.

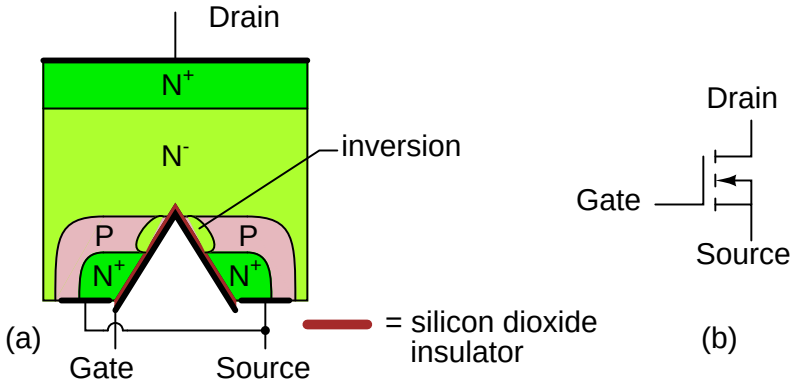


Figure 2.48: *N-channel "V-MOS" transistor: (a) Cross-section, (b) schematic symbol.*

The *V-MOS* device in (Figure 2.48) is an improved power MOSFET with the doping profile arranged for lower on-state source to drain resistance. VMOS takes its name from the V-shaped gate region, which increases the cross-sectional area of the source-drain path. This minimizes losses and allows switching of higher levels of power. UMOS, a variation using a

U-shaped grove, is more reproducible in manufacture.

- **REVIEW:**
- MOSFET's are unipoar conduction devices, conduction with one type of charge carrier, like a FET, but unlike a BJT.
- A MOSFET is a voltage controlled device like a FET. A gate voltage input controls the source to drain current.
- The MOSFET gate draws no continuous current, except leakage. However, a considerable initial surge of current is required to charge the gate capacitance.

2.11 Thyristors

Thyristors are a broad classification of bipolar-conducting semiconductor devices having four (or more) alternating N-P-N-P layers. Thyristors include: silicon controlled rectifier (SCR), TRIAC, gate turn off switch (GTO), silicon controlled switch (SCS), AC diode (DIAC), unijunction transistor (UJT), programmable unijunction transistor (PUT). Only the SCR is examined in this section; though the GTO is mentioned.

Shockley proposed the four layer diode thyristor in 1950. It was not realized until years later at General Electric. SCR's are now available to handle power levels spanning watts to megawatts. The smallest devices, packaged like small-signal transistors, switch 100's of milliamps at near 100 VAC. The largest packaged devices are 172 mm in diameter, switching 5600 Amps at 10,000 VAC. The highest power SCR's may consist of a whole semiconductor wafer several inches in diameter (100's of mm).

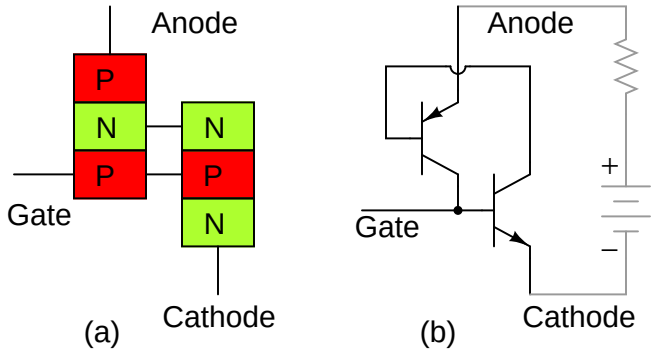


Figure 2.49: *Silicon controlled rectifier (SCR): (a) doping profile, (b) BJT equivalent circuit.*

The silicon controlled rectifier is a four layer diode with a gate connection as in Figure 2.49 (a). When turned on, it conducts like a diode, for one polarity of current. If not triggered on, it is nonconducting. Operation is explained in terms of the compound connected transistor equivalent in Figure 2.49 (b). A positive trigger signal is applied between the gate and cathode

terminals. This causes the NPN equivalent transistor to conduct. The collector of the conducting NPN transistor pulls low, moving the PNP base towards its collector voltage, which causes the PNP to conduct. The collector of the conducting PNP pulls high, moving the NPN base in the direction of its collector. This positive feedback (regeneration) reinforces the NPN's already conducting state. Moreover, the NPN will now conduct even in the absence of a gate signal. Once an SCR conducts, it continues for as long as a positive anode voltage is present. For the DC battery shown, this is forever. However, SCR's are most often used with an alternating current or pulsating DC supply. Conduction ceases with the expiration of the positive half of the sinewave at the anode. Moreover, most practical SCR circuits depend on the AC cycle going to zero to cutoff or *commutate* the SCR.

Figure 2.50 (a) shows the doping profile of an SCR. Note that the cathode, which corresponds to an equivalent emitter of an NPN transistor is heavily doped as N^+ indicates. The anode is also heavily doped (P^+). It is the equivalent emitter of a PNP transistor. The two middle layers, corresponding to base and collector regions of the equivalent transistors, are less heavily doped: N^- and P. This profile in high power SCR's may be spread across a whole semiconductor wafer of substantial diameter.

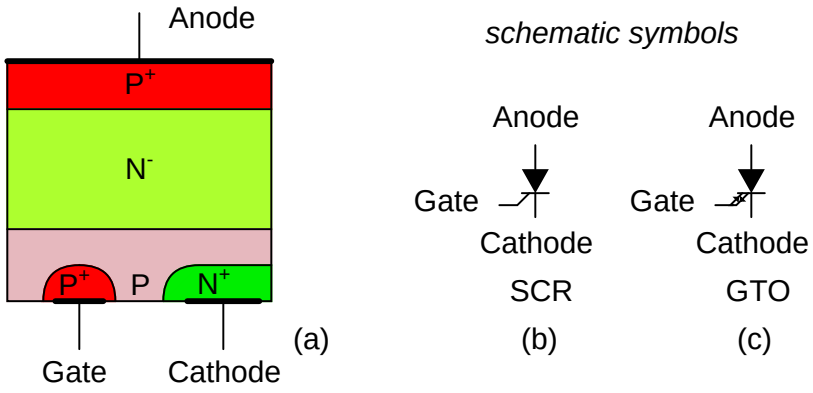


Figure 2.50: *Thyristors: (a) Cross-section, (b) silicon controlled rectifier (SCR) symbol, (c) gate turn-off thyristor (GTO) symbol.*

The schematic symbols for an SCR and GTO are shown in Figures 2.50 (b & c). The basic diode symbol indicates that cathode to anode conduction is unidirectional like a diode. The addition of a gate lead indicates control of diode conduction. The gate turn off switch (GTO) has bidirectional arrows about the gate lead, indicating that the conduction can be disabled by a negative pulse, as well as initiated by a positive pulse.

In addition to the ubiquitous silicon based SCR's, experimental silicon carbide devices have been produced. Silicon carbide (SiC) operates at higher temperatures, and is more conductive of heat than any metal, second to diamond. This should allow for either physically smaller or higher power capable devices.

- **REVIEW:**
- SCR's are the most prevalent member of the thyristor four layer diode family.

- A positive pulse applied to the gate of an SCR triggers it into conduction. Conduction continues even if the gate pulse is removed. Conduction only ceases when the anode to cathode voltage drops to zero.
- SCR's are most often used with an AC supply (or pulsating DC) because of the continuous conduction.
- A gate turn off switch (GTO) may be turned off by application of a negative pulse to the gate.
- SCR's switch megawatts of power, up to 5600 A and 10,000 V.

2.12 Semiconductor manufacturing techniques

The manufacture of only silicon based semiconductors is described in this section; most semiconductors are silicon. Silicon is particularly suitable for integrated circuits because it readily forms an oxide coating, useful in patterning integrated components like transistors.

Silicon is the second most common element in the Earth's crust in the form of silicon dioxide, SiO_2, otherwise known as silica sand. Silicon is freed from silicon dioxide by reduction with carbon in an electric arc furnace

$$SiO_2 + C = CO_2 + Si$$

Such metalurgical grade silicon is suitable for use in silicon steel transformer laminations, but not nearly pure enough for semiconductor applications. Conversion to the chloride $SiCl_4$ (or $SiHCl_3$) allows purification by fractional distillation. Reduction by ultrapure zinc or magnesium yields sponge silicon, requiring further purification. Or, thermal decomposition on a hot polycrystalline silicon rod heater by hydrogen yields ultra pure silicon.

$$Si + 3HCl = SiHCl_3 + H_2$$
$$SiHCl_3 + H_2 = Si + 3HCl_2$$

The polycrystalline silicon is melted in a fused silica crucible heated by an induction heated graphite suceptor. The graphite heater may alternately be directly driven by a low voltage at high current. In the *Czochralski process*, the silicon melt is solidified on to a pencil sized monocrystal silicon rod of the desired crystal lattice orientation. (Figure 2.51) The rod is rotated and pulled upward at a rate to encourage the diameter to expand to several inches. Once this diameter is attained, the *boule* is automatically pulled at a rate to maintain a constant diameter to a length of a few feet. Dopants may be added to the crucible melt to create, for example, a P-type semiconductor. The growing apparatus is enclosed within an inert atmosphere.

The finished boule is ground to a precise final diameter, and the ends trimmed. The boule is sliced into wafers by an inside diameter diamond saw. The wafers are ground flat and polished. The wafers could have an N-type *epitaxial* layer grown atop the wafer by thermal deposition for higher quality. Wafers at this stage of manufacture are delivered by the silicon wafer manufacturer to the semiconductor manufacturer.

The processing of semiconductors involves photo lithography, a process for making metal lithographic printing plates by acid etching. The electronics based version of this is the processing of copper printed circuit boards. This is reviewed in Figure 2.53 as an easy introduction to the photo lithography involved in semiconductor processing.

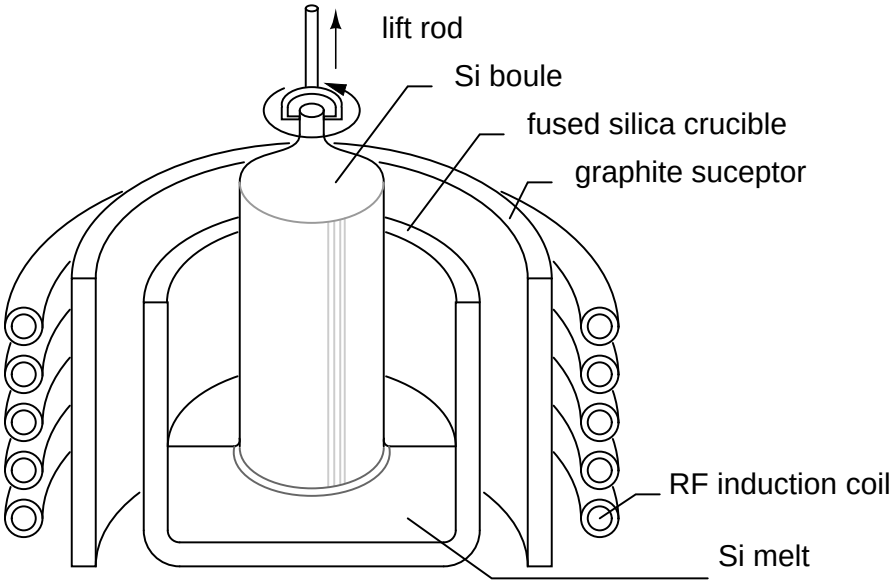


Figure 2.51: *Czochralski monocrystalline silicon growth.*

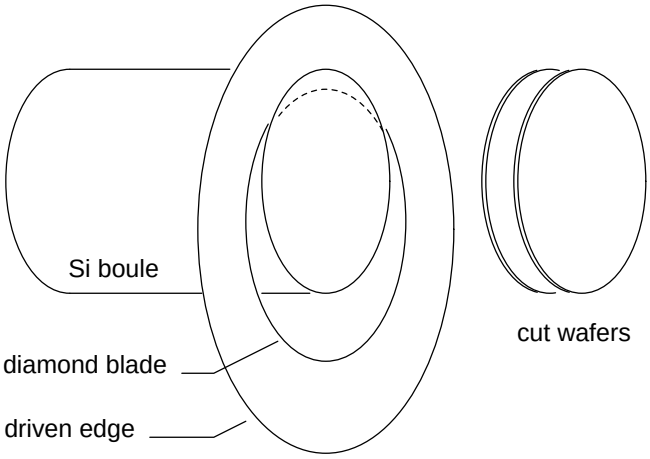


Figure 2.52: *Silicon boule is diamond sawed into wafers.*

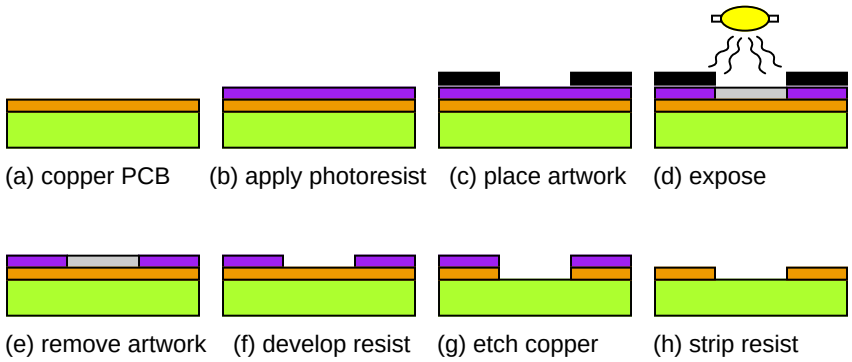


Figure 2.53: *Processing of copper printed circuit boards is similar to the photo lithographic steps of semiconductor processing.*

We start with a copper foil laminated to an epoxy fiberglass board in Figure 2.53 (a). We also need positive artwork with black lines corresponding to the copper wiring lines and pads that are to remain on the finished board. Positive artwork is required because positive acting resist is used. Though, negative resist is available for both circuit boards and semiconductor processing. At (b) the liquid positive photo resist is applied to the copper face of the printed circuit board (PCB). It is allowed to dry and may be baked in an oven. The artwork may be a plastic film positive reproduction of the original artwork scaled to the required size. The artwork is placed in contact with the circuit board under a glass plate at (c). The board is exposed to ultraviolet light (d) to form a *latent* image of softened photo resist. The artwork is removed (e) and the softened resist washed away by an alkaline solution (f). The rinsed and dried (baked) circuit board has a hardened resist image atop the copper lines and pads that are to remain after etching. The board is immersed in the etchant (g) to remove copper not protected by hardened resist. The etched board is rinsed and the resist removed by a solvent.

The major difference in the patterning of semiconductors is that a silicon dioxide layer atop the wafer takes the place of the resist during the high temperature processing steps. Though, the resist is required in low temperature wet processing to pattern the silicon dioxide.

An N-type doped silicon wafer in Figure 2.54 (a) is the starting material in the manufacture of semiconductor junctions. A silicon dioxide layer (b) is grown atop the wafer in the presence of oxygen or water vapor at high temperature (over 1000^o C in a diffusion furnace. A pool of resist is applied to the center of the cooled wafer, then spun in a vacuum chuck to evenly distribute the resist. The baked on resist (c) has a chrome on glass mask applied to the wafer at (d). This mask contains a pattern of windows which is exposed to ultraviolet light (e).

After the mask is removed in Figure 2.54 (f), the positive resist can be developed (g) in an alkaline solution, opening windows in the UV softened resist. The purpose of the resist is to protect the silicon dioxide from the hydrofluoric acid etch (h), leaving only open windows corresponding to the mask openings. The remaining resist (i) is stripped from the wafer before returning to the diffusion furnace. The wafer is exposed to a gaseous P-type dopant at high temperature in a diffusion furnace (j). The dopant only diffuses into the silicon through the openings in the silicon dioxide layer. Each P-diffusion through an opening produces a PN

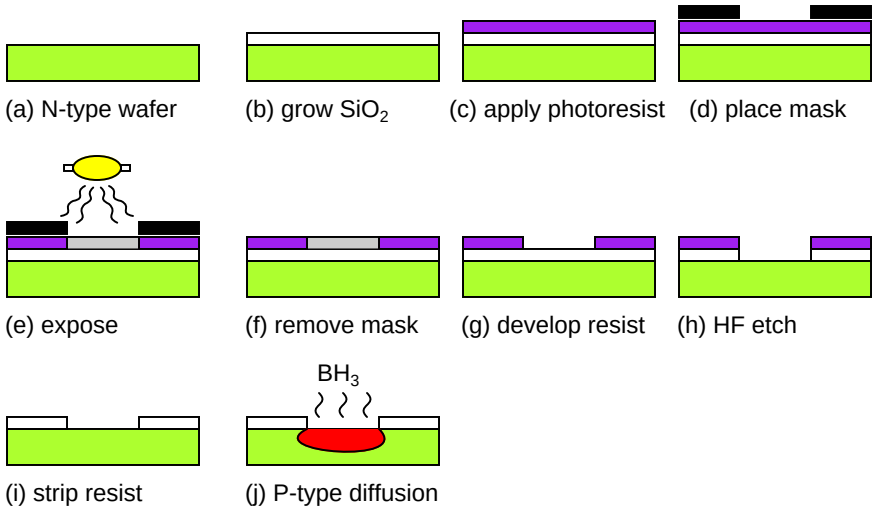


Figure 2.54: *Manufacture of a silicon diode junction.*

junction. If diodes were the desired product, the wafer would be diamond scribed and broken into individual diode chips. However, the whole wafer may be processed further into bipolar junction transistors.

To convert the diodes into transistors, a small N-type diffusion in the middle of the existing P-region is required. Repeating the previous steps with a mask having smaller openings accomplishes this. Though not shown in Figure 2.54 (j), an oxide layer was probably formed in that step during the P-diffusion. The oxide layer over the P-diffusion is shown in Figure 2.55 (k). Positive photo resist is applied and dried (l). The chrome on glass emitter mask is applied (m), and UV exposed (n). The mask is removed (o). The UV softened resist in the emitter opening is removed with an alkaline solution (p). The exposed silicon dioxide is etched away with hydrofluoric acid (HF) at (q)

After the unexposed resist is stripped from the wafer (r), it is placed in a diffusion furnace (Figure 2.55 (s) for high temperature processing. An N-type gaseous dopant, such phosphorus oxychloride (POCl) diffuses through the small emitter window in the oxide (s). This creates NPN layers corresponding to the emitter, base, and collector of a BJT. It is important that the N-type emitter not be driven all the way through the P-type base, shorting the emitter and collector. The base region between the emitter and collector also needs to be thin so that the transistor has a useful β. Otherwise, a thick base region could form a pair of diodes rather than a transistor. At (t) metalization is shown making contact with the transistor regions. This requires a repeat of the previous steps (not shown here) with a mask for contact openings through the oxide. Another repeat with another mask defines the metalization pattern atop the oxide and contacting the transistor regions through the openings.

The metalization could connect numerous transistors and other components into an *integrated circuit*. Though, only one transistor is shown. The finished wafer is diamond scribed and broken into individual dies for packaging. Fine gauge aluminum wire bonds the metalized contacts on the die to a *lead frame*, which brings the contacts out of the final package.

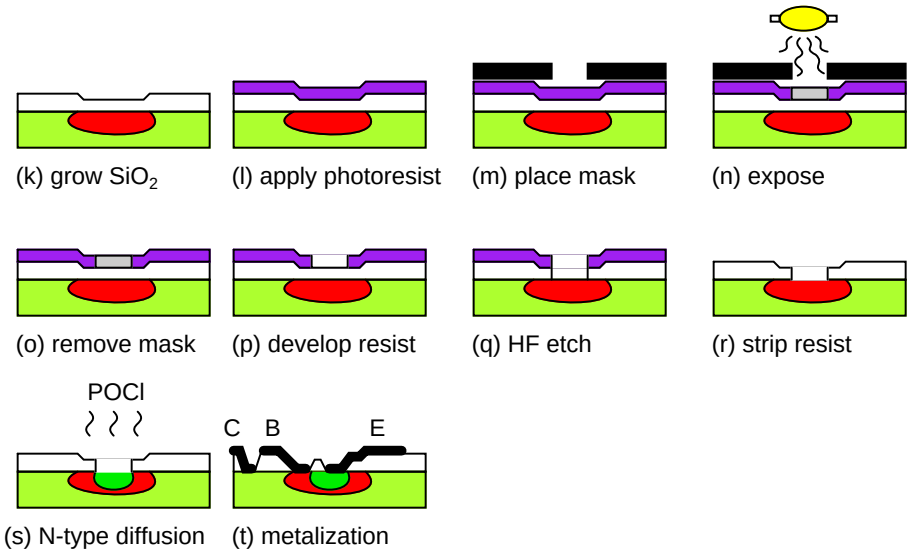


Figure 2.55: *Manufacture of a bipolar junction transistor, continuation of Manufacture of a silicon diode junction.*

- **REVIEW:**
- Most semiconductor are based on ultra pure silicon because it forms a glass oxide atop the wafer. This oxide can be patterned with photo lithography, making complex integrated circuits possible.
- Sausage shaped single crystals of silicon are grown by the Czochralski process, These are diamond sawed into wafers.
- The patterning of silicon wafers by photo lithography is similar to patterning copper printed circuit boards. Photo resist is applied to the wafer, which is exposed to UV light through a mask. The resist is developed, then the wafer is etched.
- hydrofluoric acid etching opens windows in the protective silicon dioxide atop the wafer.
- Exposure to gaseous dopants at high temperature produces semiconductor junctions as defined by the openings in the silicon dioxide layer.
- The photo lithography is repeated for more diffusions, contacts, and metalization.
- The metalization may interconnect multiple components into an integrated circuit.

2.13 Superconducting devices

Superconducting devices, though not widely used, have some unique characteristics not available in standard semiconductor devices. High sensitivity with respect to amplification of electrical signals, detection of magnetic fields, and detection of light are prized applications. High speed switching is also possible, though not applied to computers at this time. Conventional superconducting devices must be cooled to within a few degrees of 0 Kelvin (-273 ^{o} C). Though, work is proceeding at this time on *high temperature superconductor* based devices, usable at 90 K and below. This is significant because inexpensive liquid nitrogen may be used for cooling.

Superconductivity: Heike Onnes discovered *superconductivity* in mercury (Hg) in 1911, for which he won a Nobel prize. Most metals decrease electrical resistance with decreasing temperature. Though, most do not decrease to zero resistance as 0 Kelvin is approached. Mercury is unique in that its resistance abruptly drops to zero Ω at 4.2 K. Superconductors lose all resistance abruptly when cooled below their *critical temperature,* T_c A property of superconductivity is no power loss in conductors. Current may flow in a loop of superconducting wire for thousands of years. Super conductors include lead (Pb), aluminum, (Al), tin (Sn) and niobium (Nb).

Cooper pair: Lossless conduction in superconductors is not by ordinary electron flow. Electron flow in normal conductors encounters opposition as collisions with the rigid ionic metal crystal lattice. Decreasing vibrations of the crystal lattice with decreasing temperature accounts for decreasing resistance– up to a point. Lattice vibrations cease at absolute zero, but not the energy dissipating collisions of electrons with the lattice. Thus, normal conductors do not lose all resistance at absolute zero.

Electrons in superconductors form a pair of electrons called a *cooper pair*, as temperature drops below the critical temperature at which superconductivity begins. The cooper pair exists because it is at a lower energy level than unpaired electrons. The electrons are attracted to each other due to the exchange of *phonons*, very low energy particles related to vibrations. This cooper pair, quantum mechanical entity (particle or wave) is not subject to the normal laws of physics. This entity propagates through the lattice without encountering the metal ions comprising the fixed lattice. Thus, it dissipates no energy. The quantum mechanical nature of the cooper pair only allows it to exchange discrete amounts of energy, not continuously variable amounts. An absolute minimum quantum of energy is acceptable to the cooper pair. If the vibrational energy of the crystal lattice is less, (due to the low temperature), the cooper pair cannot accept it, cannot be scattered by the lattice. Thus, under the critical temperature, the cooper pairs flow unimpeded through the lattice.

Josephson junctions: Brian Josephson won a Nobel prize for his 1962 prediction of the *Josepheson junction*. A Josephson junction is a pair of superconductors bridged by a thin insulator, as in Figure 2.56 (a), through which electrons can tunnel. The first Josephson junctions were lead superconductors bridged by an insulator. These days a tri-layer of aluminum and niobium is preferred. Electrons can tunnel through the insulator even with zero voltage applied across the superconductors.

If a voltage is applied across the junction, the current decreases and oscillates at a high frequency proportional to voltage. The relationship between applied voltage and frequency is so precise that the standard volt is now defined in terms of Josephson junction oscillation frequency. The Josephson junction can also serve as a hyper-sensitive detector of low level magnetic fields. It is also very sensitive to electromagnetic radiation from microwaves to gamma

rays.

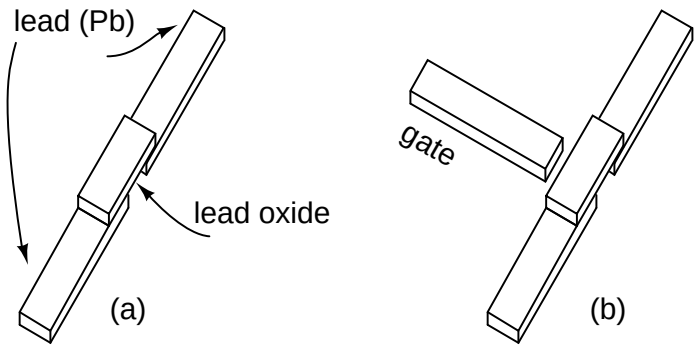


Figure 2.56: *(a) Josephson junction, (b) Josephson transistor.*

Josephson transistor: An electrode close to the oxide of the Josephson junction can influence the junction by capacitive coupling. Such an assembly in Figure 2.56 (b) is a Josephson transistor. A major feature of the Josephson transistor is low power dissipation applicable to high density circuitry, for example, computers. This transistor is generally part of a more complex superconducting device like a SQUID or RSFQ.

SQUID: A *Superconducting quantum interference device* or *SQUID* is an assembly of Josephson junctions within a superconducting ring. Only the DC SQUID is considered in this discussion. This device is highly sensitive to low level magnetic fields.

A constant current bias is forced across the ring in parallel with both Josephson junctions in Figure 2.57. The current divides equally between the two junctions in the absence of an applied magnetic field and no voltage is developed across across the ring. [4] While any value of Magnetic flux (Φ) may be applied to the SQUID, only a quantized value (a multiple of the flux quanta) can flow through the opening in the superconducting ring.[2] If the applied flux is not an exact multiple of the flux quanta, the excess flux is cancelled by a circulating current around the ring which produces a fractional flux quanta. The circulating current will flow in that direction which cancels any excess flux above a multiple of the flux quanta. It may either add to, or subtract from the applied flux, up to $\pm(1/2)$ a flux quanta. If the circulating current flows clockwise, the current adds to the top Josepheson junction and subtracts from the lower one. Changing applied flux linearly causes the circulating current to vary as a sinusoid.[3] This can be measured as a voltage across the SQUID. As the applied magnetic field is increased, a voltage pulse may be counted for each increase by a flux quanta.[19]

A SQUID is said to be sensitive to 10^{-14} Tesla, It can detect the magnetic field of neural currents in the brain at 10^{-13} Tesla. Compare this with the 30 x 10^{-6} Tesla strength of the Earth's magnetic field.

Rapid single flux quantum (RSFQ): Rather than mimic silicon semiconductor circuits, RSFQ circuits rely upon new concepts: magnetic flux quantization within a superconductor, and movement of the flux quanta produces a picosecond quantized voltage pulse. Magnetic flux can only exist within a section of superconductor quantized in discrete multiples. The lowest flux quanta allowed is employed. The pulses are switched by Josephson junctions instead of conventional transistors. The superconductors are based on a triple layer of aluminum and

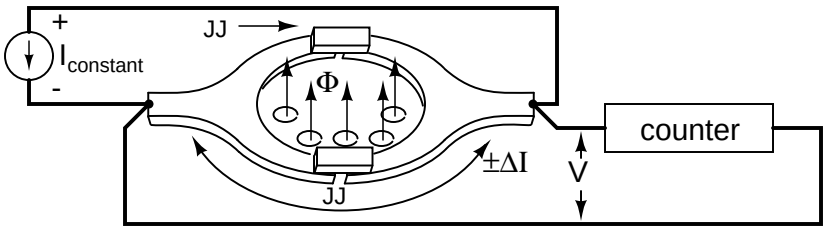


Figure 2.57: *Superconduction quantum interference device (SQUID): Josephson junction pair within a superconducting ring. A change in flux produces a voltage variation across the JJ pair.*

niobium with a critical temperature of 9.5 K, cooled to 5 K.

RSQF's operate at over 100 gHz with very little power dissipation. Manufacture is simple with existing photolithographic techniques. Though, operation requires refrigeration down to 5 K . Real world commercial applications include analog-to-digital and digital to analog converters, toggle flip-flops, shift registers, memory, adders, and multipliers.[5]

High temperature superconductors: *High temperature superconductors* are compounds exhibiting superconductivity above the liquid nitrogen boiling point of 77 K. This is significant because liquid nitrogen is readily available and inexpensive. Most conventional superconductors are metals; widely used high temperature superconductors are *cuprates*, mixed oxides of copper (Cu), for example $YBa_2Cu_3O_{7-x}$, critical temperature, T_c = 90 K . A list of others is available.[23] Most of the devices described in this section are being developed in high temperature superconductor versions for less critical applications. Though they do not have the performance of the conventional metal superconductor devices, the liquid nitrogen cooling is more available.

- **REVIEW:**
- Most metals decrease resistance as they approach absolute 0; though, the resistance does not drop to 0. Superconductors experience a rapid drop to zero resistance at their critical temperature on being cooled. Typically T_c is within 10 K of absolute zero.
- A Cooper pair, electron pair, a quantum mechanical entity, moves unimpeded through the metal crystal lattice.
- Electrons are able to tunnel through a Josephson junction, an insulating gap across a pair of superconductors.
- The addition of a third electrode, or gate, near the junction constitutes a Josephson transistor.
- A SQUID, Superconduction quantum interference device, is a highly sensitive detector of magnetic fields. It counts quantum units of a magnetic field within a superconducting ring.
- RSFQ, Rapid single flux quantum is a high speed switching device based on switching the magnetic quanta existing withing a superconducting loop.

- High temperature superconductors, T_c above liquid nitrogen boiling point, may also be used to build the superconducting devices in this section.

2.14 Quantum devices

Most integrated circuits are digital, based on MOS (CMOS) transistors. Every couple of years since the late 1960's a geometry shrink has taken place, increasing the circuit density– more circuitry at lower cost in the same space. As of this writing (2006), the MOS transistor gate length is 65-nm for leading edge production, with 45-nm anticipated within a year. At 65-nm leakage currents were becoming evident. At 45-nm, heroic innovations were required to minimize this leakage. The end of shrinkage in MOS transistors is expected at 20- to 30-nm. Though some think that 1- to 2-nm is the limit. Photolithography, or other lithographic techniques, will continue to improve, providing ever smaller geometry. However, conventional MOS transistors are not expected to be usable at these smaller geometries below 20- to 30-nm.

Improved photolithography will have to be applied to other than the conventional transistors, dimensions (under 20- to 30-nm). The objectional MOS leakage currents are due to quantum mechanical effects–electron tunneling through gate oxide, and the narrow channel. In summary, quantum mechanical effects are a hindrance to ever smaller conventional MOS transistors. The path to ever smaller geometry devices involves unique active devices which make practical use of quantum mechanical principles. As physical geometry becomes very small, electrons may be treated as the quantum mechanical equivalent: a wave. Devices making use of quantum mechanical principles include: resonant tunneling diodes, quantum tunneling transistors, metal insulator metal diodes, and quantum dot transistors.

Quantum tunneling: is the passing of electrons through an insulating barrier which is thin compared to the de Broglie (page 31) electron wavelength. If the "electron wave" is large compared to the barrier, there is a possibility that the wave appears on both sides of the barrier.

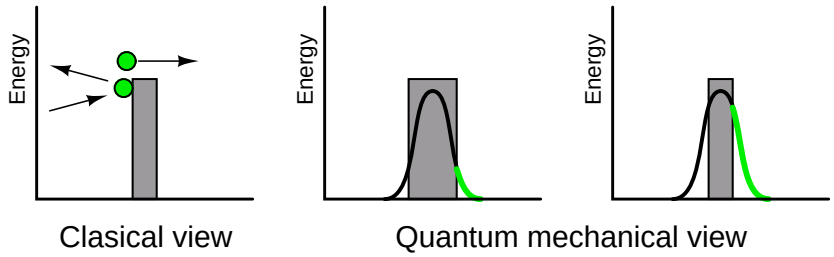


Figure 2.58: *Classical view of an electron surmounting a barrier, or not. Quantum mechanical view allows an electron to tunnel through a barrier. The probability (green) is related to the barrier thickness. After Figure 1 [22]*

In classical physics, an electron must have sufficient energy to surmount a barrier. Otherwise, it recoils from the barrier. (Figure 2.58) Quantum mechanics allows for a probability of the electron being on the other side of the barrier. If treated as a wave, the electron may look quite large compared to the thickness of the barrier. Even when treated as a wave, there is only a small probability that it will be found on the other side of a thick barrier. See green

portion of curve, Figure 2.58. Thinning the barrier increases the probability that the electron is found on the other side of the barrier. [22]

Tunnel diode: The unqualified term *tunnel diode* refers to the *esaki tunnel diode*, an early quantum device. A reverse biased diode forms a depletion region, an insulating region, between the conductive anode and cathode. This depletion region is only thin as compared to the electron wavelength when heavily doped– 1000 times the doping of a rectifier diode. With proper biasing, quantum tunneling is possible. See (page 144)for details.

RTD, resonant tunneling diode: This is a quantum device not to be confused with the Esaki tunnel diode, (page 144) , a conventional heavily doped bipolar semiconductor. Electrons *tunnel* through two barriers separated by a well in flowing source to drain in a *resonant tunneling diode*. Tunneling is also known as quantum mechanical tunneling. The flow of electrons is controlled by diode bias. This matches the energy levels of the electrons in the source to the quantized level in the well so that electrons can tunnel through the barriers. The energy level in the well is quantized because the well is small. When the energy levels are equal, a *resonance* occurs, allowing electron flow through the barriers as shown in Figure 2.59 (b). No bias or too much bias, in Figures 2.59 (a) and (c) respectively, yields an energy mismatch between the source and the well, and no conduction.

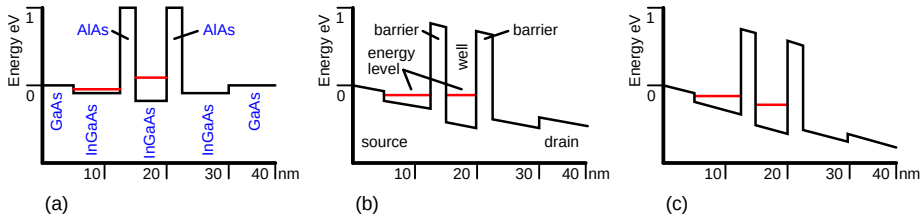


Figure 2.59: *Resonant tunneling diode (RTD): (a) No bias, source and well energy levels not matched, no conduction. (b) Small bias causes matched energy levels (resonance); conduction results. (c) Further bias mismatches energy levels, decreasing conduction.*

As bias is increased from zero across the RTD, the current increases and then decreases, corresponding to off, on, and off states. This makes simplification of conventional transistor circuits possible by substituting a pair of RTD's for two transistors. For example, two back-to-back RTD's and a transistor form a memory cell, using fewer components, less area and power compared with a conventional circuit. The potential application of RTD's is to reduce the component count, area, and power dissipation of conventional transistor circuits by replacing some, though not all, transistors. [11] RTD's have been shown to oscillate up to 712 gHz. [8]

Double-layer tunneling transistor: The *Deltt*, otherwise known as the *Double-layer tunneling transistor* is constructed of a pair of conductive wells separated by an insulator or high band gap semiconductor. (Figure 2.60) The wells are so thin that electrons are confined to two dimensions. These are known as *quantum wells*. A pair of these quantum wells are insulated by a thin GaAlAs, high band gap (does not easily conduct) layer. Electrons can *tunnel* through the insulating layer if the electrons in the two quantum wells have the same momentum and energy. The wells are so thin that the electron may be treated as a wave– the quantum mechanical duality of particles and waves. The top and optional bottom control gates may be adjusted to equalize the energy levels (resonance) of the electrons to allow conduction from

source to drain. Figure 2.60, barrier diagram red bars show unequal energy levels in the wells, an "off-state" condition. Proper biasing of the gates equalizes the energy levels of electrons in the wells, the "on-state" condition. The bars would be at the same level in the energy level diagram.

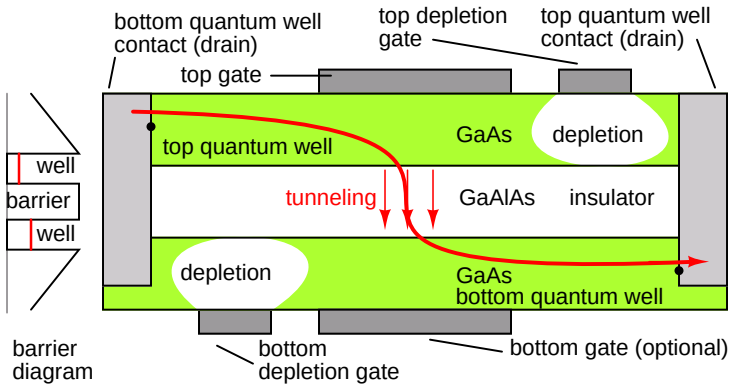


Figure 2.60: *Double-layer tunneling transistor (Deltt) is composed of two electron containing wells separated by a nonconducting barrier. The gate voltages may be adjusted so that the energy and momentum of the electrons in the wells are equal which permits electrons to tunnel through the nonconductive barrier. (The energy levels are shown as unequal in the barrier diagram.)*

If gate bias is increased beyond that required for tunneling, the energy levels in the quantum wells no longer match, tunneling is inhibited, source to drain current decreases. To summarize, increasing gate bias from zero results in on, off, on conditions. This allows a pair of Deltt's to be stacked in the manner of a CMOS complementary pair; though, different p- and n-type transistors are not required. Power supply voltage is about 100 mV. Experimental Deltt's have been produced which operate near 4.2 K, 77 K, and 0^o C. Room temperature versions are expected.[11] [13] [20]

MIIM diode: The *metal-insulator-insulator-metal* (MIIM) diode is a quantum tunneling device, not based on semiconductors. See "MIIM diode section" Figure 2.61. The insulator layers must be thin compared to the de Broglie (page 31) electron wavelength, for quantum tunneling to be possible. For diode action, there must be a prefered tunneling direction, resulting in a sharp bend in the diode forward characteristic curve. The MIIM diode has a sharper forward curve than the metal insulator metal (MIM) diode, not considered here.

The energy levels of M1 and M2 are equal in "no bias" Figure 2.61. However, (thermal) electrons cannot flow due to the high I1 and I2 barriers. Electrons in metal M2 have a higher energy level in "reverse bias" Figure 2.61, but still cannot overcome the insulator barrier. As "forward bias" Figure 2.61 is increased, a *quantum well*, an area where electrons may exist, is formed between the insulators. Electrons may pass through insulator I1 if M1 is bised at the same energy level as the quantum well. A simple explanation is that the distance through the insulators is shorter. A longer explanation is that as bias increases, the probability of the electron wave overlapping from M1 to the quantum well increases. For a more detailed

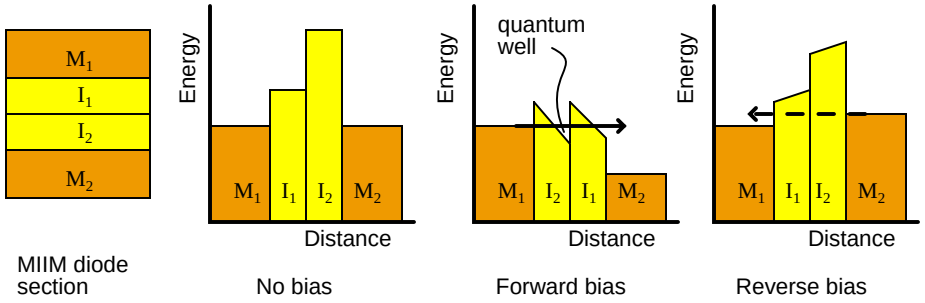


Figure 2.61: *Metal insulator insulator metal (MIIM) diode: Cross section of diode. Energy levels for no bias, forward bias, and reverse bias. After Figure 1 [21].*

explanation see Phiar Corp. [21]

MIIM devices operate at higher frequencies (3.7 THz) than microwave transistors. [16] The addition of a third electrode to a MIIM diode produces a transistor.

Quantum dot transistor: An isolated conductor may take on a charge, measured in coulombs for large objects. For a nano-scale isolated conductor known as a *quantum dot*, the charge is measured in electrons. A quantum dot of 1- to 3-nm may take on an incremental charge of a single electron. This is the basis of the *quantum dot transistor*, also known as a *single electron transistor*.

A quantum dot placed atop a thin insulator over an electron rich source is known as a *single electron box*. (Figure 2.62 (a)) The energy required to transfer an electron is related to the size of the dot and the number of electrons already on the dot.

A gate electrode above the quantum dot can adjust the energy level of the dot so that quantum mechanical tunneling of an electron (as a wave) from the source through the insulator is possible. (Figure 2.62 (b)) Thus, a single electron may tunnel to the dot.

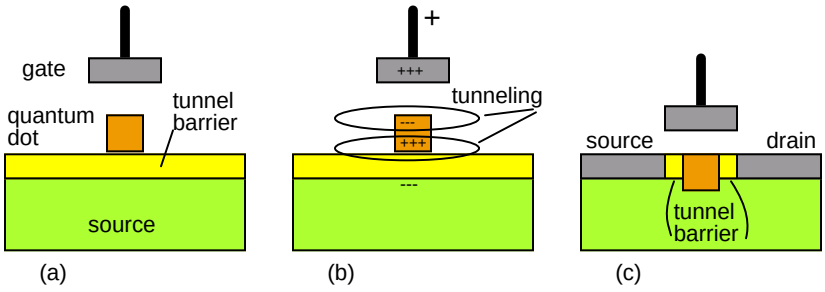


Figure 2.62: *(a) Single electron box, an isolated quantum dot separated from an electron source by an insulator. (b) Positive charge on the gate polarizes quantum dot, tunneling an electron from the source to the dot. (c) Quantum transistor: channel is replaced by quantum dot surrounded by tunneling barrier.*

If the quantum dot is surrounded by a tunnel barrier and embedded between the source

and drain of a conventional FET, as in Figure 2.62 (c) , the charge on the dot can modulate the flow of electrons from source to drain. As gate voltage increases, the source to drain current increases, up to a point. A further increase in gate voltage decreases drain current. This is similar to the behavior of the RTD and Deltt resonant devices. Only one kind of transistor is required to build a complementary logic gate.[11]

Single electron transistor: If a pair of conductors, superconductors, or semiconductors are separated by a pair of tunnel barriers (insulator), surrounding a tiny conductive island, like a quantum dot, the flow of a single charge (a Cooper pair for superconductors) may be controlled by a gate. This is a *single electron transistor* similar to Figure 2.62 (c). Increasing the positive charge on the gate, allows an electron to tunnel to the island. If it is sufficiently small, the low capacitance will cause the dot potential to rise substantially due to the single electron. No more electrons can tunnel to the island due the electron charge. This is known at the *coulomb blockade*. The electron which tunneled to the island, can tunnel to the drain.

Single electron transistors operate at near absolute zero. The exception is the graphene single electron transistor, having a graphene island. They are all experimental devices.

Graphene transistor: Graphite, an allotrope of carbon, does not have the rigid interlocking crystalline structure of diamond. None the less, it has a crystalline structure– one atom thick, a so called two-dimensional structure. A graphite is a three-dimensional crystal. However, it cleaves into thin sheets. Experimenters, taking this to the extreme, produce micron sized specks as thin as a single atom known as *graphene*. (Figure 2.63 (a)) These membranes have unique electronic properties. Highly conductive, conduction is by either electrons or holes, without doping of any kind. [12]

Graphene sheets may be cut into transistor structures by lithographic techniques. The transistors bear some resemblance to a MOSFET. A gate capacitively coupled to a graphene channel controls conduction.

As silicon transistors scale to smaller sizes, leakage increases along with power dissipation. And they get smaller every couple of years. Graphene transistors dissipate little power. And, they switch at high speed. Graphene might be a replacement for silicon someday.

Graphene can be fashioned into devices as small as sixty atoms wide. Graphene quantum dots within a transistor this small serve as *single electron transistors*. Previous single electron transistors fashioned from either superconductors or conventional semiconductors operate near absolute zero. Graphene single electron transistors uniquely function at room temperature.[24]

Graphene transistors are laboratory curiosities at this time. If they are to go into production two decades from now, graphene wafers must be produced. The first step, production of graphene by chemical vapor deposition (CVD) has been accomplished on an experimental scale. Though, no wafers are available to date.

Carbon nanotube transistor: If a 2-D sheet of graphene is rolled, the resulting 1-D structure is known as a *carbon nanotube*. (Figure 2.63 (b)) The reason to treat it as 1-dimensional is that it is highly conductive. Electrons traverse the carbon nanotube without being scattered by a crystal lattice. Resistance in normal metals is caused by scattering of electrons by the metallic crystalline lattice. If electrons avoid this scattering, conduction is said to be by *ballistic transport*. Both metallic (acting) and semiconducting carbon nanotubes have been produced. [6]

Field effect transistors may be fashioned from a carbon nanotubes by depositing source and drain contacts on the ends, and capacitively coupling a gate to the nanotube between the

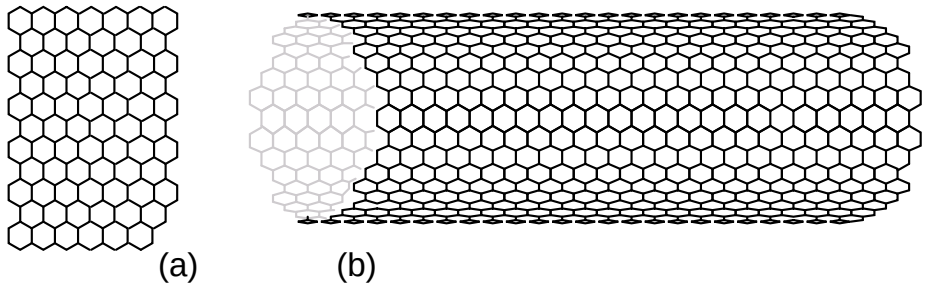


Figure 2.63: *(a) Graphene: A single sheet of the graphite allotrope of carbon. The atoms are arranged in a hexagonal pattern with a carbon at each intersection. (b) Carbon nanotube: A rolled-up sheet of graphene.*

contacts. Both p- and n-type transistors have been fabricated. Why the interest in carbon nanotube transistors? Nanotube semiconductors are Smaller, faster, lower power compared with silicon transistors. [7]

Spintronics: Conventional semiconductors control the flow of electron charge, current. Digital states are represented by "on" or "off" flow of current. As semiconductors become more dense with the move to smaller geometry, the power that must be dissipated as heat increases to the point that it is difficult to remove. Electrons have properties other than charge such as spin. A tentative explanation of *electron spin* is the rotation of distributed electron charge about the spin axis, analogous to diurnal rotation of the Earth. The loops of current created by charge movement, form a magnetic field. However, the electron is more like a point charge than a distributed charge, Thus, the rotating distributed charge analogy is not a correct explanation of spin. Electron spin may have one of two states: up or down which may represent digital states. More precisely the spin (m_s) quantum number may be $\pm 1/2$ the angular momentum (l) quantum number. [1]

Controlling electron spin instead of charge flow considerably reduces power dissipation and increases switching speed. *Spintronics*, an acronym for *SPIN TRansport electrONICS,* is not widely applied because of the difficulty of generating, controlling, and sensing electron spin. However, high density, non-volatile magnetic spin memory is in production using modified semiconductor processes. This is related to the *spin valve* magnetic read head used in computer harddisk drives, not mentioned further here.

A simple *magnetic tunnel junction (MTJ)* is shown in Figure 2.64 (a), consisting of a pair of *ferromagnetic*, strong magnetic properties like iron (Fe), layers separated by a thin insulator. Electrons can tunnel through a sufficiently thin insulator due to the quantum mechanical properties of electrons– the wave nature of electrons. The current flow through the MTJ is a function of the magnetization, spin polarity, of the ferromagnetic layers. The resistance of the MTJ is low if the magnetic spin of the top layer is in the same direction (polarity) as the bottom layer. If the magnetic spins of the two layers oppose, the resistance is higher. [9]

The change in resistance can be enhanced by the addition of an *antiferromagnet*, material having spins aligned but opposing, below the bottom layer in Figure 2.64 (b). This bias magnet *pins* the lower ferromagnetic layer spin to a single unchanging polarity. The top layer magnetization (spin) may be flipped to represent data by the application of an external magnetic field

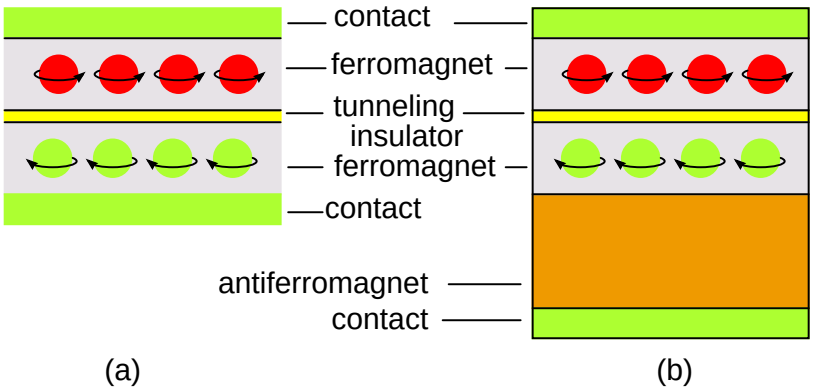


Figure 2.64: *(a) Magnetic tunnel junction (MTJ): Pair of ferromagnetic layers separated by a thin insulator. The resistance varies with the magnetization polarity of the top layer (b) Antiferromagnetic bias magnet and pinned bottom ferromagnetic layer increases resistance sensitivity to changes in polarity of the top ferromagnetic layer. Adapted from [9] Figure 3.*

not shown in the figure. The pinned layer is not affected by external magnetic fields. Again, the MTJ resistance is lowest when the spin of the top ferromagnetic layer is the same sense as the bottom pinned ferromagnetic layer. [9]

The MTJ may be improved further by splitting the pinned ferromagnetic layer into two layers separated by a buffer layer in Figure 2.65 (a). This isolates the top layer. The bottom ferromagnetic layer is pinned by the antiferromagnet as in the previous figure. The ferromagnetic layer atop the buffer is attracted by the bottom ferromagnetic layer. Opposites attract. Thus, the spin polarity of the additional layer is opposite of that in the bottom layer due to attraction. The bottom and middle ferromagnetic layers remain fixed. The top ferromagnetic layer may be set to either spin polarity by high currents in proximate conductors (not shown). This is how data are stored. Data are read out by the difference in current flow through the tunnel junction. Resistance is lowest if the layers on both sides of the insulting layer are of the same spin. [9]

An array of magnetic tunnel junctions may be embedded in a silicon wafer with conductors connecting the top and bottom terminals for reading data bits from the MTJ's with conventional CMOS circuitry. One such MTJ is shown in Figure 2.65 (b) with the read conductors. Not shown, another crossed array of conductors carrying heavy write currents switch the magnetic spin of the top ferromagnetic layer to store data. A current is applied to one of many "X" conductors and a "Y" conductor. One MTJ in the array is magnetized under the conductors' cross-over. Data are read out by sensing the MTJ current with conventional silicon semiconductor circuitry. [10]

The main reason for interest in magnetic tunnel junction memory is that it is *nonvolatile*. It does not lose data when powered "off". Other types of nonvolatile memory are capable of only limited storage cycles. MTJ memory is also higher speed than most semiconductor memory types. It is now (2006) a commercial product. [18]

Not a commercial product, or even a laboratory device, is the theoretical spin transistor

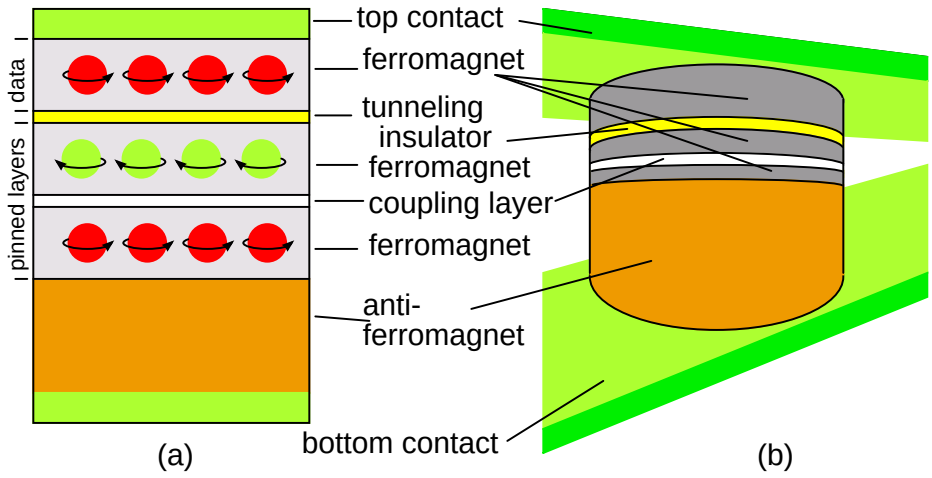


Figure 2.65: *(a)Splitting the pinned ferromagnetic layer of (b) by a buffer layer improves stability and isolates the top ferromagnetic unpinned layer. Data are stored in the top ferromagnetic layer based on spin polarity (b) MTJ cell embedded in read lines of a semiconductor die– one of many MTJ's. Adapted from [10]*

which might one day make spin logic gates possible. The spin transistor is a derivative of the theoretical spin diode.

It has been known for some time that electrons flowing through a cobalt-iron ferromagnet become spin polarized. The ferromagnet acts as a filter passing electrons of one spin preferentially. These electrons may flow into an adjacent nonmagnetic conductor (or semiconductor) retaining the spin polarization for a short time, nano-seconds. Though, spin polarized electrons may propagate a considerable distance compared with semiconductor dimensions. The spin polarized electrons may be detected by a nickel-iron ferromagnetic layer adjacent to the semiconductor. [1] [15]

It has also been shown that electron spin polarization occurs when circularly polarized light illuminates some semiconductor materials. Thus, it should be possible to inject spin polarized electrons into a semiconductor diode or transistor. The interest in spin based transistors and gates is because of the non-dissipative nature of spin propagation, compared with dissipative charge flow. As conventional semiconductors are scaled down in size, power dissipation increases. At some point the scaling down will no longer be practical. Researchers are looking for a replacement for the conventional charge flow based transistor. That device may be based on spintronics. [14]

- **REVIEW:**
- As MOS gate oxide thins with each generation of smaller transistors, excessive gate leakage causes unacceptable power dissipation and heating. The limit of scaling down conventional semiconductor geometry is within sight.
- Resonant tunneling diode (RTD): Quantum mechanical effects, which degrade conven-

tional semiconductors, are employed in the RTD. The flow of electrons through a sufficiently thin insulator, is by the wave nature of the electron– particle wave duality. The RTD functions as an amplifier.

- Double layer tunneling transistor (Deltt): The Deltt is a transistor version of the RTD. Gate bias controls the ability of electrons to tunnel through a thin insulator from one quantum well to another (source to drain).

- Quantum dot transistor: A quantum dot, capable of holding a charge, is surrounded by a thin tunnel barrier replacing the gate of a conventional FET. The charge on the quantum dot controls source to drain current flow.

- Spintronics: Electrons have two basic properties: charge and spin. Conventional electronic devices control the flow of charge, dissipating energy. Spintronic devices manipulate electron spin, a propagative, non-dissipative process.

2.15 Semiconductor devices in SPICE

The SPICE (**s**imulation **p**rogram, **i**ntegrated **c**ircuit **e**mphesis) electronic simulation program provides circuit elements and models for semiconductors. The SPICE element names begin with d, q, j, or m correspond to diode, BJT, JFET and MOSFET elements, respectively. These elements are accompanied by corresponding "models" These models have extensive lists of parameters describing the device. Though, we do not list them here. In this section we provide a very brief listing of simple spice models for semiconductors, just enough to get started. For more details on models and an extensive list of model parameters see Kuphaldt. [17] This reference also gives instructions on using SPICE.

Diode: The diode statement begins with a diode element name which must begin with "d" plus optional characters. Some example diode element names include: d1, d2, dtest, da, db, d101, etc. Two node numbers specify the connection of the anode and cathode, respectively, to other components. The node numbers are followed by a model name, referring to a ".model" statement.

The model statement line begins with ".model", followed by the model name matching one or more diode statements. Next is a "d" indicating that a diode is being modeled. The remainder of the model statement is a list of optional diode parameters of the form ParameterName=ParameterValue. None are shown in the example below. For a list, see reference, "diodes". [17]

```
General form:  d[name] [anode] [cathode] [model]
               .model  [modelname] d ( [parmtr1=x] [parmtr2=y] . . .)

Example:       d1 1 2 mod1
                .model mod1 d
```

Models for specific diode part numbers are often furnished by the semiconductor diode manufacturer. These models include parameters. Otherwise, the parameters default to so called "default values", as in the example.

BJT, bipolar junction transistor: The BJT element statement begins with an element name which must begin with "q" with associated circuit symbol designator characters, example: q1, q2, qa, qgood. The BJT node numbers (connections) identify the wiring of the collector, base, emitter respectively. A model name following the node numbers is associated with a model statement.

```
General form:  q[name] [collector] [base] [emitter] [model]
               .model [modelname] [npn or pnp] ([parmtr1=x] . . .)

Example:       q1 2 3 0 mod1
              .model mod1 pnp
Example:       q2 7 8 9 q2n090
              .model q2n090  npn ( bf=75 )
```

The model statement begins with ".model", followed by the model name, followed by one of "npn" or "pnp". The optional list of parameters follows, and may continue for a few lines beginning with line continuation symbol "+", plus. Shown above is the forward β parameter set to 75 for the hypothetical q2n090 model. Detailed transistor models are often available from semiconductor manufacturers.

FET, field effect transistor The field effect transistor element statement begins with an element name beginning with "j" for JFET associated with some unique characters, example: j101, j2b, jalpha, etc. The node numbers follow for the drain, gate and source terminals, respectively. The node numbers define connectivity to other circuit components. Finally, a model name indicates the JFET model to use.

```
General form:  j[name] [drain] [gate] [source] [model]
               .model [modelname] [njf or pjf] ( [parmtr1=x] . . .)

Example:       j1 2 3 0 mod1
               .model mod1 pjf
               j3 4 5 0 mod2
               .model mod2 njf ( vto=-4.0 )
```

The ".model" in the JFET model statement is followed by the model name to identify this model to the JFET element statement(s) using it. Following the model name is either pjf or njf for p-channel or n-channel JFET's respectively. A long list of JFET parameters may follow. We only show how to set V_p, pinch off voltage, to -4.0 V for an n-channel JFET model. Otherwise, this vto parameter defaults to -2.5 V or 2.5V for n-channel or p-channel devices, respectively.

MOSFET, metal oxide field effect transistor The MOSFET element name must begin with "m", and is the first word in the element statement. Following are the four node numbers for the drain, gate, source, and substrate, respectively. Next is the model name. Note that the source and substrate are both connected to the same node "0" in the example. Discrete MOSFET's are packaged as three terminal devices, the source and substrate are the same physical terminal. Integrated MOSFET's are four terminal devices; the substrate is a fourth terminal. Integrated MOSFET's may have numerous devices sharing the same substrate, separate from the sources. Though, the sources might still be connected to the common substrate.

```
General form:  m[name] [drain] [gate] [source] [substrate] [model]
               .model [modelname] [nmos or pmos] ( [parmtr1=x] . . . )
```

```
Example:        m1 2 3 0 0 mod1
                m5 5 6 0 0 mod4
                .model mod1 pmos
                .model mod4 nmos ( vto=1 )
```

The MOSFET model statement begins with ".model" followed by the model name followed by either "pmos" or "nmos". Optional MOSFET model parameters follow. The list of possible parameters is long. See Volume 5, "MOSFET" for details. [17] MOSFET manufacturers provide detailed models. Otherwise, defaults are in effect.

The bare minimum semiconductor SPICE information is provided in this section. The models shown here allow simulation of basic circuits. In particular, these models do not account for high speed or high frequency operation. Simulations are shown in the Volume 5 Chapter 7, "Using SPICE ...".

- **REVIEW:**
- Semiconductors may be computer simulated with SPICE.
- SPICE provides element statements and models for the diode, BJT, JFET, and MOSFET.

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.

Bibliography

[1] David D. Awschalom, Michael E. Flatte, Nitin Samarth, "Spintronics", Scientific American, June 2002 at http://www.sciam.com

[2] John Bland, "The Fluxoid" in "A Mossbauer Spectroscopy and Magnetometry Study of Magnetic Multilayers and Oxides", Oliver Lodge Laboratory, Department of Physics, University of Liverpool, 2002, at http://www.cmp.liv.ac.uk/frink/thesis/thesis/node45.html

[3] John Bland, "Superconducting Quantum Interference Device (SQUID)" in "A Mossbauer Spectroscopy and Magnetometry Study of Magnetic Multilayers and Oxides", Oliver Lodge Laboratory, Department of Physics, University of Liverpool, 2002, at http://www.cmp.liv.ac.uk/frink/thesis/thesis/node48.html

[4] John Bland, "SQUID Magnetometer" in "A Mossbauer Spectroscopy and Magnetometry Study of Magnetic Multilayers and Oxides", Oliver Lodge

Laboratory, Department of Physics, University of Liverpool, 2002, at http://www.cmp.liv.ac.uk/frink/thesis/thesis/node48.html

[5] Darren K. Brock, "RSFQ Technology: Circuits and Systems", Hypres, Inc., NY, at http://www.hypres.com/papers/Brock-RSFQ-CirSys-Rev01.pdf

[6] Matthew Broersma , "Nanotubes break semiconducting record", Cnet News, December 19, 2003, at http://news.com.com/2100-1006-5129761.html

[7] "Carbon Nanotube Transistor", Physics News Graphics, May 13, 1998, at http://www.aip.org/mgr/png/html/tubefet.htm

[8] E. R. Brown, C. D. Parker, "Resonant Tunnel Diodes as Submillimetre-Wave Sources", Philosophical Transactions: Mathematical, Physical and Engineering Sciences, Vol. 354, No. 1717, The Current Status of Semiconductor Tunnelling Devices (Oct. 15, 1996), pp. 2365-2381 at http://links.jstor.org/sici?sici=1364-503X(19961015)354%3A1717%3C2365%3ARTDASS%3E2.0.CO%3B2-Q

[9] W. J. Gallagher, S. S. P. Parkin, "Development of the magnetic tunnel junction MRAM at IBM: From first junctions to a 16-Mb MRAM demonstrator chip", IBM Research and Development, Spintronics, Volume 50, Number 1, 2006, at http://www.research.ibm.com/journal/rd/501/gallagher.html

[10] "IBM, Infineon Develop Most Advanced MRAM Technology to Date", IBM Research, at http://domino.research.ibm.com/comm/pr.nsf/pages/news.20030610_mram.html

[11] Linda Geppert "Quantum Transistors: toward nanoectronic", IEEE Spectrum, September 2000, at http://www.ece.osu.edu/~berger/press/quan0900.pdf

[12] A. K. Geim1 and K. S. Novoselov1 , "The rise of graphene", Nature Materials, 6, 2007, at http://www.nature.com/nmat/journal/v6/n3/full/nmat1849.html

[13] Ilan Greenberg, "Transistor Technology Takes a Quantum Leap", Wired News, December 8, 1997, at http://www.wired.com/news/technology/0,1282,8994,00.html

[14] R. Colin Johnson, "Spintronics approach advances toward live chips," EE Times, 11/06/2006, at http://www.eetimes.com/showArticle.jhtml?articleID=193500309

[15] R. Colin Johnson " U. of Delaware researchers edge closer to spintronics," EE Times, 07/26/2007, at http://www.eetimes.com/news/design/showArticle.jhtml?articleID=201201400

[16] R. Colin Johnson, "Can metal-insulator electronics do it better, sans semiconductors?" http://www.eetimes.com/showArticle.jhtml?articleID=201200024

[17] Tony R. Kuphaldt, "Lessons in Electricity", Reference, Vol. 5, Ch 7, 2007 at http://www.ibiblio.org/obp/electricCircuits/Ref/spice.html

[18] Tom Lee, "Is nonvolatile MRAM right for your consumer embedded device application? ", Freescale Semiconductor at http://www.acumeninfo.com/subscriber/article/getArticle.jhtml?articleId=197006965

[19] HyperPhysics, "SQUID Magnetometer", HyperPhysics at http://hyperphysics.phy-astr.gsu.edu/hbase/solids/squid.html

[20] Phillip F. Schewe, Ben Stein, "A Quantum Tunneling Transistor", Physics Nessw Update, Number 357, February 4, 1998, at http://www.aip.org/pnu/1998/physnews.357.htm

[21] "Why MIIM?", Phiar Corporation, at http://www.phiar.com/whyMIIM.php4

[22] "What is Quantum Tunneling?", Phiar Corporation, at http://www.phiar.com/whatQuantum.php4

[23] Oxford University, "Theory, Superconductor Synthesis", Oxford University, 1996, at http://www.chem.ox.ac.uk/vrchemistry/super/theory.htm

[24] John Walko, "Graphene transistor to rival silicon, say researchers", EE Times Europe, 03/02/2007, at http://www.eetimes.com/news/design/showArticle.jhtml?articleID=197700700

[25] Ying-Yu Tzou,"Power Electronics: An Introduction", Institute of Control Engineering, National Chiao Tung University, at http://pemclab.cn.nctu.edu.tw/peclub/w3cnotes

Chapter 3

DIODES AND RECTIFIERS

Contents

3.1 Introduction . . . 98
3.2 Meter check of a diode . . . 103
3.3 Diode ratings . . . 107
3.4 Rectifier circuits . . . 108
3.5 Peak detector . . . 115
3.6 Clipper circuits . . . 117
3.7 Clamper circuits . . . 121
3.8 Voltage multipliers . . . 123
3.9 Inductor commutating circuits . . . 130
3.10 Diode switching circuits . . . 132
3.10.1 Logic . . . 132
3.10.2 Analog switch . . . 133
3.11 Zener diodes . . . 134
3.12 Special-purpose diodes . . . 143
3.12.1 Schottky diodes . . . 143
3.12.2 Tunnel diodes . . . 144
3.12.3 Light-emitting diodes . . . 146
3.12.4 Laser diodes . . . 150
3.12.5 Photodiodes . . . 151
3.12.6 Solar cells . . . 153
3.12.7 Varicap or varactor diodes . . . 159
3.12.8 Snap diode . . . 159
3.12.9 PIN diodes . . . 160
3.12.10 IMPATT diode . . . 160
3.12.11 Gunn diode . . . 161
3.12.12 Shockley diode . . . 162
3.12.13 Constant-current diodes . . . 162

3.13 Other diode technologies 163
3.13.1 SiC diodes 163
3.13.2 Polymer diode 163
3.14 SPICE models 163
Bibliography 171

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 kind of diode in modern circuit design is the *semiconductor* diode, although other diode technologies exist. Semiconductor diodes are symbolized in schematic diagrams such as Figure 3.1. The term "diode" is customarily reserved for small signal devices, $I \leq 1$ A. The term *rectifier* is used for power devices, $I > 1$ A.

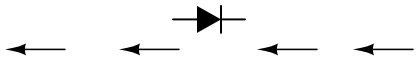


Figure 3.1: *Semiconductor diode schematic symbol: Arrows indicate the direction of electron current flow.*

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. (Figure 3.2)

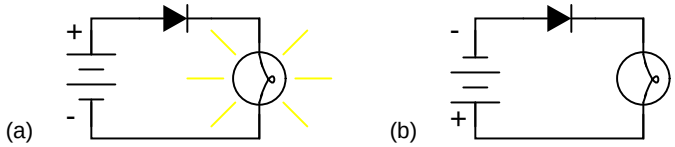


Figure 3.2: *Diode operation: (a) Current flow is permitted; the diode is forward biased. (b) Current flow is prohibited; the diode is reversed biased.*

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 like a 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 only one direction as in Figure 3.3.

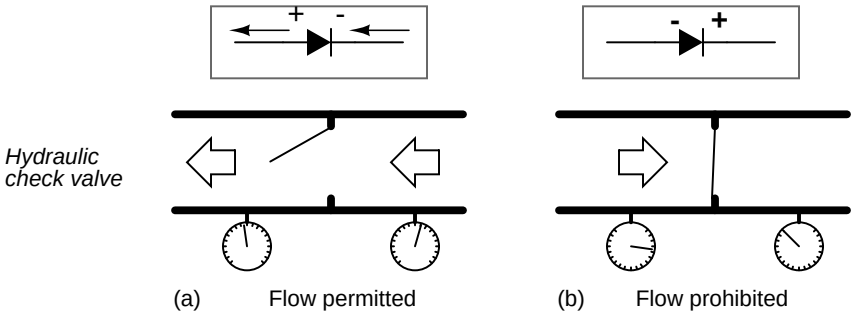


Figure 3.3: *Hydraulic check valve analogy: (a) Electron current flow permitted. (b) Current flow prohibited.*

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 in Figure 3.4.

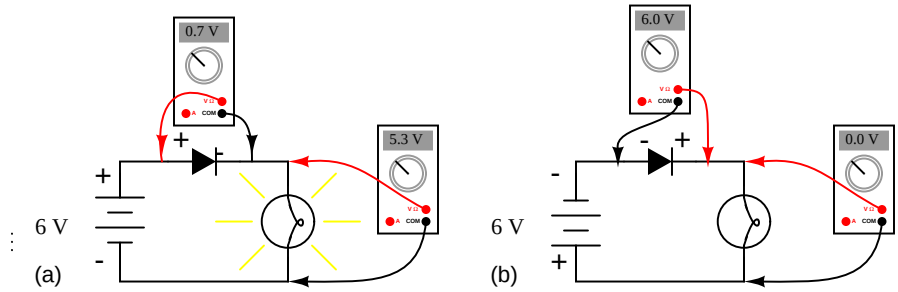


Figure 3.4: *Diode circuit voltage measurements: (a) Forward biased. (b) Reverse biased.*

A forward-biased diode conducts current and drops a small voltage across it, leaving most of the battery voltage dropped across the lamp. If the battery's polarity is reversed, the diode becomes reverse-biased, and drops *all* of the battery's voltage leaving none for the lamp. If we consider the diode to be a self-actuating switch (closed in the forward-bias mode and open in the reverse-bias mode), this behavior makes sense. The most substantial difference 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. If no voltage

applied is across a semiconductor diode, a thin depletion region exists around the region of the P-N junction, preventing current flow. (Figure 3.5 (a)) The depletion region is almost devoid of available charge carriers, and acts as an insulator:

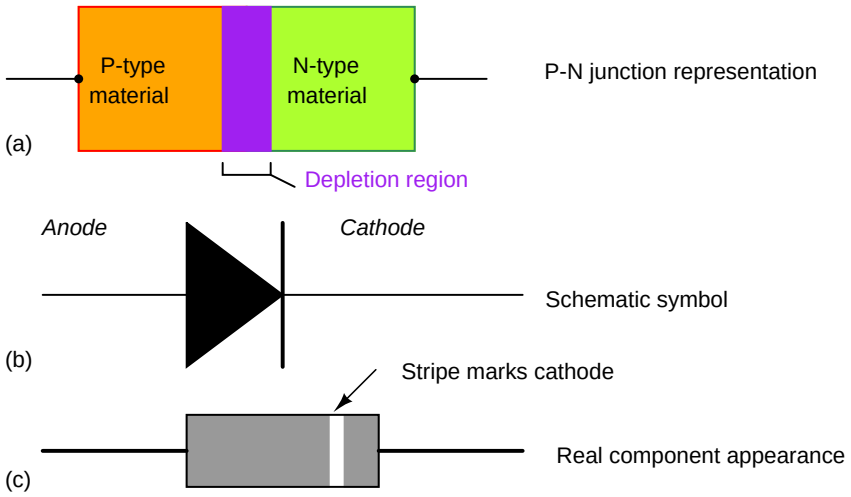


Figure 3.5: *Diode representations: PN-junction model, schematic symbol, physical part.*

The schematic symbol of the diode is shown in Figure 3.5 (b) such that the anode (pointing end) corresponds to the P-type semiconductor at (a). The cathode bar, non-pointing end, at (b) corresponds to the N-type material at (a). Also note that the cathode stripe on the physical part (c) corresponds to the cathode on the symbol.

If a reverse-biasing voltage is applied across the P-N junction, this depletion region expands, further resisting any current through it. (Figure 3.6)

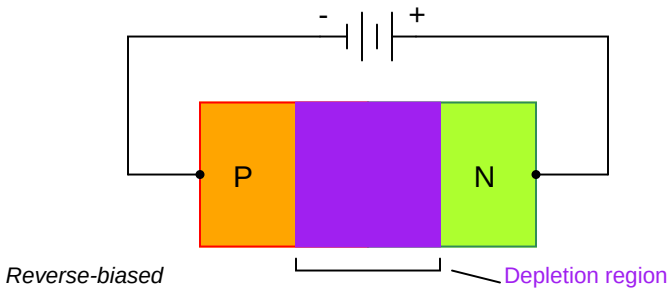


Figure 3.6: *Depletion region expands with reverse bias.*

Conversely, if a forward-biasing voltage is applied across the P-N junction, the depletion region collapses becoming thinner. 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* as illustrated in Figure 3.7.

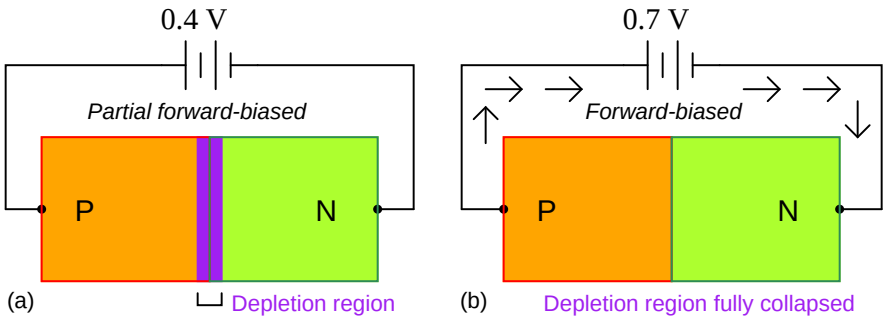


Figure 3.7: *Inceasing forward bias from (a) to (b) decreases depletion region thickness.*

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 constant for a wide range of diode currents, meaning that diode voltage drop is not like that of a resistor or even a normal (closed) switch. For most simplified circuit analysis, the voltage drop across a conducting diode may be considered constant at the nominal figure and not related to the amount of current.

Actually, forward voltage drop is more complex. An equation describes 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 Kelvins

The term 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 Number (~ 2.718281828)

V_D = Voltage applied across diode in volts

You need not be familiar with the "diode equation" 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, as it is for any insulator. If the applied reverse-bias voltage becomes too great, the diode will experience a condition known as *breakdown* (Figure 3.8), 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

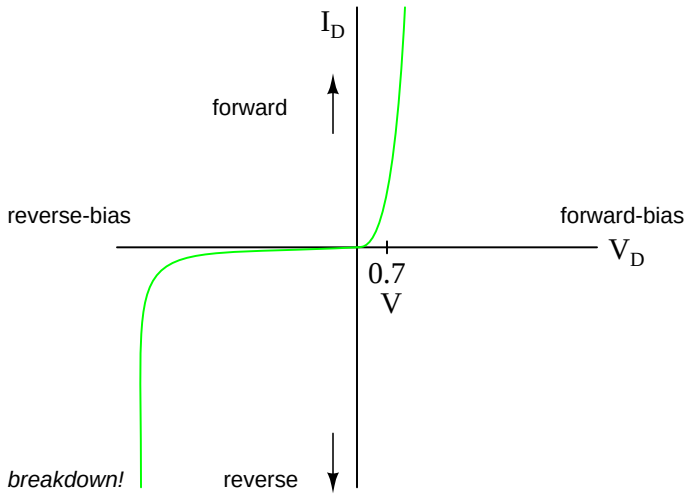


Figure 3.8: *Diode curve: showing knee at 0.7 V forward bias for Si, and reverse breakdown.*

temperature, and is fixed 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 as in Figure 3.9. Connected one way across the diode, the meter should show a very low resistance at (a). Connected the other way across the diode, it should show a very high resistance at (b) ("OL" on some digital meter models).

Of course, 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!

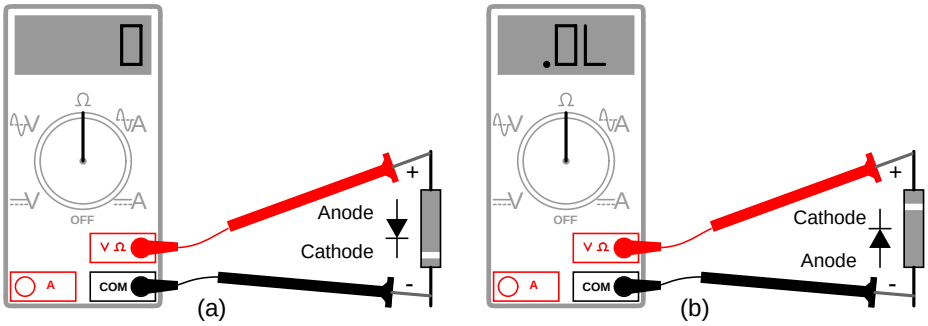


Figure 3.9: *Determination of diode polarity: (a) Low resistance indicates forward bias, black lead is cathode and red lead anode (for most meters) (b) Reversing leads shows high resistance indicating reverse bias.*

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. (Figure 3.10)

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 circuit of Figure 3.11 may be constructed using a battery, resistor, and voltmeter

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 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"

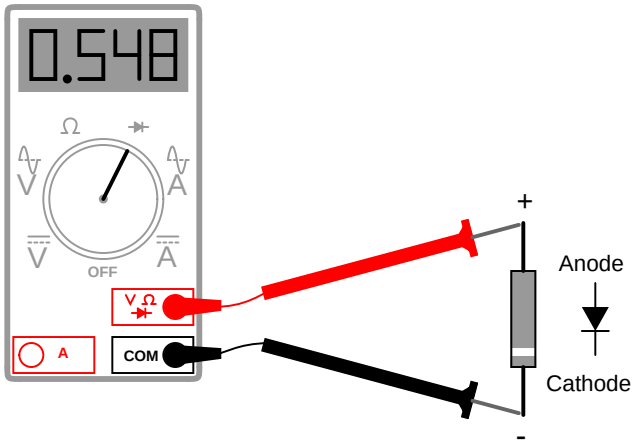


Figure 3.10: *Meter with a "Diode check" function displays the forward voltage drop of 0.548 volts instead of a low resistance.*

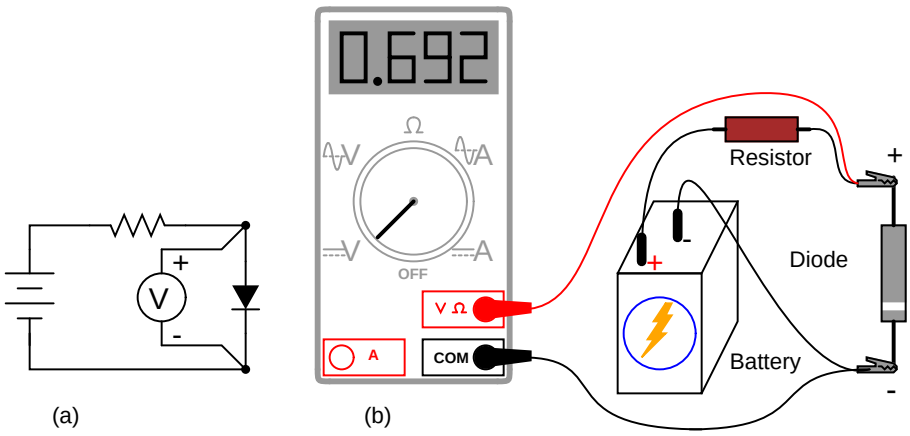


Figure 3.11: *Measuring forward voltage of a diode without"diode check" meter function: (a) Schematic diagram. (b) Pictorial diagram.*

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 measuring its resistance, otherwise any parallel-connected components would affect the reading obtained. When using a multimeter which 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 will only pass negligible current. Consequently, the meter "sees" the diode as an open (no continuity), and only registers the resistor's resistance. (Figure 3.12)

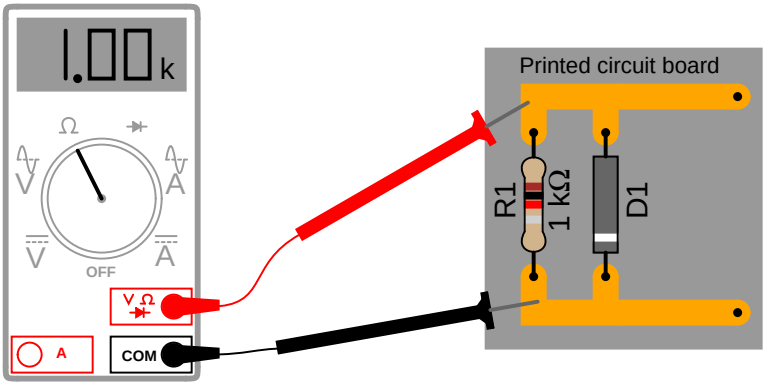


Figure 3.12: *Ohmmeter equipped with a low test voltage (<0.7 V) does not see diodes allowing it to measure parallel resistors.*

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. (Figure 3.13)

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. Special types of diodes, though, which are designed to "break down" in reverse-bias mode without damage (called *zener diodes*), which are tested with the same 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.

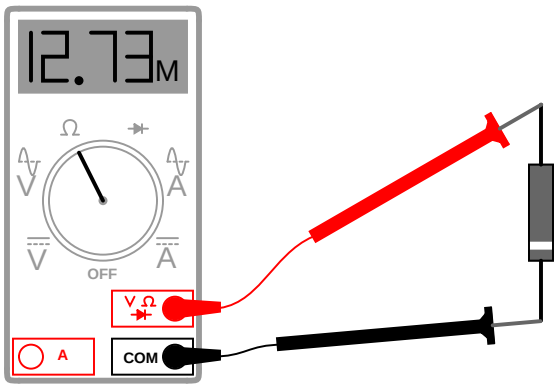


Figure 3.13: *Ohmmeter equipped with a low test voltage, too low to forward bias diodes, does not see diodes.*

- Some multimeters provide a "diode check" function that displays the actual forward voltage of the diode when its 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 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 (^oC). 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 its 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 involves a

device that only allows one-way flow of electrons. As we have seen, this is exactly what a semiconductor diode does. The simplest kind of rectifier circuit is the *half-wave* rectifier. It only allows one half of an AC waveform to pass through to the load. (Figure 3.14)

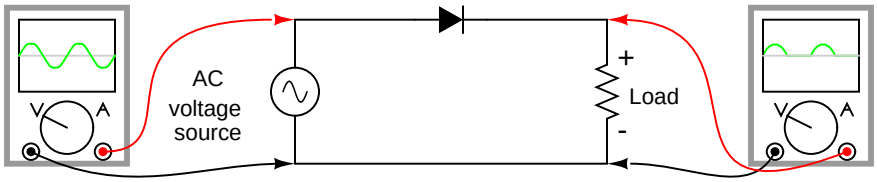


Figure 3.14: *Half-wave rectifier circuit.*

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, the AC power source only supplies power to the load one half every full cycle, meaning that half 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. (Figure 3.15)

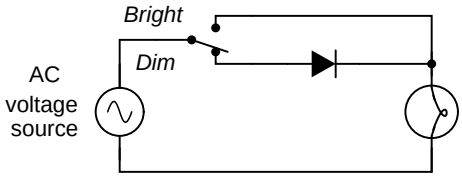


Figure 3.15: *Half-wave rectifier application: Two level lamp dimmer.*

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 to control the electrical power sent to it is 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 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 kind of full-wave rectifier, called the *center-tap* design, uses a transformer with a center-tapped secondary winding and two diodes, as in Figure 3.16.

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

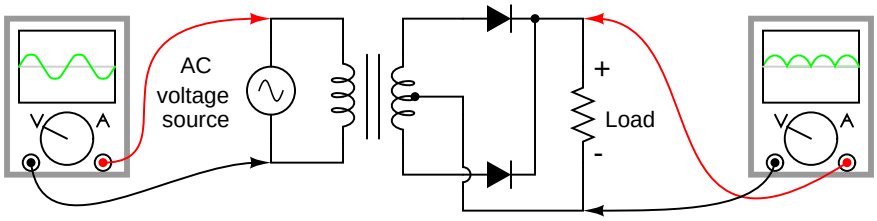


Figure 3.16: *Full-wave rectifier, center-tapped design.*

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 as in Figure 3.17.

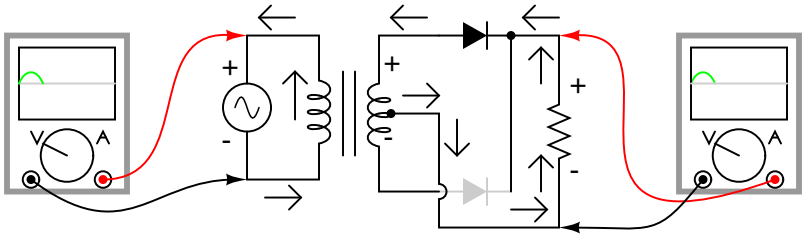


Figure 3.17: *Full-wave center-tap rectifier: Top half of secondary winding conducts during positive half-cycle of input, delivering positive half-cycle to load..*

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. (Figure 3.18)

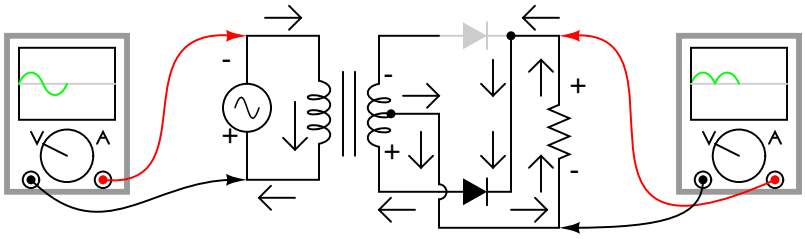


Figure 3.18: *Full-wave center-tap rectifier: During negative input half-cycle, bottom half of secondary winding conducts, delivering a positive half-cycle to the load.*

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 only seen in low-power applications.

The full-wave center-tapped rectifier polarity at the load may be reversed by changing the direction of the diodes. Furthermore, the reversed diodes can be paralleled with an existing positive-output rectifier. The result is dual-polarity full-wave center-tapped rectifier in Figure 3.19. Note that the connectivity of the diodes themselves is the same configuration as a bridge.

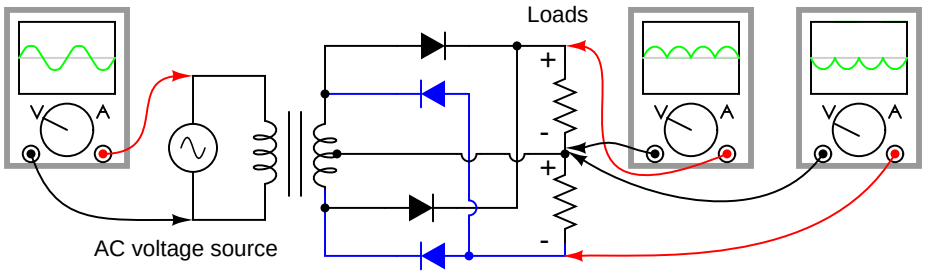


Figure 3.19: *Dual polarity full-wave center tap rectifier*

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*. (Figure 3.20)

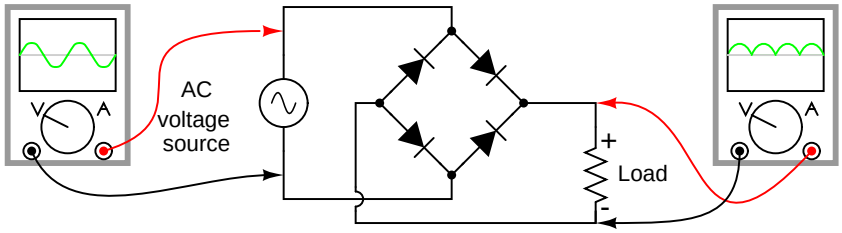


Figure 3.20: *Full-wave bridge rectifier.*

Current directions for the full-wave bridge rectifier circuit are as shown in Figure 3.21 for positive half-cycle and Figure 3.22 for negative half-cycles of the AC source waveform. Note that regardless of the polarity of the input, the current flows in the same direction through the load. That is, the negative half-cycle of source is a positive half-cycle at the load. The current flow is through two diodes in series for both polarities. Thus, two diode drops of the source voltage are lost (0.7·2=1.4 V for Si) in the diodes. This is a disadvantage compared with a full-wave center-tap design. This disadvantage is only a problem in very low voltage power supplies.

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. (Figure 3.23)

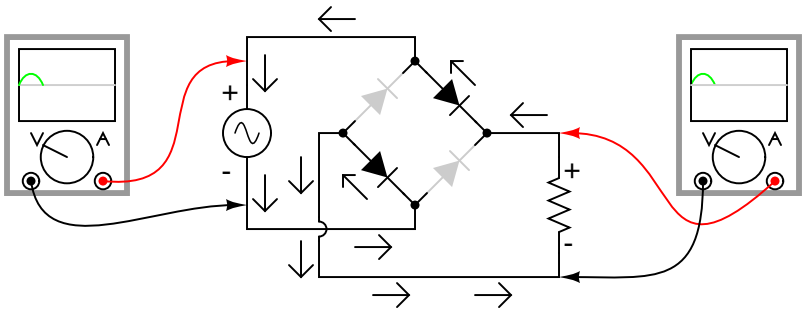


Figure 3.21: *Full-wave bridge rectifier: Electron flow for positive half-cycles.*

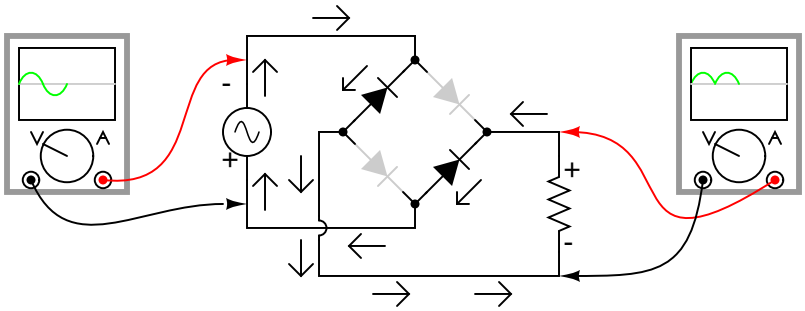


Figure 3.22: *Full-wave bridge rectifier: Electron flow for negative half=cycles.*

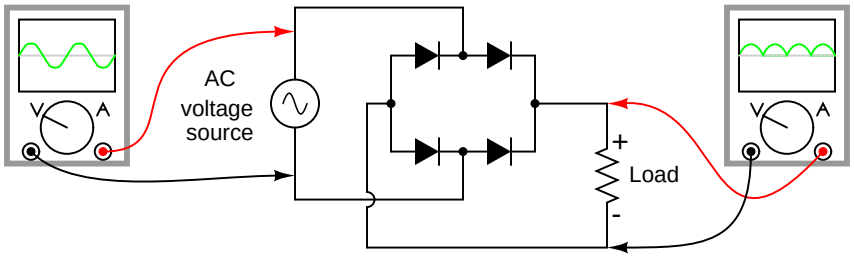


Figure 3.23: *Alternative layout style for Full-wave bridge rectifier.*

One advantage of remembering this layout for a bridge rectifier circuit is that it expands easily into a polyphase version in Figure 3.24.

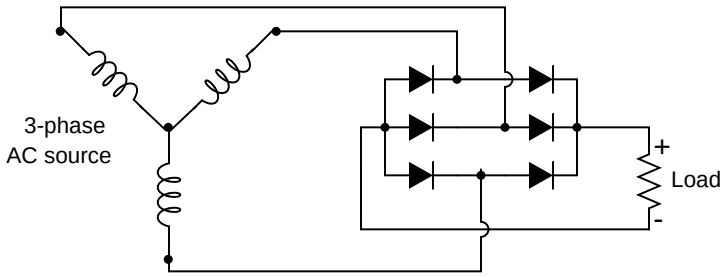


Figure 3.24: *Three-phase full-wave bridge rectifier circuit.*

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 the six-phase bridge rectifier circuit in Figure 3.25.

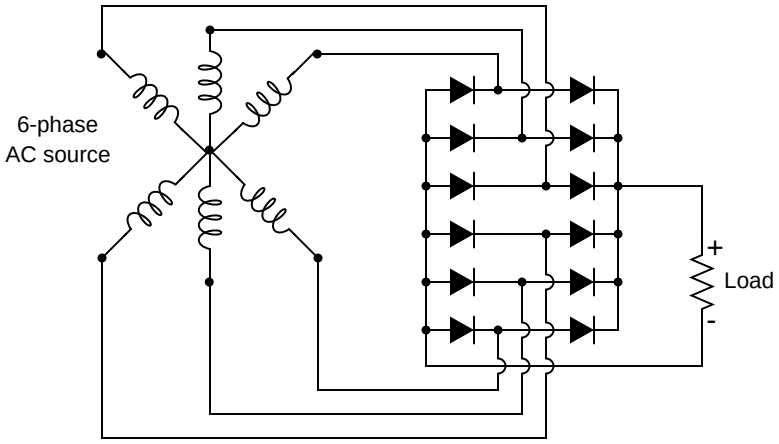


Figure 3.25: *Six-phase full-wave bridge rectifier circuit.*

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 diagram in Figure 3.26 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.

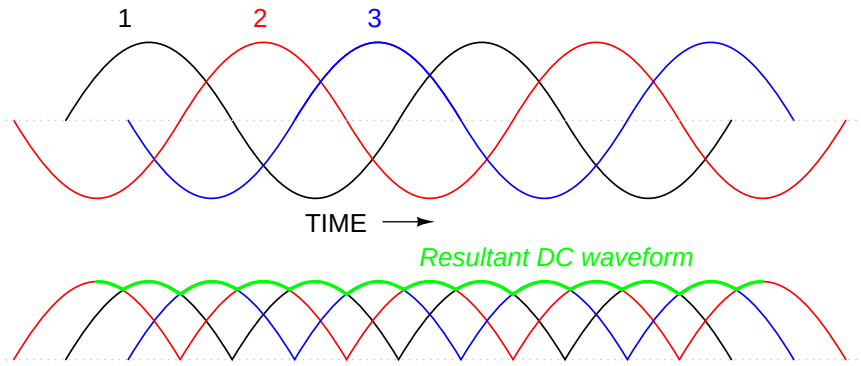


Figure 3.26: *Three-phase AC and 3-phase full-wave rectifier output.*

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 only one direction (way), 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. (Figure 3.27) 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:

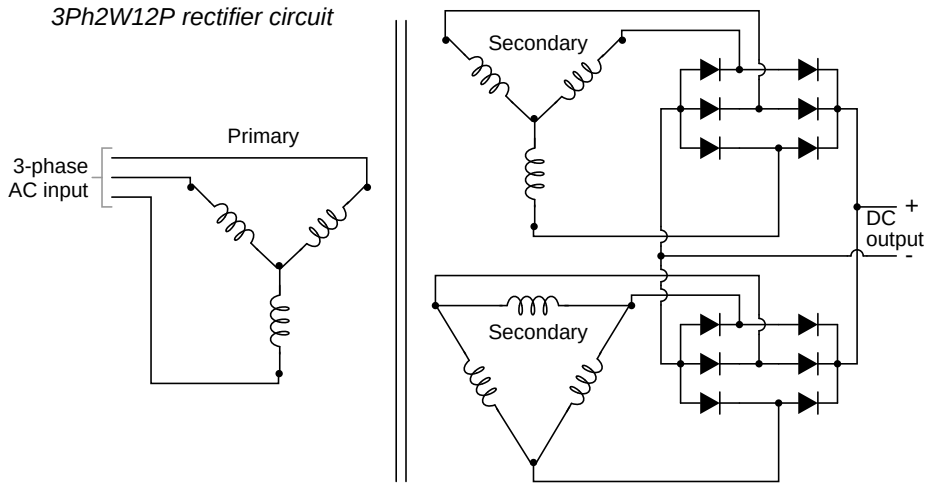


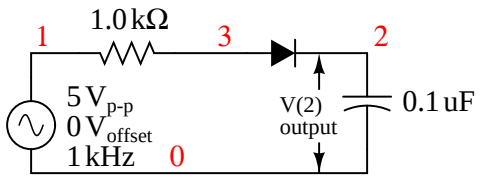
Figure 3.27: *Polyphase rectifier circuit: 3-phase 2-way 12-pulse (3Ph2W12P)*

- **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 Peak detector

A *peak detector* is a series connection of a diode and a capacitor outputting a DC voltage equal to the peak value of the applied AC signal. The circuit is shown in Figure 3.28 with the corresponding SPICE net list. An AC voltage source applied to the peak detector, charges the capacitor to the peak of the input. The diode conducts positive "half cycles," charging the capacitor to the waveform peak. When the input waveform falls below the DC "peak" stored on the capacitor, the diode is reverse biased, blocking current flow from capacitor back to the source. Thus, the capacitor retains the peak value even as the waveform drops to zero. Another view of the peak detector is that it is the same as a half-wave rectifier with a filter capacitor added to the output.

It takes a few cycles for the capacitor to charge to the peak as in Figure 3.29 due to the series resistance (RC "time constant"). Why does the capacitor not charge all the way to 5 V?




```
*SPICE 03441.eps
C1 2 0 0.1u
R1 1 3 1.0k
V1 1 0 SIN(0 5 1k)
D1 3 2 diode
.model diode d
.tran 0.01m 50mm
.end
```

Figure 3.28: *Peak detector: Diode conducts on positive half cycles charging capacitor to the peak voltage (less diode forward drop).*

It would charge to 5 V if an "ideal diode" were obtainable. However, the silicon diode has a forward voltage drop of 0.7 V which subtracts from the 5 V peak of the input.

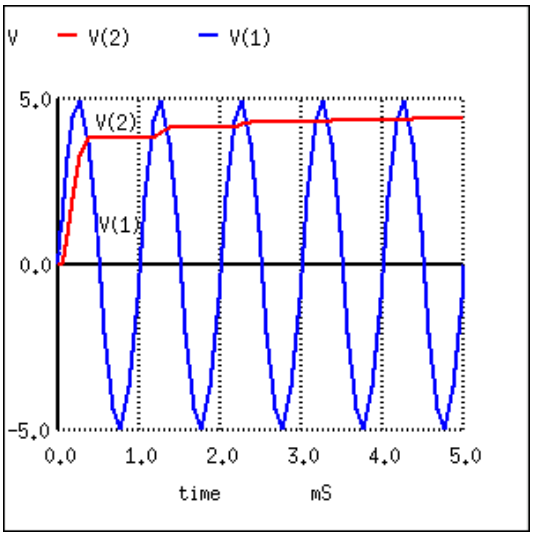


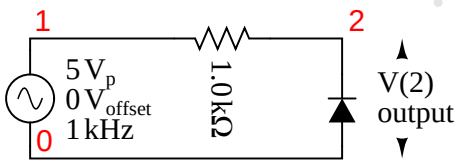
Figure 3.29: *Peak detector: Capacitor charges to peak within a few cycles.*

The circuit in Figure 3.28 could represent a DC power supply based on a half-wave rectifier. The resistance would be a few Ohms instead of 1 kΩ due to a transformer secondary winding replacing the voltage source and resistor. A larger "filter" capacitor would be used. A power supply based on a 60 Hz source with a filter of a few hundred μF could supply up to 100 mA. Half-wave supplies seldom supply more due to the difficulty of filtering a half-wave.

The peak detector may be combined with other components to build a crystal radio (page 435).

3.6 Clipper circuits

A circuit which removes the peak of a waveform is known as a *clipper*. A negative clipper is shown in Figure 3.30. This schematic diagram was produced with Xcircuit schematic capture program. Xcircuit produced the SPICE net list Figure 3.30, except for the second, and next to last pair of lines which were inserted with a text editor.




```
*SPICE 03437.eps
*  A K ModelName
D1 0 2 diode
R1 2 1 1.0k
V1 1 0 SIN(0 5 1k)
.model diode d
.tran .05m 3m
.end
```

Figure 3.30: *Clipper: clips negative peak at -0.7 V.*

During the positive half cycle of the 5 V peak input, the diode is reversed biased. The diode does not conduct. It is as if the diode were not there. The positive half cycle is unchanged at the output V(2) in Figure 3.31. Since the output positive peaks actually overlays the input sinewave V(1), the input has been shifted upward in the plot for clarity. In Nutmeg, the SPICE display module, the command "plot v(1)+1" accomplishes this.

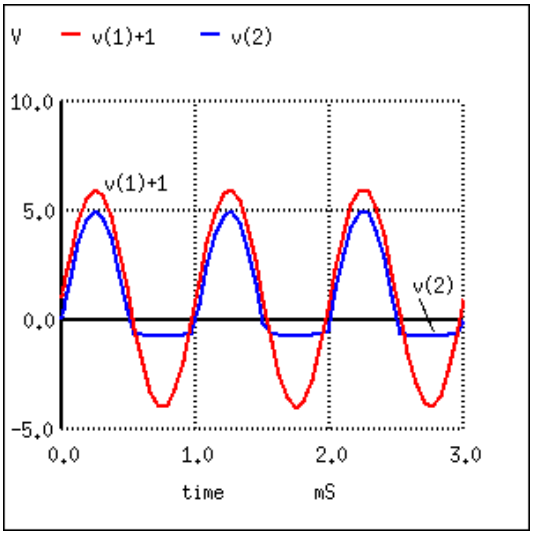


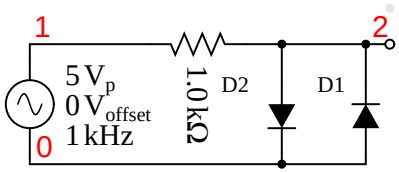
Figure 3.31: *V(1)+1 is actually V(1), a 10 Vptp sinewave, offset by 1 V for display clarity. V(2) output is clipped at -0.7 V, by diode D1.*

During the negative half cycle of sinewave input of Figure 3.31, the diode is forward biased, that is, conducting. The negative half cycle of the sinewave is shorted out. The negative half

cycle of V(2) would be clipped at 0 V for an ideal diode. The waveform is clipped at -0.7 V due to the forward voltage drop of the silicon diode. The spice model defaults to 0.7 V unless parameters in the model statement specify otherwise. Germanium or Schottky diodes clip at lower voltages.

Closer examination of the negative clipped peak (Figure 3.31) reveals that it follows the input for a slight period of time while the sinewave is moving toward -0.7 V. The clipping action is only effective after the input sinewave exceeds -0.7 V. The diode is not conducting for the complete half cycle, though, during most of it.

The addition of an anti-parallel diode to the existing diode in Figure 3.30 yields the symmetrical clipper in Figure 3.32.




```
*SPICE 03438.eps
D1 0 2 diode
D2 2 0 diode
R1 2 1 1.0k
V1 1 0 SIN(0 5 1k)
.model diode d
.tran 0.05m 3m
.end
```

Figure 3.32: *Symmetrical clipper: Anti-parallel diodes clip both positive and negative peak, leaving a ± 0.7 V output.*

Diode D1 clips the negative peak at -0.7 V as before. The additional diode D2 conducts for positive half cycles of the sine wave as it exceeds 0.7 V, the forward diode drop. The remainder of the voltage drops across the series resistor. Thus, both peaks of the input sinewave are clipped in Figure 3.33. The net list is in Figure 3.32

The most general form of the diode clipper is shown in Figure 3.34. For an ideal diode, the clipping occurs at the level of the clipping voltage, V1 and V2. However, the voltage sources have been adjusted to account for the 0.7 V forward drop of the real silicon diodes. D1 clips at 1.3V +0.7V=2.0V when the diode begins to conduct. D2 clips at -2.3V -0.7V=-3.0V when D2 conducts.

The clipper in Figure 3.34 does not have to clip both levels. To clip at one level with one diode and one voltage source, remove the other diode and source.

The net list is in Figure 3.34. The waveforms in Figure 3.35 show the clipping of v(1) at output v(2).

There is also a zener diode clipper circuit in the "Zener diode" section. A zener diode replaces both the diode and the DC voltage source.

A practical application of a clipper is to prevent an amplified speech signal from overdriving a radio transmitter in Figure 3.36. Over driving the transmitter generates spurious radio signals which causes interference with other stations. The clipper is a protective measure.

A sinewave may be squared up by overdriving a clipper. Another clipper application is the protection of exposed inputs of integrated circuits. The input of the IC is connected to a pair of diodes as at node "2" of Figure 3.34. The voltage sources are replaced by the power supply rails of the IC. For example, CMOS IC's use 0V and +5 V. Analog amplifiers might use ±12V for the V1 and V2 sources.

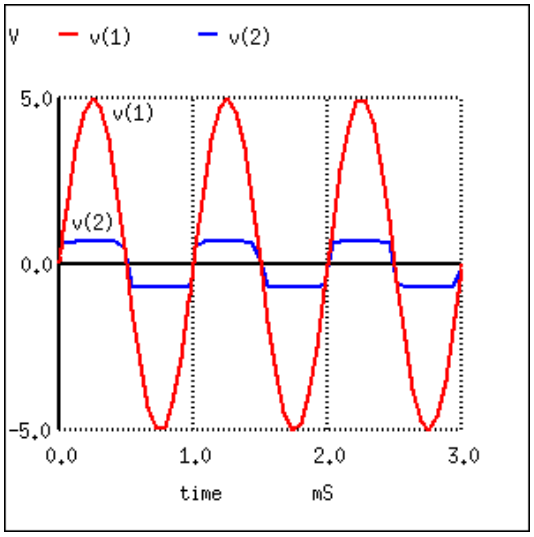
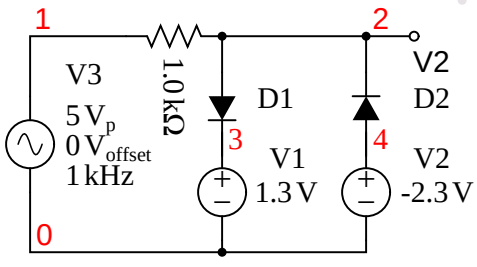


Figure 3.33: *Diode D1 clips at -0.7 V as it conducts during negative peaks. D2 conducts for positive peaks, clipping at 0.7V.*




```
*SPICE 03439.eps
V1 3 0 1.3
V2 4 0 -2.3
D1 2 3 diode
D2 4 2 diode
R1 2 1 1.0k
V3 1 0 SIN(0 5 1k)
.model diode d
.tran 0.05m 3m
.end
```

Figure 3.34: *D1 clips the input sinewave at 2V. D2 clips at -3V.*

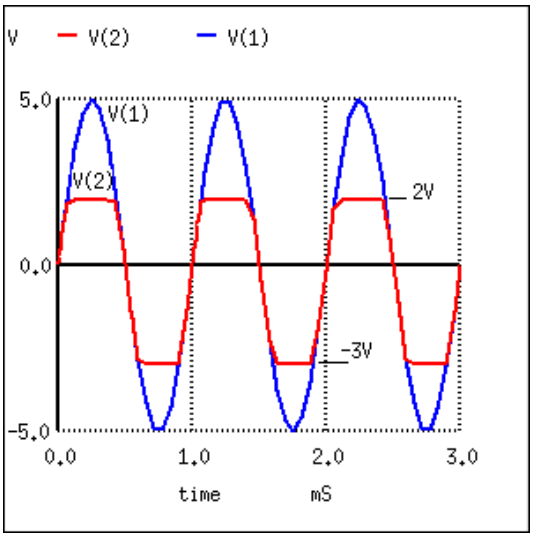


Figure 3.35: *D1 clips the sinewave at 2V. D2 clips at -3V.*

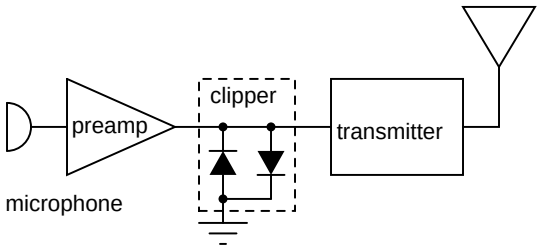


Figure 3.36: *Clipper prevents over driving radio transmitter by voice peaks.*

- **REVIEW**
- A resistor and diode driven by an AC voltage source clips the signal observed across the diode.
- A pair of anti-parallel Si diodes clip symmetrically at ± 0.7V
- The grounded end of a clipper diode(s) can be disconnected and wired to a DC voltage to clip at an arbitrary level.
- A clipper can serve as a protective measure, preventing a signal from exceeding the clip limits.

3.7 Clamper circuits

The circuits in Figure 3.37 are known as *clampers* or *DC restorers*. The corresponding netlist is in Figure 3.38. These circuits clamp a peak of a waveform to a specific DC level compared with a capacitively coupled signal which swings about its average DC level (usually 0V). If the diode is removed from the clamper, it defaults to a simple coupling capacitor– no clamping.

What is the clamp voltage? And, which peak gets clamped? In Figure 3.37 (a) the clamp voltage is 0 V ignoring diode drop, (more exactly 0.7 V with Si diode drop). In Figure 3.38, the positive peak of V(1) is clamped to the 0 V (0.7 V) clamp level. Why is this? On the first positive half cycle, the diode conducts charging the capacitor left end to +5 V (4.3 V). This is -5 V (-4.3 V) on the right end at V(1,4). Note the polarity marked on the capacitor in Figure 3.37 (a). The right end of the capacitor is -5 V DC (-4.3 V) with respect to ground. It also has an AC 5 V peak sinewave coupled across it from source V(4) to node 1. The sum of the two is a 5 V peak sine riding on a - 5 V DC (-4.3 V) level. The diode only conducts on successive positive excursions of source V(4) if the peak V(4) exceeds the charge on the capacitor. This only happens if the charge on the capacitor drained off due to a load, not shown. The charge on the capacitor is equal to the positive peak of V(4) (less 0.7 diode drop). The AC riding on the negative end, right end, is shifted down. The positive peak of the waveform is clamped to 0 V (0.7 V) because the diode conducts on the positive peak.

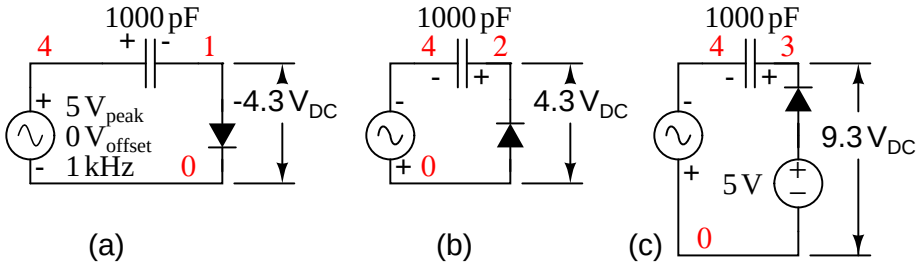
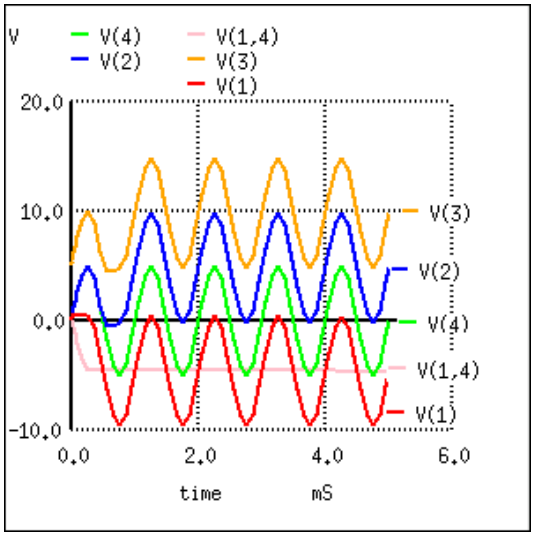


Figure 3.37: *Clampers: (a) Positive peak clamped to 0 V. (b) Negative peak clamped to 0 V. (c) Negative peak clamped to 5 V.*




```
*SPICE 03443.eps
V1 6 0 5
D1 6 3 diode
C1 4 3 1000p
D2 0 2 diode
C2 4 2 1000p
C3 4 1 1000p
D3 1 0 diode
V2 4 0 SIN(0 5 1k)
.model diode d
.tran 0.01m 5m
.end
```

Figure 3.38: *V(4) source voltage 5 V peak used in all clampers. V(1) clamper output from Figure 3.37 (a). V(1,4) DC voltage on capacitor in Figure (a). V(2) clamper output from Figure (b). V(3) clamper output from Figure (c).*

Suppose the polarity of the diode is reversed as in Figure 3.37 (b)? The diode conducts on the negative peak of source V(4). The negative peak is clamped to 0 V (-0.7 V). See V(2) in Figure 3.38.

The most general realization of the clamper is shown in Figure 3.37 (c) with the diode connected to a DC reference. The capacitor still charges during the negative peak of the source. Note that the polarities of the AC source and the DC reference are series aiding. Thus, the capacitor charges to the sum to the two, 10 V DC (9.3 V). Coupling the 5 V peak sinewave across the capacitor yields Figure 3.38 V(3), the sum of the charge on the capacitor and the sinewave. The negative peak appears to be clamped to 5 V DC (4.3V), the value of the DC clamp reference (less diode drop).

Describe the waveform if the DC clamp reference is changed from 5 V to 10 V. The clamped waveform will shift up. The negative peak will be clamped to 10 V (9.3). Suppose that the amplitude of the sine wave source is increased from 5 V to 7 V? The negative peak clamp level will remain unchanged. Though, the amplitude of the sinewave output will increase.

An application of the clamper circuit is as a “DC restorer” in “composite video” circuitry in both television transmitters and receivers. An NTSC (US video standard) video signal “white level” corresponds to minimum (12.5%) transmitted power. The video “black level” corresponds to a high level (75% of transmitter power. There is a “blacker than black level” corresponding to 100% transmitted power assigned to synchronization signals. The NTSC signal contains both video and synchronization pulses. The problem with the composite video is that its average DC level varies with the scene, dark vs light. The video itself is supposed to vary. However, the sync must always peak at 100%. To prevent the sync signals from drifting with changing scenes, a “DC restorer” clamps the top of the sync pulses to a voltage corresponding to 100%

transmitter modulation. [2]

- REVIEW:
- A capacitively coupled signal alternates about its average DC level (0 V).
- The signal out of a clamper appears the have one peak clamped to a DC voltage. Example: The negative peak is clamped to 0 VDC, the waveform appears to be shifted upward. The polarity of the diode determines which peak is clamped.
- An application of a clamper, or DC restorer, is in clamping the sync pulses of composite video to a voltage corresponding to 100% of transmitter power.

3.8 Voltage multipliers

A *voltage multiplier* is a specialized rectifier circuit producing an output which is theoretically an integer times the AC peak input, for example, 2, 3, or 4 times the AC peak input. Thus, it is possible to get 200 VDC from a 100 V_{peak} AC source using a doubler, 400 VDC from a quadrupler. Any load in a practical circuit will lower these voltages.

A voltage doubler application is a DC power supply capable of using either a 240 VAC or 120 VAC source. The supply uses a switch selected full-wave bridge to produce about 300 VDC from a 240 VAC source. The 120 V position of the switch rewires the bridge as a doubler producing about 300 VDC from the 120 VAC. In both cases, 300 VDC is produced. This is the input to a switching regulator producing lower voltages for powering, say, a personal computer.

The half-wave voltage doubler in Figure 3.39 (a) is composed of two circuits: a clamper at (b) and peak detector (half-wave rectifier) in Figure 3.28, which is shown in modified form in Figure 3.39 (c). C2 has been added to a peak detector (half-wave rectifier).

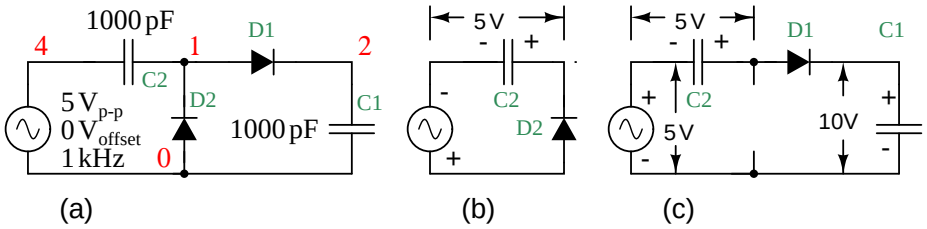
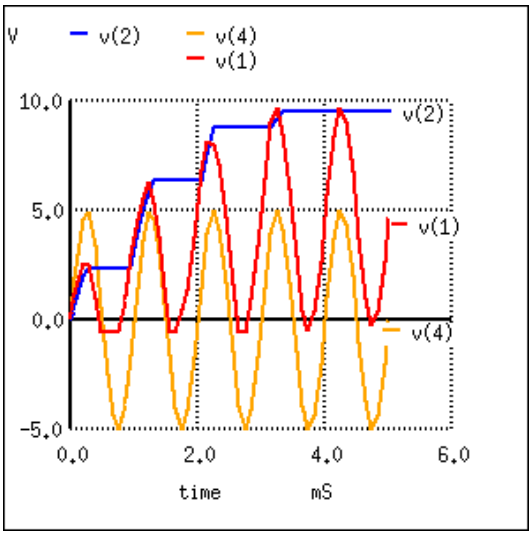


Figure 3.39: *Half-wave voltage doubler (a) is composed of (b) a clamper and (c) a half-wave rectifier.*

Referring to Figure 3.39 (b), C2 charges to 5 V (4.3 V considering the diode drop) on the negative half cycle of AC input. The right end is grounded by the conducting D2. The left end is charged at the negative peak of the AC input. This is the operation of the clamper.

During the positive half cycle, the half-wave rectifier comes into play at Figure 3.39 (c). Diode D2 is out of the circuit since it is reverse biased. C2 is now in series with the voltage source. Note the polarities of the generator and C2, series aiding. Thus, rectifier D1 sees a total of 10 V at the peak of the sinewave, 5 V from generator and 5 V from C2. D1 conducts waveform

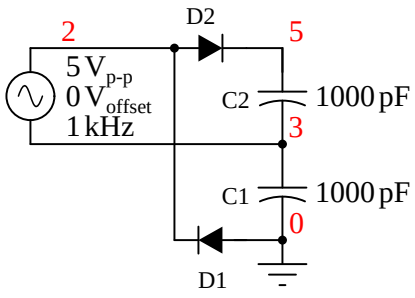
v(1) (Figure 3.40), charging C1 to the peak of the sine wave riding on 5 V DC (Figure 3.40 v(2)). Waveform v(2) is the output of the doubler, which stabilizes at 10 V (8.6 V with diode drops) after a few cycles of sinewave input.




```
*SPICE 03255.eps
C1 2 0 1000p
D1 1 2 diode
C2 4 1 1000p
D2 0 1 diode
V1 4 0 SIN(0 5 1k)
.model diode d
.tran 0.01m 5m
.end
```

Figure 3.40: *Voltage doubler: v(4) input. v(1) clamper stage. v(2) half-wave rectifier stage, which is the doubler output.*

The *full-wave voltage doubler* is composed of a pair of series stacked half-wave rectifiers. (Figure 3.41) The corresponding netlist is in Figure 3.41. The bottom rectifier charges C1 on the negative half cycle of input. The top rectifier charges C2 on the positive halfcycle. Each capacitor takes on a charge of 5 V (4.3 V considering diode drop). The output at node 5 is the series total of C1 + C2 or 10 V (8.6 V with diode drops).




```
*SPICE 03273.eps
*R1 3 0 100k
*R2 5 3 100k
D1 0 2 diode
D2 2 5 diode
C1 3 0 1000p
C2 5 3 1000p
V1 2 3 SIN(0 5 1k)
.model diode d
.tran 0.01m 5m
.end
```

Figure 3.41: *Full-wave voltage doubler consists of two half-wave rectifiers operating on alternating polarities.*

Note that the output v(5) Figure 3.42 reaches full value within one cycle of the input v(2) excursion.

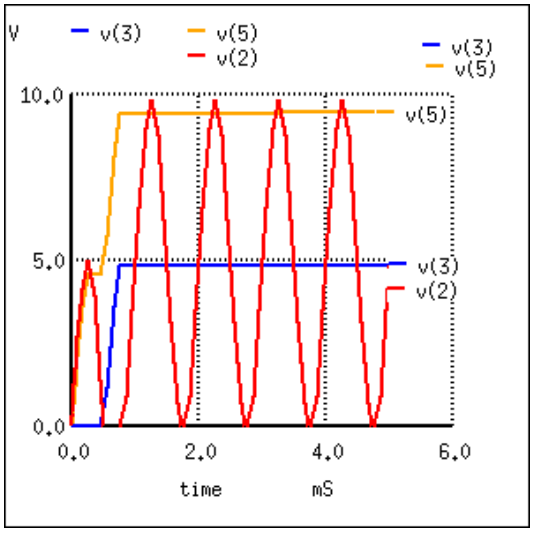


Figure 3.42: *Full-wave voltage doubler: v(2) input, v(3)voltage at mid point, v(5) voltage at output*

Figure 3.43 illustrates the derivation of the full-wave doubler from a pair of opposite polarity half-wave rectifiers (a). The negative rectifier of the pair is redrawn for clarity (b). Both are combined at (c) sharing the same ground. At (d) the negative rectifier is re-wired to share one voltage source with the positive rectifier. This yields a ± 5 V (4.3 V with diode drop) power supply; though, 10 V is measurable between the two outputs. The ground reference point is moved so that +10 V is available with respect to ground.

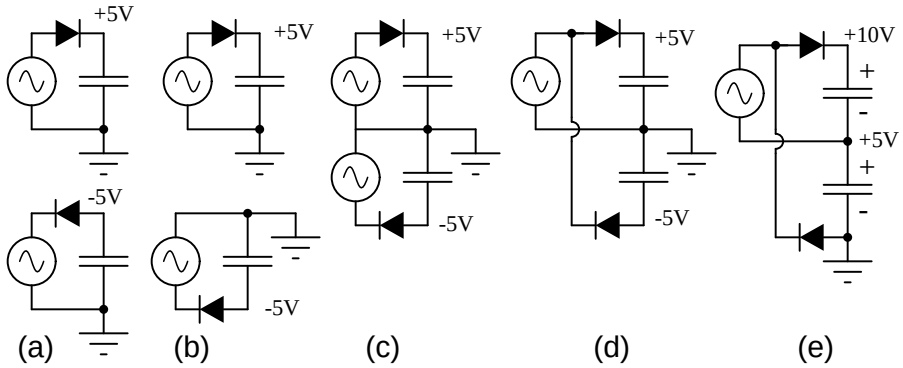


Figure 3.43: *Full-wave doubler: (a) Pair of doublers, (b) redrawn, (c) sharing the ground, (d) share the same voltage source. (e) move the ground point.*

A *voltage tripler* (Figure 3.44) is built from a combination of a doubler and a half wave rectifier (C3, D3). The half-wave rectifier produces 5 V (4.3 V) at node 3. The doubler provides another 10 V (8.4 V) between nodes 2 and 3. for a total of 15 V (12.9 V) at the output node 2 with respect to ground. The netlist is in Figure 3.45.

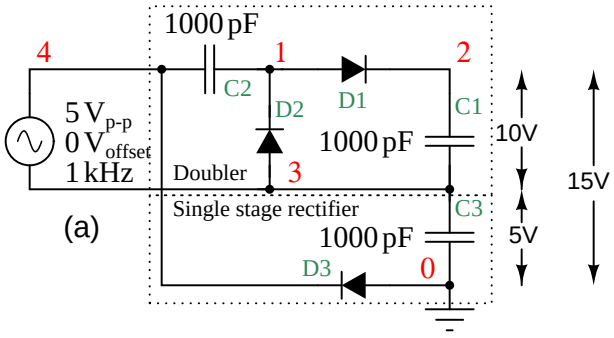
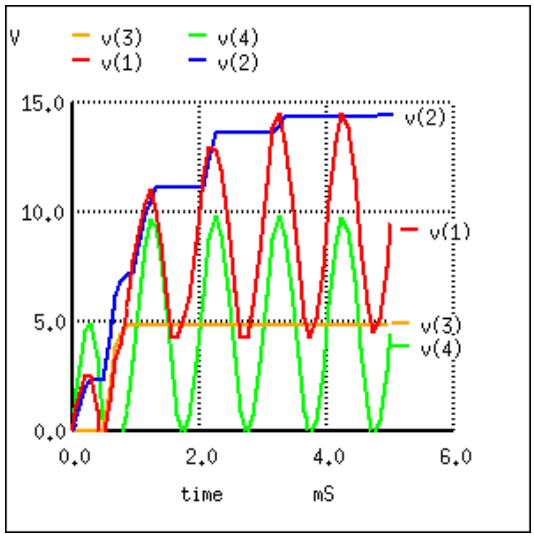


Figure 3.44: *Voltage tripler composed of doubler stacked atop a single stage rectifier.*

Note that V(3) in Figure 3.45 rises to 5 V (4.3 V) on the first negative half cycle. Input v(4) is shifted upward by 5 V (4.3 V) due to 5 V from the half-wave rectifier. And 5 V more at v(1) due to the clamper (C2, D2). D1 charges C1 (waveform v(2)) to the peak value of v(1).




```
*SPICE 03283.eps
C3 3 0 1000p
D3 0 4 diode
C1 2 3 1000p
D1 1 2 diode
C2 4 1 1000p
D2 3 1 diode
V1 4 3 SIN(0 5 1k)
.model diode d
.tran 0.01m 5m
.end
```

Figure 3.45: *Voltage tripler: v(3) half-wave rectifier, v(4) input+ 5 V, v(1) clamper, v(2) final output.*

A *voltage quadrupler* is a stacked combination of two doublers shown in Figure 3.46. Each doubler provides 10 V (8.6 V) for a series total at node 2 with respect to ground of 20 V (17.2

V). The netlist is in Figure 3.47.

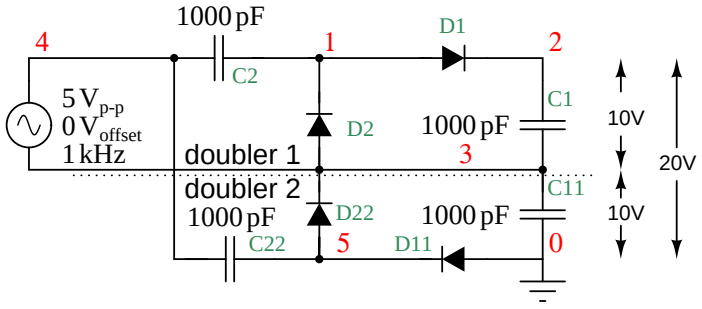
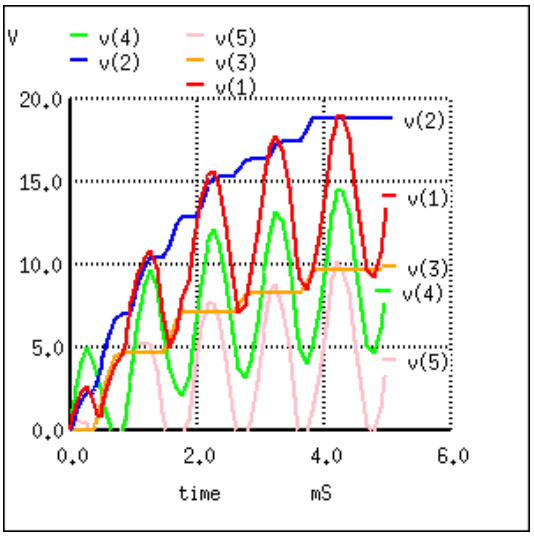


Figure 3.46: *Voltage quadrupler, composed of two doublers stacked in series, with output at node 2.*

The waveforms of the quadrupler are shown in Figure 3.47. Two DC outputs are available: v(3), the doubler output, and v(2) the quadrupler output. Some of the intermediate voltages at clampers illustrate that the input sinewave (not shown), which swings by ¡plusminus)¿5 V, is successively clamped at higher levels: at v(5), v(4) and v(1). Strictly v(4) is not a clamper output. It is simply the AC voltage source in series with the v(3) the doubler output. None the less, v(1) is a clamped version of v(4)




```
*SPICE 03441.eps
*SPICE 03286.eps
C22 4 5 1000p
C11 3 0 1000p
D11 0 5 diode
D22  5 3 diode
C1 2 3 1000p
D1  1 2 diode
C2 4 1 1000p
D2  3 1 diode
V1 4 3 SIN(0 5 1k)
.model diode d
.tran 0.01m 5m
.end
```

Figure 3.47: *Voltage quadrupler: DC voltage available at v(3) and v(2). Intermediate waveforms: Clampers: v(5), v(4), v(1).*

Some notes on voltage multipliers are in order at this point. The circuit parameters used in the examples (V= 5 V 1 kHz, C=1000 pf) do not provide much current, microamps. Further-

more, load resistors have been omitted. Loading reduces the voltages from those shown. If the circuits are to be driven by a kHz source at low voltage, as in the examples, the capacitors are usually 0.1 to 1.0 μF so that milliamps of current are available at the output. If the multipliers are driven from 50/60 Hz, the capacitor are a few hundred to a few thousand microfarads to provide hundreds of milliamps of output current. If driven from line voltage, pay attention to the polarity and voltage ratings of the capacitors.

Finally, any direct line driven power supply (no transformer) is dangerous to the experimenter and line operated test equipment. Commercial direct driven supplies are safe because the hazardous circuitry is in an enclosure to protect the user. When breadboarding these circuits with electrolytic capacitors of any voltage, the capacitors will explode if the polarity is reversed. Such circuits should be powered up behind a safety shield.

A voltage multiplier of cascaded half-wave doublers of arbitrary length is known as a *Cockcroft-Walton* multiplier as shown in Figure 3.48. This multiplier is used when a high voltage at low current is required. The advantage over a conventional supply is that an expensive high voltage transformer is not required– at least not as high as the output.

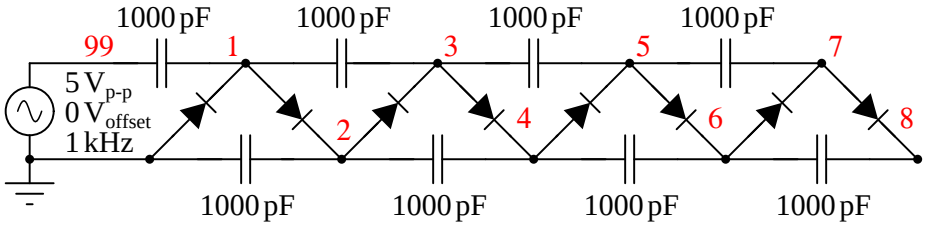
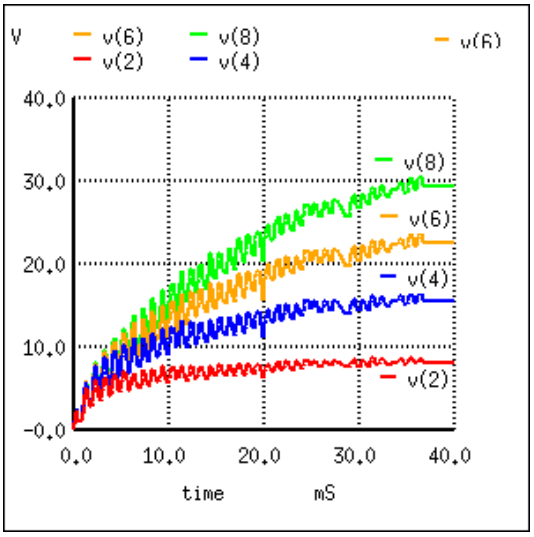


Figure 3.48: *Cockcroft-Walton x8 voltage multiplier; output at v(8).*

The pair of diodes and capacitors to the left of nodes 1 and 2 in Figure 3.48 constitute a half-wave doubler. Rotating the diodes by 45^o counterclockwise, and the bottom capacitor by 90^o makes it look like Figure 3.39 (a). Four of the doubler sections are cascaded to the right for a theoretical x8 multiplication factor. Node 1 has a clamper waveform (not shown), a sinewave shifted up by 1x (5 V). The other odd numbered nodes are sinewaves clamped to successively higher voltages. Node 2, the output of the first doubler, is a 2x DC voltage v(2) in Figure 3.49. Successive even numbered nodes charge to successively higher voltages: v(4), v(6), v(8)

Without diode drops, each doubler yields 2Vin or 10 V, considering two diode drops (10-1.4)=8.6 V is realistic. For a total of 4 doublers one expects 4·8.6=34.4 V out of 40 V. Consulting Figure 3.49, v(2) is about right;however, v(8) is $<$30 V instead of the anticipated 34.4 V. The bane of the Cockcroft-Walton multiplier is that each additional stage adds less than the previous stage. Thus, a practical limit to the number of stages exist. It is possible to overcome this limitation with a modification to the basic circuit. [3] Also note the time scale of 40 msec compared with 5 ms for previous circuits. It required 40 msec for the voltages to rise to a terminal value for this circuit. The netlist in Figure 3.49 has a ".tran 0.010m 50m" command to extend the simulation time to 50 msec; though, only 40 msec is plotted.

The Cockcroft-Walton multiplier serves as a more efficient high voltage source for photomultiplier tubes requiring up to 2000 V. [3] Moreover, the tube has numerous *dynodes*, terminals requiring connection to the lower voltage "even numbered" nodes. The series string of multiplier taps replaces a heat generating resistive voltage divider of previous designs.




```
D1 7 8 diode
C1 8 6 1000p
D2 6 7 diode
C2 5 7 1000p
D3 5 6 diode
C3 4 6 1000p
D4 4 5 diode
C4 3 5 1000p
D5 3 4 diode
C5 2 4 1000p
D6 2 3 diode
D7 1 2 diode
C6 1 3 1000p
C7 2 0 1000p
C8 99 1 1000p
D8 0 1 diode
V1 99 0 SIN(0 5
1k)
.model diode d
.tran 0.01m 50m
.end
```

Figure 3.49: *Cockcroft-Walton (x8) waveforms. Output is v(8).*

An AC line operated Cockcroft-Walton multiplier provides high voltage to "ion generators" for neutralizing electrostatic charge and for air purifiers.

- **REVIEW:**
- A voltage multiplier produces a DC multiple (2,3,4, etc) of the AC peak input voltage.
- The most basic multiplier is a half-wave doubler.
- The full-wave double is a superior circuit as a doubler.
- A tripler is a half-wave doubler and a conventional rectifier stage (peak detector).
- A quadrupler is a pair of half-wave doublers
- A long string of half-wave doublers is known as a Cockcroft-Walton multiplier.

3.9 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 in Figure 3.50 with no protection against inductive kickback.

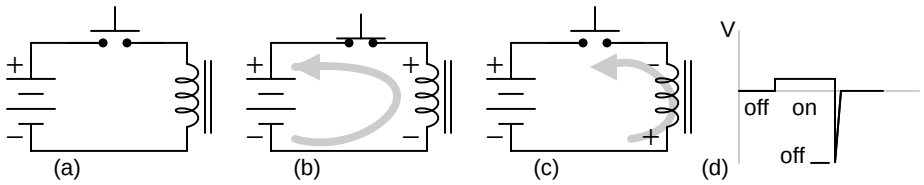


Figure 3.50: *Inductive kickback: (a) Switch open. (b) Switch closed, electron current flows from battery through coil which has polarity matching battery. Magnetic field stores energy. (c) Switch open, Current still flows in coil due to collapsing magnetic field. Note polarity change on coil. (d) Coil voltage vs time.*

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 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 causes excessive arcing at the contacts, greatly reducing their service life. Of the practical methods for mitigating the high voltage transient created when the switch is opened, none so simple as the so-called *commutating diode* in Figure 3.51.

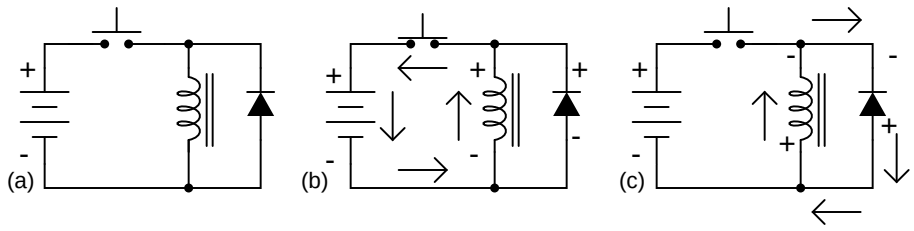


Figure 3.51: *Inductive kickback with protection: (a) Switch open. (b)Switch closed, storing energy in magnetic field. (c) Switch open, inductive kickback is shorted by diode.*

In this circuit, the diode is placed in parallel with the coil, such 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 in Figure 3.51 (b).

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 in Figure 3.51 (c).

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, for example, on an inductor coil when current through it is interrupted. 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, one cannot 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 schematic in Figure 3.52 shows how this can be done.

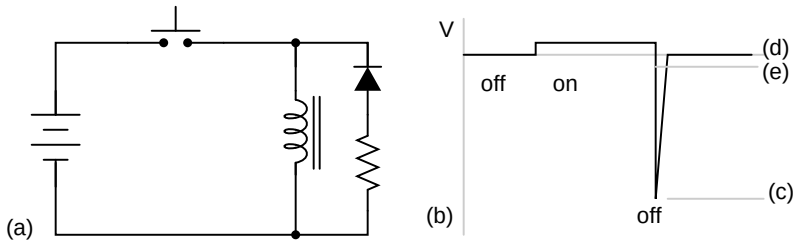


Figure 3.52: *(a) Commutating diode with series resistor. (b) Voltage waveform. (c) Level with no diode. (d) Level with diode, no resistor. (e) Compromise level with diode and resistor.*

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 demagnetization. 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.

3.10 Diode switching circuits

Diodes can perform switching and digital logic operations. Forward and reverse bias switch a diode between the low and high impedance states, respectively. Thus, it serves as a switch.

3.10.1 Logic

Diodes can perform digital logic functions: AND, and OR. Diode logic was used in early digital computers. It only finds limited application today. Sometimes it is convenient to fashion a single logic gate from a few diodes.

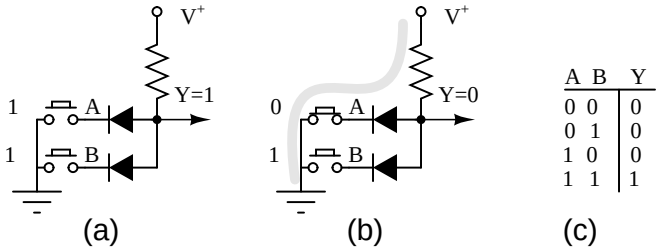


Figure 3.53: *Diode AND gate*

An AND gate is shown in Figure 3.53. Logic gates have inputs and an output (Y) which is a function of the inputs. The inputs to the gate are high (logic 1), say 10 V, or low, 0 V (logic 0). In the figure, the logic levels are generated by switches. If a switch is up, the input is

effectively high (1). If the switch is down, it connects a diode cathode to ground, which is low (0). The output depends on the combination of inputs at A and B. The inputs and output are customarily recorded in a "truth table" at (c) to describe the logic of a gate. At (a) all inputs are high (1). This is recorded in the last line of the truth table at (c). The output, Y, is high (1) due to the V^+ on the top of the resistor. It is unaffected by open switches. At (b) switch A pulls the cathode of the connected diode low, pulling output Y low (0.7 V). This is recorded in the third line of the truth table. The second line of the truth table describes the output with the switches reversed from (b). Switch B pulls the diode and output low. The first line of the truth table recordes the Output=0 for both input low (0). The truth table describes a logical AND function. Summary: both inputs A and B high yields a high (1) out.

A two input OR gate composed of a pair of diodes is shown in Figure 3.54. If both inputs are logic low at (a) as simulated by both switches "downward," the output Y is pulled low by the resistor. This logic zero is recorded in the first line of the truth table at (c). If one of the inputs is high as at (b), or the other input is high, or both inputs high, the diode(s) conduct(s), pulling the output Y high. These results are reordered in the second through fourth lines of the truth table. Summary: any input "high" is a high out at Y.

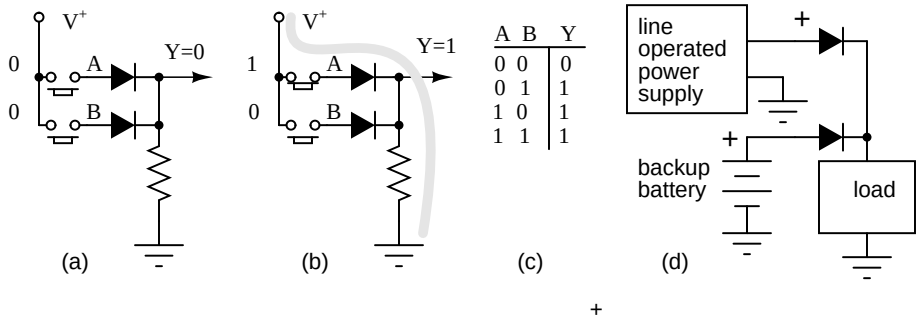


Figure 3.54: *OR gate: (a) First line, truth table (TT). (b) Third line TT. (d) Logical OR of power line supply and back-up battery.*

A backup battery may be OR-wired with a line operated DC power supply in Figure 3.54 (d) to power a load, even during a power failure. With AC power present, the line supply powers the load, assuming that it is a higher voltage than the battery. In the event of a power failure, the line supply voltage drops to 0 V; the battery powers the load. The diodes must be in series with the power sources to prevent a failed line supply from draining the battery, and to prevent it from over charging the battery when line power is available. Does your PC computer retain its BIOS setting when powered off? Does your VCR (video cassette recorder) retain the clock setting after a power failure? (PC Yes, old VCR no, new VCR yes.)

3.10.2 Analog switch

Diodes can switch analog signals. A reverse biased diode appears to be an open circuit. A forward biased diode is a low resistance conductor. The only problem is isolating the AC signal being switched from the DC control signal. The circuit in Figure 3.55 is a parallel resonant

network: resonant tuning inductor paralleled by one (or more) of the switched resonator capacitors. This parallel LC resonant circuit could be a preselector filter for a radio receiver. It could be the frequency determining network of an oscillator (not shown). The digital control lines may be driven by a microprocessor interface.

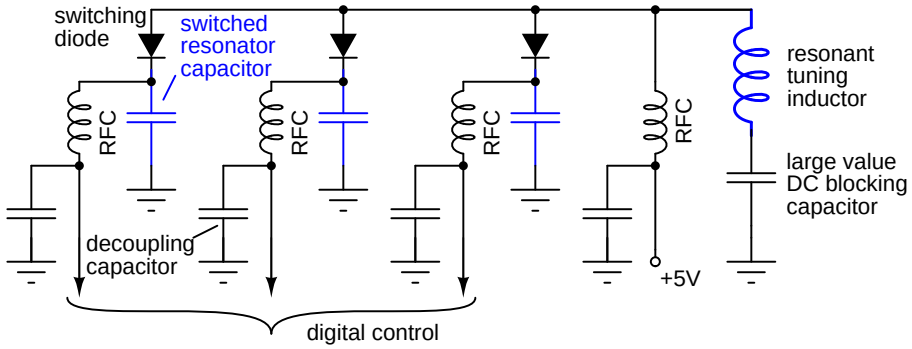


Figure 3.55: *Diode switch: A digital control signal (low) selects a resonator capacitor by forward biasing the switching diode.*

The large value DC blocking capacitor grounds the resonant tuning inductor for AC while blocking DC. It would have a low reactance compared to the parallel LC reactances. This prevents the anode DC voltage from being shorted to ground by the resonant tuning inductor. A switched resonator capacitor is selected by pulling the corresponding digital control low. This forward biases the switching diode. The DC current path is from +5 V through an RF choke (RFC), a switching diode, and an RFC to ground via the digital control. The purpose of the RFC at the +5 V is to keep AC out of the +5 V supply. The RFC in series with the digital control is to keep AC out of the external control line. The decoupling capacitor shorts the little AC leaking through the RFC to ground, bypassing the external digital control line.

With all three digital control lines high ($\geq$+5 V), no switched resonator capacitors are selected due to diode reverse bias. Pulling one or more lines low, selects one or more switched resonator capacitors, respectively. As more capacitors are switched in parallel with the resonant tuning inductor, the resonant frequency decreases.

The reverse biased diode capacitance may be substantial compared with very high frequency or ultra high frequency circuits. PIN diodes may be used as switches for lower capacitance.

3.11 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 as in Figure 3.56 (a).

According to the “diode equation” (page 101), 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 in Figure 3.56 (a), 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 in Figure 3.56 (b).

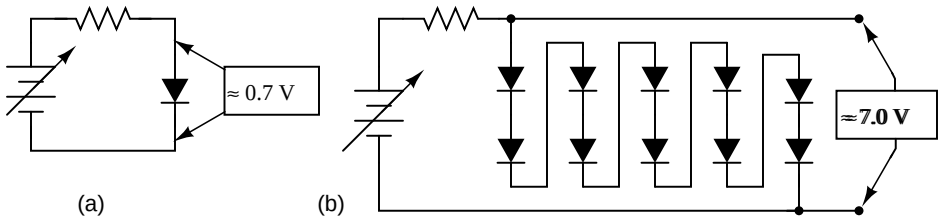


Figure 3.56: *Forward biased Si reference: (a) single diode, 0.7V, (b) 10-diodes in series 7.0V.*

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 as in Figure 3.57 (a).

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 Figure 3.57 (b).

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.

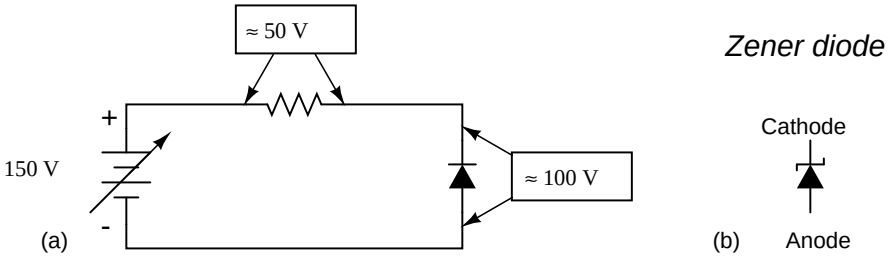


Figure 3.57: *(a) Reverse biased Si small-signal diode breaks down at about 100V. (b) Symbol for Zener diode.*

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 in Figure 3.58.

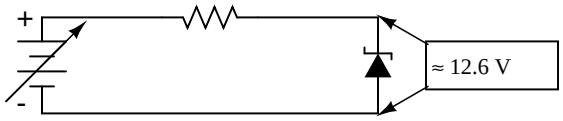


Figure 3.58: *Zener diode regulator circuit, Zener voltage = 12.6V).*

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 readily detected: 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" in Figure 3.59 (a)

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. (Figure 3.59 (b))

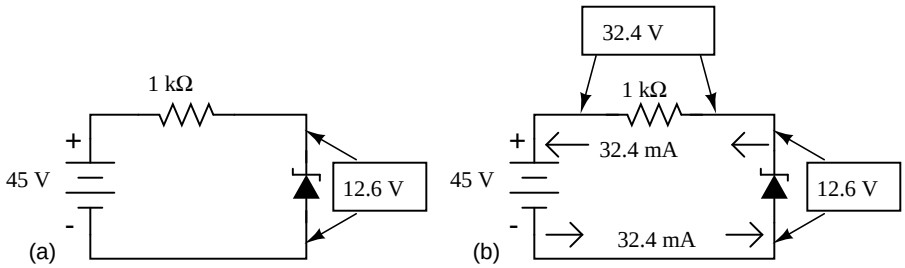


Figure 3.59: *(a) Zener Voltage regulator with 1000 Ω resistor. (b) Calculation of voltage drops and current.*

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:

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

$$P_{resistor} = 1.0498 \text{ W}$$

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

$$P_{diode} = 408.24 \text{ mW}$$

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 in Figure 3.60 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:

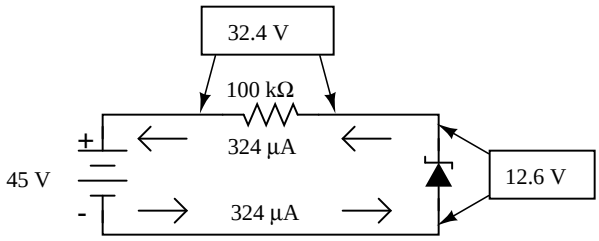


Figure 3.60: *Zener regulator with 100 kΩ resistor.*

$P_{resistor}$ = (324 µA)(32.4 V)

$P_{resistor}$ = 10.498 mW

P_{diode} = (324 µA)(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 in Figure 3.61.

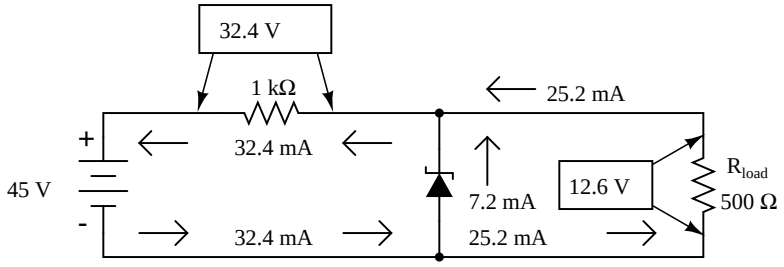


Figure 3.61: *Zener regulator with 1000 Ω series resistor and 500 Ω load.*

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.

(Figure 3.62)

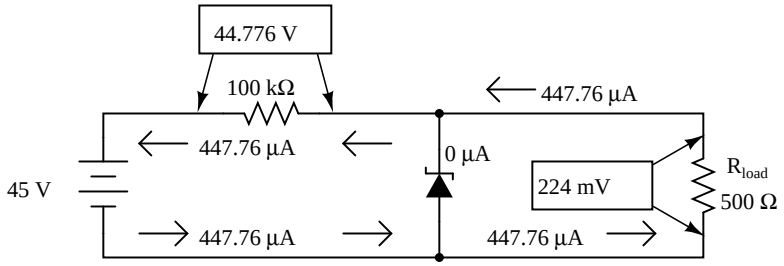


Figure 3.62: *Zener non-regulator with 100 KΩ series resistor with 500 Ω load.¿*

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 in Figure 3.63.

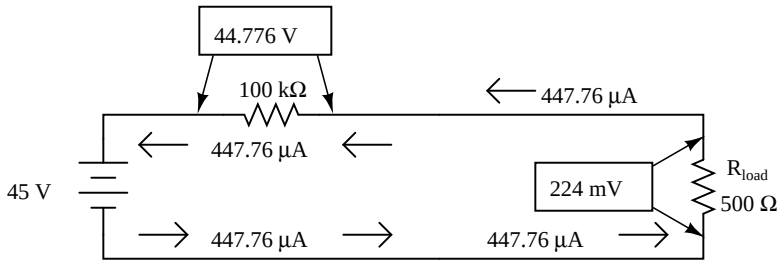


Figure 3.63: *Non-regulator with Zener removed.*

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. At least 12.6 volts must be dropped across to "activate" it.

The analytical technique of removing a zener diode from a circuit and seeing whether or not enough voltage is 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 circuit controlling the main power.

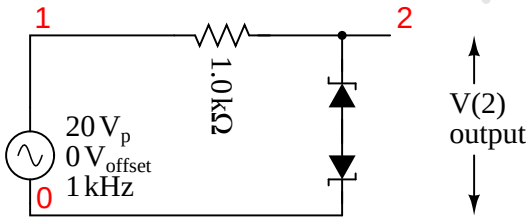
Zener diodes are manufactured in standard voltage ratings listed in Table 3.1. The table "Common zener diode voltages" lists common voltages for 0.3W and 1.3W parts. The wattage corresponds to die and package size, and is the power that the diode may dissipate without

damage.

Table 3.1: *Common zener diode voltages*

0.5W						
2.7V	3.0V	3.3V	3.6V	3.9V	4.3V	4.7V
5.1V	5.6V	6.2V	6.8V	7.5V	8.2V	9.1V
10V	11V	12V	13V	15V	16V	18V
20V	24V	27V	30V			
1.3W						
4.7V	5.1V	5.6V	6.2V	6.8V	7.5V	8.2V
9.1V	10V	11V	12V	13V	15V	16V
18V	20V	22V	24V	27V	30V	33V
36V	39V	43V	47V	51V	56V	62V
68V	75V	100V	200V			

Zener diode clipper: A clipping circuit which clips the peaks of waveform at approximately the zener voltage of the diodes. The circuit of Figure 3.64 has two zeners connected series opposing to symmetrically clip a waveform at nearly the Zener voltage. The resistor limits current drawn by the zeners to a safe value.




```
*SPICE 03445.eps
D1 4 0 diode
D2 4 2 diode
R1 2 1 1.0k
V1 1 0 SIN(0 20
1k)
.model diode d
bv=10
.tran 0.001m 2m
.end
```

Figure 3.64: *Zener diode clipper:*

The zener breakdown voltage for the diodes is set at 10 V by the diode model parameter "bv=10" in the spice net list in Figure 3.64. This causes the zeners to clip at about 10 V. The back-to-back diodes clip both peaks. For a positive half-cycle, the top zener is reverse biased, breaking down at the zener voltage of 10 V. The lower zener drops approximately 0.7 V since it is forward biased. Thus, a more accurate clipping level is 10+0.7=10.7V. Similar negative half-cycle clipping occurs a -10.7 V. (Figure 3.65) shows the clipping level at a little over $\pm$10 V.

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

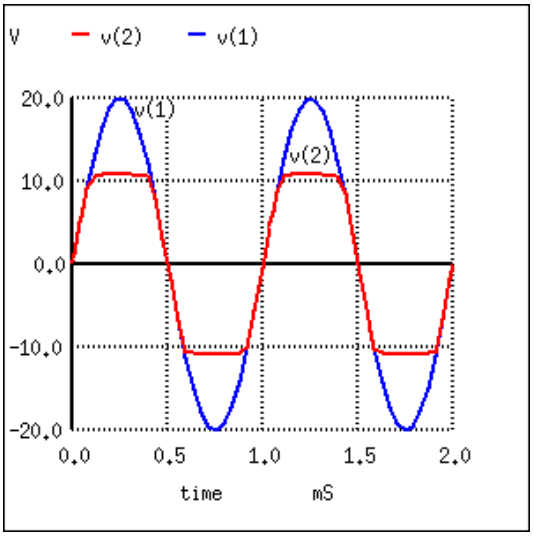


Figure 3.65: *Zener diode clipper: v(1) input is clipped at waveform v(2).*

- 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.

3.12 Special-purpose diodes

3.12.1 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 in Figure 3.66.

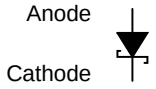


Figure 3.66: *Schottky diode schematic symbol.*

The forward voltage drop (V_F), reverse-recovery time (t_{rr}), and junction capacitance (C_J) of 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. Though they are used in low voltage switching regulator power supplies.

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.

Switching regulator power supplies operating at 100's of kHz cannot use conventional silicon diodes as rectifiers because of their slow switching speed . When the signal applied to a diode changes from forward to reverse bias, conduction continues for a short time, while carriers are being swept out of the depletion region. Conduction only ceases after this t_r *reverse recovery time* has expired. Schottky diodes have a shorter reverse recovery time.

Regardless of switching speed, the 0.7 V forward voltage drop of silicon diodes causes poor efficiency in low voltage supplies. This is not a problem in, say, a 10 V supply. In a 1 V supply the 0.7 V drop is a substantial portion of the output. One solution is to use a schottky power diode which has a lower forward drop.

3.12.2 Tunnel diodes

Tunnel diodes exploit a strange quantum phenomenon called *resonant tunneling* to provide a negative resistance forward-bias characteristics. When a small forward-bias voltage is applied across a tunnel diode, it begins to conduct current. (Figure 3.67(b)) 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." The schematic symbol for the tunnel diode shown in Figure 3.67(a).

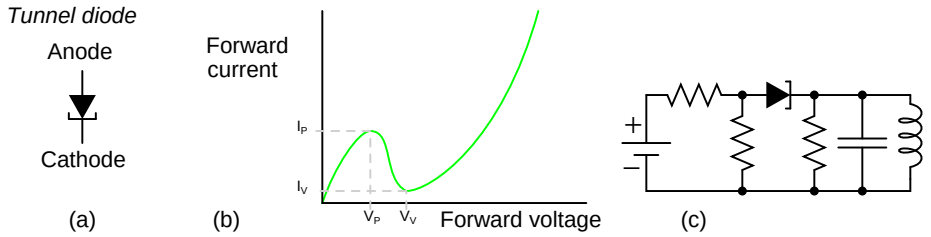


Figure 3.67: *Tunnel diode (a) Schematic symbol. (b) Current vs voltage plot (c) Oscillator.*

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.

Tunnel diodes are heavily doped in both the P and N regions, 1000 times the level in a rectifier. This can be seen in Figure 3.68. Standard diodes are to the far left, zener diodes near to the left, and tunnel diodes to the right of the dashed line. The heavy doping produces an

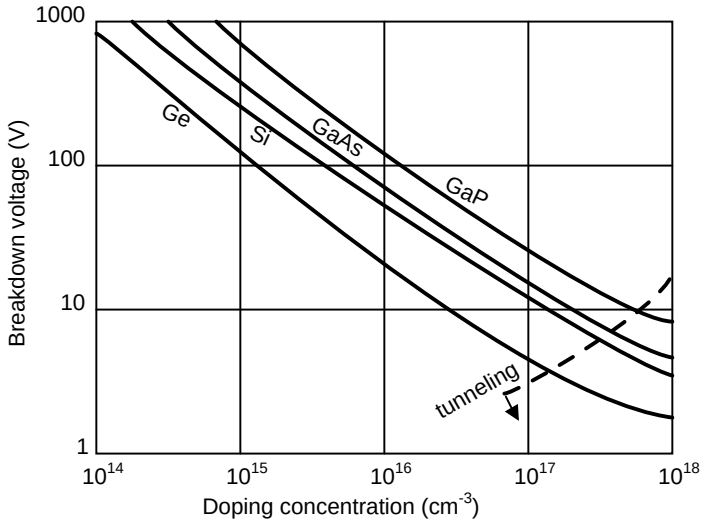


Figure 3.68: *Reverse breakdown voltage versus doping level. After Sze [22]*

unusually thin depletion region. This produces an unusually low reverse breakdown voltage with high leakage. The thin depletion region causes high capacitance. To overcome this, the tunnel diode junction area must be tiny. The forward diode characteristic consists of two regions: a normal forward diode characteristic with current rising exponentially beyond V_F, 0.3 V for Ge, 0.7 V for Si. Between 0 V and V_F is an additional "negative resistance" characteristic peak. This is due to quantum mechanical tunneling involving the dual particle-wave nature of electrons. The depletion region is thin enough compared with the equivalent wavelength of the electron that they can tunnel through. They do not have to overcome the normal forward diode voltage V_F. The energy level of the conduction band of the N-type material overlaps the level of the valence band in the P-type region. With increasing voltage, tunneling begins; the levels overlap; current increases, up to a point. As current increases further, the energy levels overlap less; current decreases with increasing voltage. This is the "negative resistance" portion of the curve.

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. 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 as in Figure 3.67(c), where it allows 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. The resistors bias the tunnel diode at a few tenths of a volt centered on the negative resistance portion of the characteristic curve. The L-C resonant circuit may be a section of waveguide for microwave operation. Oscillation to 5 GHz is possible.

At one time the tunnel diode was the only solid-state microwave amplifier available. Tunnel diodes were popular starting in the 1960's. They were longer lived than traveling wave tube amplifiers, an important consideration in satellite transmitters. Tunnel diodes are also resistant to radiation because of the heavy doping. Today various transistors operate at microwave frequencies. Even small signal tunnel diodes are expensive and difficult to find today. There is one remaining manufacturer of germanium tunnel diodes, and none for silicon devices. They are sometimes used in military equipment because they are insensitive to radiation and large temperature changes.

There has been some research involving possible integration of silicon tunnel diodes into CMOS integrated circuits. They are thought to be capable of switching at 100 GHz in digital circuits. The sole manufacturer of germanium devices produces them one at a time. A batch process for silicon tunnel diodes must be developed, then integrated with conventional CMOS processes. [21]

The Esaki tunnel diode should not be confused with the *resonant tunneling diode* (page 84), of more complex construction from compound semiconductors. The RTD is a more recent development capable of higher speed.

3.12.3 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 its *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 change 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*.

Forward biased silicon diodes give off heat as electron and holes from the N-type and P-type regions, respectively, recombine at the junction. In a forward biased LED, the recombination of electrons and holes in the active region in Figure 3.69 (c) yields photons. This process is known as *electroluminescence*. To give off photons, the potential barrier through which the electrons fall must be higher than for a silicon diode. The forward diode drop can range to a few volts for some color LEDs.

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. Early generations of LEDs were red, green, yellow, orange, and infra-red, later generations included blue and ultraviolet, with violet being the latest color added to the selection. 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. This allowed for multicolor LEDs, such as tricolor LEDs (commercially available in the 1980's) using red and green (which can create yellow) and later RGB LEDs (red, green, and blue), which cover the entire color spectrum.

The schematic symbol for an LED is a regular diode shape inside of a circle, with two small arrows pointing away (indicating emitted light), shown in Figure 3.69.

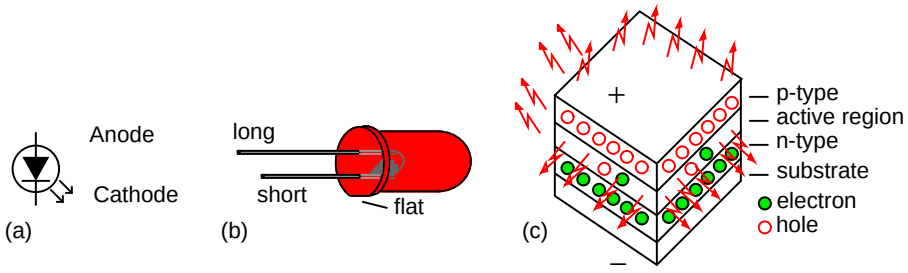


Figure 3.69: *LED, Light Emitting Diode: (a) schematic symbol. (b) Flat side and short lead of device correspond to cathode, as well as the internal arrangement of the cathode. (c) Cross section of Led die.*

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. LEDs can sense light. They 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 silicon 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 the example circuit in Figure 3.70 (a) using a 6 V source.

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 the example in Figure 3.70 (b) for a supply voltage of 24 volts:

Here, the dropping resistor must be increased to a size of 1.12 kΩ to drop 22.4 volts at 20

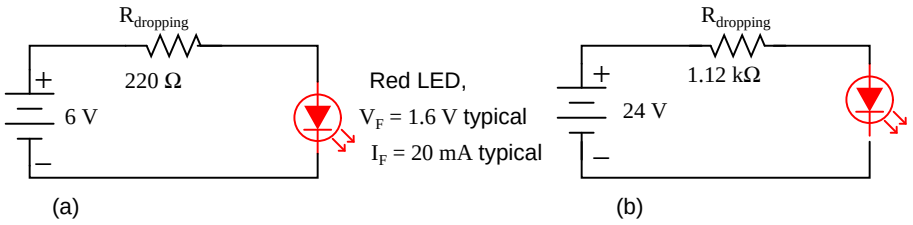


Figure 3.70: *Setting LED current at 20 ma. (a) for a 6 V source, (b) for a 24 V source.*

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.

Multiple LEDs are sometimes required, say in lighting. If LEDs are operated in parallel, each must have its own current limiting resistor as in Figure 3.71 (a) to ensure currents dividing more equally. However, it is more efficient to operate LEDs in series (Figure 3.71 (b)) with a single dropping resistor. As the number of series LEDs increases the series resistor value must decrease to maintain current, to a point. The number of LEDs in series (V_f) cannot exceed the capability of the power supply. Multiple series strings may be employed as in Figure 3.71 (c).

In spite of equalizing the currents in multiple LEDs, the brightness of the devices may not match due to variations in the individual parts. Parts can be selected for brightness matching for critical applications.

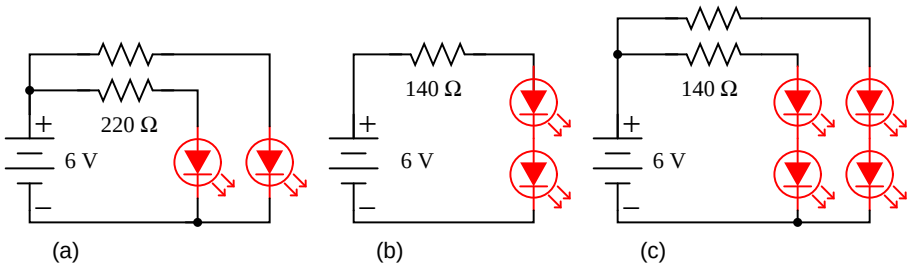


Figure 3.71: *Multiple LEDs: (a) In parallel, (b) in series, (c) series-parallel*

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,

connect a protective rectifying diode anti-parallel with the LED to prevent reverse breakdown every other half-cycle as in Figure 3.72 (a).

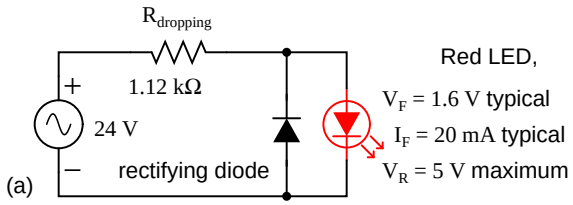


Figure 3.72: *Driving an LED with AC*

The anti-parallel diode in Figure 3.72 can be replaced with an anti-parallel LED. The resulting pair of anti-parallel LED's illuminate on alternating half-cycles of the AC sinewave. This configuration draws 20 ma, splitting it equally between the LED's on alternating AC half cycles. Each LED only receives 10 mA due to this sharing. The same is true of the LED anti-parallel combination with a rectifier. The LED only receives 10 ma. If 20 mA was required for the LED(s), The resistor value could be halved.

The forward voltage drop of LED's is inversely proportional to the wavelength (λ). As wavelength decreases going from infrared to visible colors to ultraviolet, V_f increases. While this trend is most obvious in the various devices from a single manufacturer, The voltage range for a particular color LED from various manufacturers varies. This range of voltages is shown in Table 3.2.

Table 3.2: *Optical and electrical properties of LED's*

LED	λ **nm (= 10** $^{-9}$**m)**	$\mathbf{V_f}$**(from)**	$\mathbf{V_f}$ **(to)**
infrared	940	1.2	1.7
red	660	1.5	2.4
orange	602-620	2.1	2.2
yellow, green	560-595	1.7	2.8
white, blue, violet	-	3	4
ultraviolet	370	4.2	4.8

As lamps, LEDs are superior to incandescent bulbs in many ways. First and foremost is efficiency: LEDs output far more light power per watt of electrical input 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 incandescent lamps. This is because 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).

LEDs excel in monochromatic lighting applications like traffic signals and automotive tail lights. Incandescents are abysmal in this application since they require filtering, decreasing efficiency. LEDs do not require filtering.

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. A new broad spectrum light source is the white LED. While small white panel indicators have been available for many years, illumination grade devices are still in development.

Table 3.3: *Efficiency of lighting*

Lamp type	**Efficiency lumen/watt**	**Life hrs**	**notes**
White LED	35	100,000	costly
White LED, future	100	100,000	R&D target
Incandescent	12	1000	inexpensive
Halogen	15-17	2000	high quality light
Compact fluorescent	50-100	10,000	cost effective
Sodium vapor, lp	70-200	20,000	outdoor
Mercury vapor	13-48	18,000	outdoor

A white LED is a blue LED exciting a phosphor which emits yellow light. The blue plus yellow approximates white light. The nature of the phosphor determines the characteristics of the light. A red phosphor may be added to improve the quality of the yellow plus blue mixture at the expense of efficiency. Table 3.3 compares white illumination LEDs to expected future devices and other conventional lamps. Efficiency is measured in lumens of light output per watt of input power. If the 50 lumens/watt device can be improved to 100 lumens/watt, white LEDs will be comparable to compact fluorescent lamps in efficiency.

LEDs in general have been a major subject of R&D since the 1960's. Because of this it is impractical to cover all geometries, chemistries, and characteristics that have been created over the decades. The early devices were relatively dim and took moderate currents. The efficiencies have been improved in later generations to the point it is hazardous to look closely and directly into an illuminated LED. This can result in eye damage, and the LEDs only required a minor increase in dropping voltage (Vf) and current. Modern high intensity devices have reached 180 lumens using 0.7 Amps (82 lumens/watt, Luxeon Rebel series cool white), and even higher intensity models can use even higher currents with a corresponding increase in brightness. Other developments, such as quantum dots, are the subject of current research, so expect to see new things for these devices in the future

3.12.4 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 its 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 falling 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.

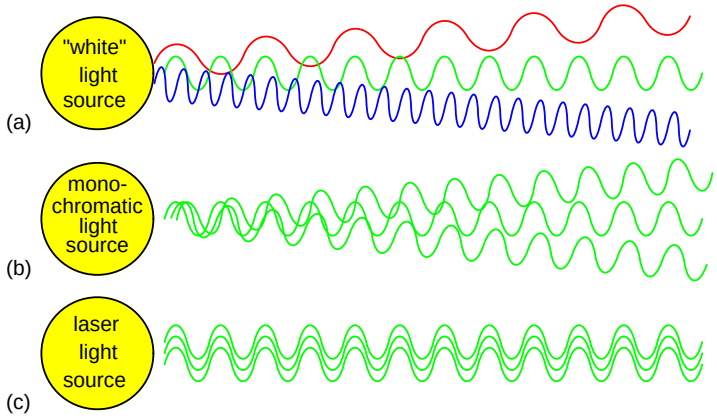


Figure 3.73: *(a) White light of many wavelengths. (b) Mono-chromatic LED light, a single wavelength. (c) Phase coherent laser light.*

Incandescent lamps produce "white" (mixed-frequency, or mixed-color) light as in Figure 3.73 (a). Regular LEDs produce monochromatic light: same frequency (color), but different phases, resulting in similar beam dispersion in Figure 3.73 (b). Laser LEDs produce *coherent light*: light that is both monochromatic (single-color) and monophasic (single-phase), resulting in precise beam confinement as in Figure 3.73 (c).

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 continuous laser, 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.

3.12.5 Photodiodes

A *photodiode* is a diode optimized to produce an electron current flow in response to irradiation by ultraviolet, visible, or infrared light. Silicon is most often used to fabricate photodiodes;

though, germanium and gallium arsenide can be used. The junction through which light enters the semiconductor must be thin enough to pass most of the light on to the active region (depletion region) where light is converted to electron hole pairs.

In Figure 3.74 a shallow P-type diffusion into an N-type wafer produces a PN junction near the surface of the wafer. The P-type layer needs to be thin to pass as much light as possible. A heavy N+ diffusion on the back of the wafer makes contact with metalization. The top metalization may be a fine grid of metallic fingers on the top of the wafer for large cells. In small photodiodes, the top contact might be a sole bond wire contacting the bare P-type silicon top.

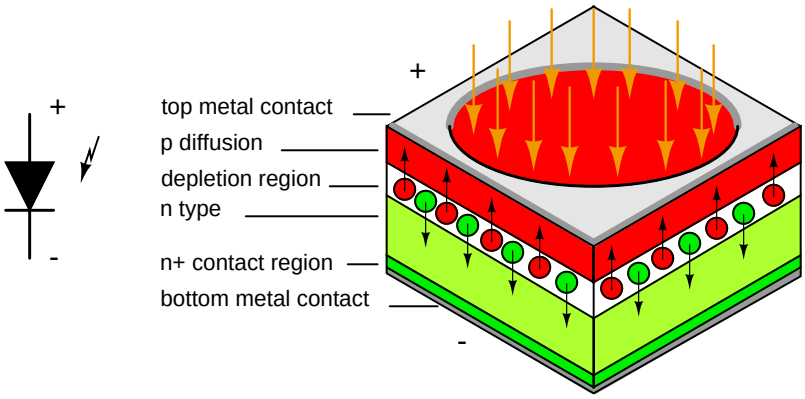


Figure 3.74: *Photodiode: Schematic symbol and cross section.*

Light entering the top of the photodiode stack fall off exponentially in with depth of the silicon. A thin top P-type layer allows most photons to pass into the depletion region where electron-hole pairs are formed. The electric field across the depletion region due to the built in diode potential causes electrons to be swept into the N-layer, holes into the P-layer. Actually electron-hole pairs may be formed in any of the semiconductor regions. However, those formed in the depletion region are most likely to be separated into the respective N and P-regions. Many of the electron-hole pairs formed in the P and N-regions recombine. Only a few do so in the depletion region. Thus, a few electron-hole pairs in the N and P-regions, and most in the depletion region contribute to *photocurrent*, that current resulting from light falling on the photodiode.

The voltage out of a photodiode may be observed. Operation in this *photovoltaic* (PV) mode is not linear over a large dynamic range, though it is sensitive and has low noise at frequencies less than 100 kHz. The preferred mode of operation is often *photocurrent (PC)* mode because the current is linearly proportional to light flux over several decades of intensity, and higher frequency response can be achieved. PC mode is achieved with reverse bias or zero bias on the photodiode. A current amplifier (transimpedance amplifier) should be used with a photodiode in PC mode. Linearity and PC mode are achieved as long as the diode does not become forward biased.

High speed operation is often required of photodiodes, as opposed to solar cells. Speed is a function of diode capacitance, which can be minimized by decreasing cell area. Thus, a

sensor for a high speed fiber optic link will use an area no larger than necessary, say 1 mm^2. Capacitance may also be decreased by increasing the thickness of the depletion region, in the manufacturing process or by increasing the reverse bias on the diode.

PIN diode The *p-i-n diode* or *PIN diode* is a photodiode with an intrinsic layer between the P and N-regions as in Figure 3.75. The **P-I**ntrinsic-**N** structure increases the distance between the P and N conductive layers, decreasing capacitance, increasing speed. The volume of the photo sensitive region also increases, enhancing conversion efficiency. The bandwidth can extend to 10's of gHz. PIN photodiodes are the preferred for high sensitivity, and high speed at moderate cost.

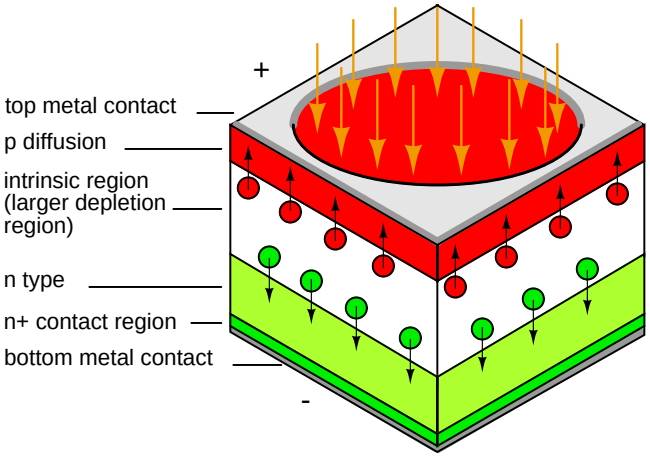


Figure 3.75: *PIN photodiode: The intrinsic region increases the thickness of the depletion region.*

Avalanche photo diode: An *avalanche photodiode (APD)*designed to operate at high reverse bias exhibits an electron multiplier effect analogous to a photomultiplier tube. The reverse bias can run from 10's of volts to nearly 2000 V. The high level of reverse bias accelerates photon created electron-hole pairs in the intrinsic region to a high enough velocity to free additional carriers from collisions with the crystal lattice. Thus, many electrons per photon result. The motivation for the APD is to achieve amplification within the photodiode to overcome noise in external amplifiers. This works to some extent. However, the APD creates noise of its own. At high speed the APD is superior to a PIN diode amplifier combination, though not for low speed applications. APD's are expensive, roughly the price of a photomultiplier tube. So, they are only competitive with PIN photodiodes for niche applications. One such application is single photon counting as applied to nuclear physics.

3.12.6 Solar cells

A photodiode optimized for efficiently delivering power to a load is the *solar cell*. It operates in photovoltaic mode (PV) because it is forward biased by the voltage developed across the load resistance.

Monocrystalline solar cells are manufactured in a process similar to semiconductor processing. This involves growing a single crystal boule from molten high purity silicon (P-type), though, not as high purity as for semiconductors. The boule is diamond sawed or wire sawed into wafers. The ends of the boule must be discarded or recycled, and silicon is lost in the saw kerf. Since modern cells are nearly square, silicon is lost in squaring the boule. Cells may be etched to texture (roughen) the surface to help trap light within the cell. Considerable silicon is lost in producing the 10 or 15 cm square wafers. These days (2007) it is common for solar cell manufacturer to purchase the wafers at this stage from a supplier to the semiconductor industry.

P-type Wafers are loaded back-to-back into fused silica boats exposing only the outer surface to the N-type dopant in the diffusion furnace. The diffusion process forms a thin n-type layer on the top of the cell. The diffusion also shorts the edges of the cell front to back. The periphery must be removed by plasma etching to unshort the cell. Silver and or aluminum paste is screened on the back of the cell, and a silver grid on the front. These are sintered in a furnace for good electrical contact. (Figure 3.76)

The cells are wired in series with metal ribbons. For charging 12 V batteries, 36 cells at approximately 0.5 V are vacuum laminated between glass, and a polymer metal back. The glass may have a textured surface to help trap light.

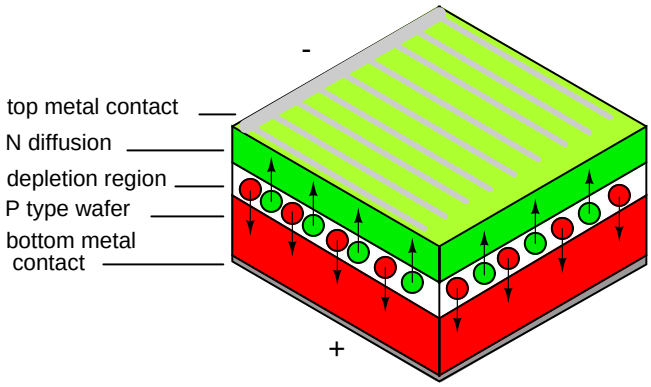


Figure 3.76: *Silicon Solar cell*

The ultimate commercial high efficiency (21.5%) single crystal silicon solar cells have all contacts on the back of the cell. The active area of the cell is increased by moving the top (-) contact conductors to the back of the cell. The top (-) contacts are normally made to the N-type silicon on top of the cell. In Figure 3.77 the (-) contacts are made to N^+ diffusions on the bottom interleaved with (+) contacts. The top surface is textured to aid in trapping light within the cell.. [18]

Multicyrstalline silicon cells start out as molten silicon cast into a rectangular mold. As the silicon cools, it crystallizes into a few large (mm to cm sized) randomly oriented crystals instead of a single one. The remainder of the process is the same as for single crystal cells. The finished cells show lines dividing the individual crystals, as if the cells were cracked. The high efficiency is not quite as high as single crystal cells due to losses at crystal grain boundaries.

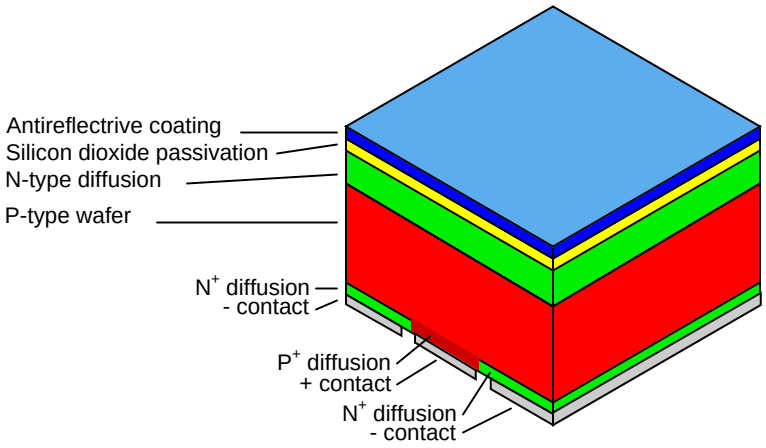


Figure 3.77: *High efficiency solar cell with all contacts on the back. Adapted from Figure 1 [18]*

The cell surface cannot be roughened by etching due to the random orientation of the crystals. However, an antireflectrive coating improves efficiency. These cells are competitive for all but space applications.

Three layer cell: The highest efficiency solar cell is a stack of three cells tuned to absorb different portions of the solar spectrum. Though three cells can be stacked atop one another, a monolithic single crystal structure of 20 semiconductor layers is more compact. At 32 % efficiency, it is now (2007) favored over silicon for space application. The high cost prevents it from finding many earth bound applications other than concentrators based on lenses or mirrors.

Intensive research has recently produced a version enhanced for terrestrial concentrators at 400 - 1000 suns and 40.7% efficiency. This requires either a big inexpensive Fresnel lens or reflector and a small area of the expensive semiconductor. This combination is thought to be competitive with inexpensive silicon cells for solar power plants. [9] [23]

Metal organic chemical vapor deposition (MOCVD) deposits the layers atop a P-type germanium substrate. The top layers of N and P-type gallium indium phosphide (GaInP) having a band gap of 1.85 eV, absorbs ultraviolet and visible light. These wavelengths have enough energy to exceed the band gap. Longer wavelengths (lower energy) do not have enough energy to create electron-hole pairs, and pass on through to the next layer. A gallium arsenide layers having a band gap of 1.42 eV, absorbs near infrared light. Finally the germanium layer and substrate absorb far infrared. The series of three cells produce a voltage which is the sum of the voltages of the three cells. The voltage developed by each material is 0.4 V less than the band gap energy listed in Table 3.4. For example, for GaInP: 1.8 eV/e - 0.4 V = 1.4 V. For all three the voltage is 1.4 V + 1.0 V + 0.3 V = 2.7 V. [4]

Crystalline solar cell arrays have a long usable life. Many arrays are guaranteed for 25 years, and believed to be good for 40 years. They do not suffer initial degradation compared with amorphous silicon.

Both single and multicrystalline solar cells are based on silicon wafers. The silicon is both the substrate and the active device layers. Much silicon is consumed. This kind of cell has

Table 3.4: *High efficiency triple layer solar cell.*

Layer	Band gap	Light absorbed
Gallium indium phosphide	1.8 eV	UV, visible
Gallium arsenide	1.4 eV	near infrared
Germanium	0.7 eV	far infrared

been around for decades, and takes approximately 86% of the solar electric market. For further information about crystalline solar cells see Honsberg. [8]

Amorphous silicon thin film solar cells use tiny amounts of the active raw material, silicon. Approximately half the cost of conventional crystalline solar cells is the solar cell grade silicon. The thin film deposition process reduces this cost. The downside is that efficiency is about half that of conventional crystalline cells. Moreover, efficiency degrades by 15-35% upon exposure to sunlight. A 7% efficient cell soon ages to 5% efficiency. Thin film amorphous silicon cells work better than crystalline cells in dim light. They are put to good use in solar powered calculators.

Non-silicon based solar cells make up about 7% of the market. These are thin-film polycrystalline products. Various compound semiconductors are the subject of research and development. Some non-silicon products are in production. Generally, the efficiency is better than amorphous silicon, but not nearly as good as crystalline silicon.

Cadmium telluride as a polycrystalline thin film on metal or glass can have a higher efficiency than amorphous silicon thin films. If deposited on metal, that layer is the negative contact to the cadmium telluride thin film. The transparent P-type cadmium sulfide atop the cadmium telluride serves as a buffer layer. The positive top contact is transparent, electrically conductive fluorine doped tin oxide. These layers may be laid down on a sacrificial foil in place of the glass in the process in the following pargraph. The sacrificial foil is removed after the cell is mounted to a permanent substrate.

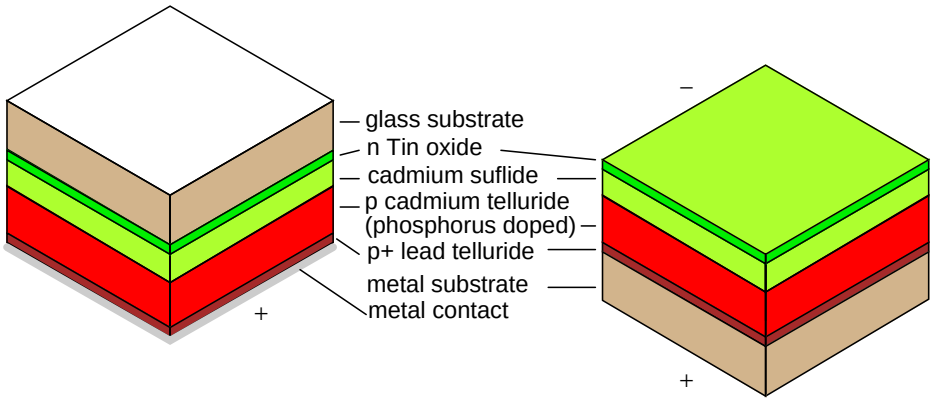


Figure 3.78: *Cadmium telluride solar cell on glass or metal.*

A process for depositing cadmium telluride on glass begins with the deposition of N-type

transparent, electrically conducive, tin oxide on a glass substrate. The next layer is P-type cadmium telluride; though, N-type or intrinsic may be used. These two layers constitute the NP junction. A P^+ (heavy P-type) layer of lead telluride aids in establishing a low resistance contact. A metal layer makes the final contact to the lead telluride. These layers may be laid down by vacuum deposition, chemical vapor deposition (CVD), screen printing, electrodeposition, or atmospheric pressure chemical vapor deposition (APCVD) in helium. [10]

A variation of cadmium telluride is mercury cadmium telluride. Having lower bulk resistance and lower contact resistance improves efficiency over cadmium telluride.

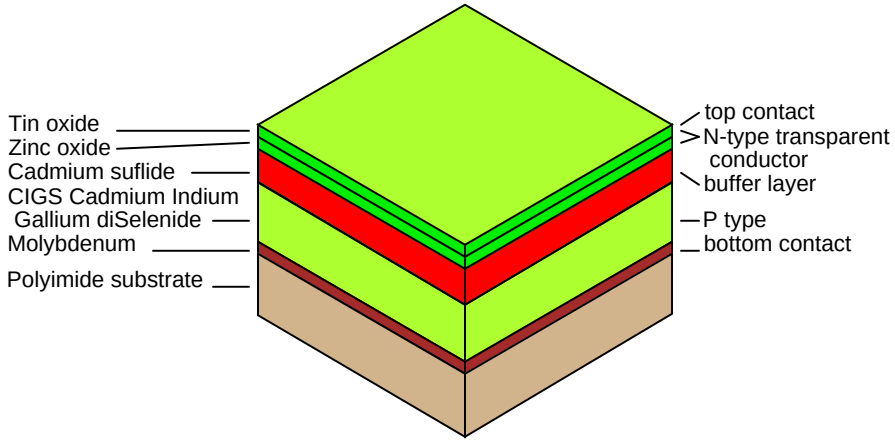


Figure 3.79: *Cadmium Indium Gallium diSelenide solar cell (CIGS)*

Cadmium Indium Gallium diSelenide: A most promising thin film solar cell at this time (2007) is manufactured on a ten inch wide roll of flexible polyimide– Cadmium Indium Gallium diSelenide (CIGS). It has a spectacular efficiency of 10%. Though, commercial grade crystalline silicon cells surpassed this decades ago, CIGS should be cost competitive. The deposition processes are at a low enough temperature to use a polyimide polymer as a substrate instead of metal or glass. (Figure 3.79) The CIGS is manufactured in a roll to roll process, which should drive down costs. GIGS cells may also be produced by an inherently low cost electrochemical process. [7]

- **REVIEW**:
- Most solar cells are silicon single crystal or multicrystal because of their good efficiency and moderate cost.
- Less efficient thin films of various amorphous or polycrystalline materials comprise the rest of the market.
- Table 3.5 compares selected solar cells.

Table 3.5: *Solar cell properties*

Solar cell type	**Maximum efficiency**	**Practical efficiency**	**Notes**
Selenium, polycrystalline	0.7%	-	1883, Charles Fritts
Silicon, single crystal	-	4%	1950's, first silicon solar cell
Silicon, single crystal PERL, terrestrial, space	25%	-	solar cars, cost=100x commercial
Silicon, single crystal, commercial terrestrial	24%	14-17%	$5-$10/peak watt
Cypress Semiconductor, Sunpower, silicon single crystal	21.5%	19%	all contacts on cell back
Gallium Indium Phosphide/ Gallium Arsenide/ Germanium, single crystal, multilayer	-	32%	Preferred for space.
Advanced terrestrial version of above.	-	40.7%	Uses optical concentrator.
Silicon, multicrystalline	18.5%	15.5%	-
Thin films,	-	-	-
Silicon, amorphous	13%	5-7%	Degrades in sun light. Good indoors for calculators or cloudy outdoors.
Cadmium telluride, polycrystalline	16%	-	glass or metal substrate
Copper indium arsenide diselenide, polycrystalline	18%	10%	10 inch flexible polymer web. [17]
Organic polymer, 100% plastic	4.5%	-	R&D project

3.12.7 Varicap or varactor diodes

A variable capacitance diode is known as a *varicap diode* or as a *varactor*. If a diode is reverse biased, an insulating depletion region forms between the two semiconductive layers. In many diodes the width of the depletion region may be changed by varying the reverse bias. This varies the capacitance. This effect is accentuated in varicap diodes. The schematic symbols is shown in Figure 3.80, one of which is packaged as common cathode dual diode.

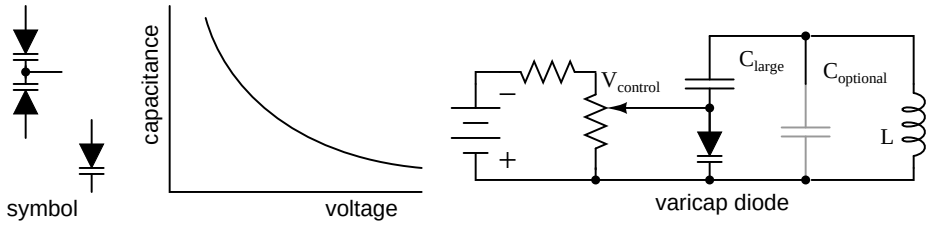


Figure 3.80: *Varicap diode: Capacitance varies with reverse bias. This varies the frequency of a resonant network.*

If a varicap diode is part of a resonant circuit as in Figure 3.80, the frequency may be varied with a control voltage, $V_{control}$. A large capacitance, low X_c, in series with the varicap prevents $V_{control}$ from being shorted out by inductor L. As long as the series capacitor is large, it has minimal effect on the frequency of resonant circuit. $C_{optional}$ may be used to set the center resonant frequency. $V_{control}$ can then vary the frequency about this point. Note that the required active circuitry to make the resonant network oscillate is not shown. For an example of a varicap diode tuned AM radio receiver see “electronic varicap diode tuning,” (page 438)

Some varicap diodes may be referred to as abrupt, hyperabrupt, or super hyper abrupt. These refer to the change in junction capacitance with changing reverse bias as being abrupt or hyper-abrupt, or super hyperabrupt. These diodes offer a relatively large change in capacitance. This is useful when oscillators or filters are swept over a large frequency range. Varying the bias of abrupt varicaps over the rated limits, changes capacitance by a 4:1 ratio, hyperabrupt by 10:1, super hyperabrupt by 20:1.

Varactor diodes may be used in frequency multiplier circuits. See “Practical analog semiconductor circuits,” Varactor multiplier

3.12.8 Snap diode

The *snap diode*, also known as the *step recovery diode* is designed for use in high ratio frequency multipliers up to 20 gHz. When the diode is forward biased, charge is stored in the PN junction. This charge is drawn out as the diode is reverse biased. The diode looks like a low impedance current source during forward bias. When reverse bias is applied it still looks like a low impedance source until all the charge is withdrawn. It then “snaps” to a high impedance state causing a voltage impulse, rich in harmonics. An applications is a comb generator, a generator of many harmonics. Moderate power 2x and 4x multipliers are another application.

3.12.9 PIN diodes

A *PIN diode* is a fast low capacitance switching diode. Do not confuse a PIN switching diode with a PIN photo diode (page 153). A PIN diode is manufactured like a silicon switching diode with an intrinsic region added between the PN junction layers. This yields a thicker depletion region, the insulating layer at the junction of a reverse biased diode. This results in lower capacitance than a reverse biased switching diode.

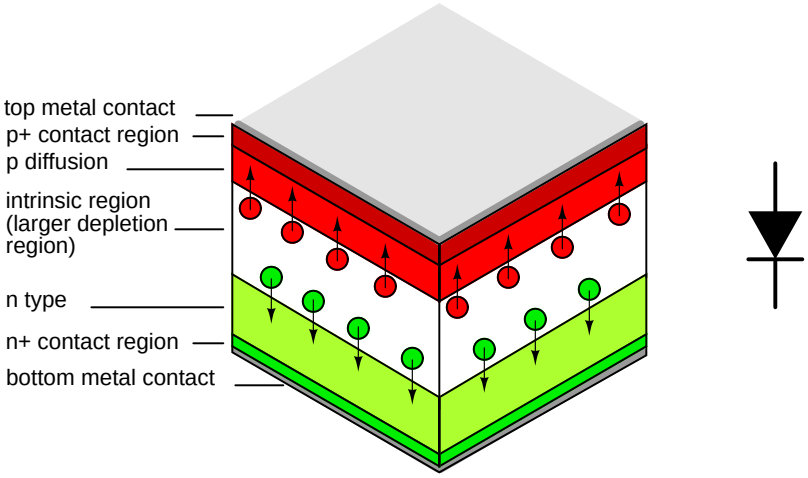


Figure 3.81: *Pin diode: Cross section aligned with schematic symbol.*

PIN diodes are used in place of switching diodes in radio frequency (RF) applications, for example, a T/R switch (page 441). The 1n4007 1000 V, 1 A general purpose power diode is reported to be usable as a PIN switching diode. The high voltage rating of this diode is achieved by the inclusion of an intrinsic layer dividing the PN junction. This intrinsic layer makes the 1n4007 a PIN diode. Another PIN diode application is as the antenna switch (page 441) for a direction finder receiver.

PIN diodes serve as variable resistors when the forward bias is varied. One such application is the voltage variable attenuator (page 441). The low capacitance characteristic of PIN diodes, extends the frequency flat response of the attenuator to microwave frequencies.

3.12.10 IMPATT diode

¡The *IMPact Avalanche Transit Time* diode is a high power radio frequency (RF) generator operating from 3 to 100 gHz. IMPATT diodes are fabricated from silicon, gallium arsenide, or silicon carbide.

An IMPATT diode is reverse biased above the breakdown voltage. The high doping levels produce a thin depletion region. The resulting high electric field rapidly accelerates carriers which free other carriers in collisions with the crystal lattice. Holes are swept into the P_+ region. Electrons drift toward the N regions. The cascading effect creates an avalanche current which increases even as voltage across the junction decreases. The pulses of current lag the

voltage peak across the junction. A “negative resistance” effect in conjunction with a resonant circuit produces oscillations at high power levels (high for semiconductors).

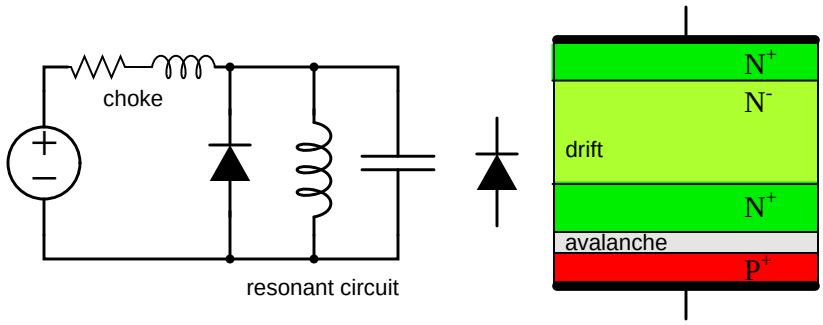


Figure 3.82: *IMPATT diode: Oscillator circuit and heavily doped P and N layers.*

The resonant circuit in the schematic diagram of Figure 3.82 is the lumped circuit equivalent of a waveguide section, where the IMPATT diode is mounted. DC reverse bias is applied through a choke which keeps RF from being lost in the bias supply. This may be a section of waveguide known as a bias Tee. Low power RADAR transmitters may use an IMPATT diode as a power source. They are too noisy for use in the receiver. [20]

3.12.11 Gunn diode

Diode, gunn Gunn diode

A *gunn diode* is solely composed of N-type semiconductor. As such, it is not a true diode. Figure 3.83 shows a lightly doped N^- layer surrounded by heavily doped N+ layers. A voltage applied across the N-type gallium arsenide gunn diode creates a strong electric field across the lightly doped N^- layer.

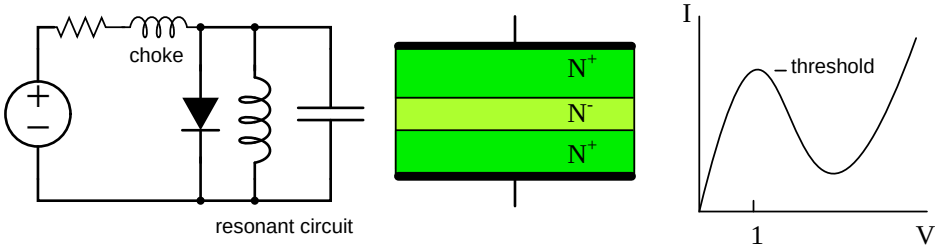


Figure 3.83: *Gunn diode: Oscillator circuit and cross section of only N-type semiconductor diode.*

As voltage is increased, conduction increases due to electrons in a low energy conduction band. As voltage is increased beyond the threshold of approximately 1 V, electrons move from the lower conduction band to the higher energy conduction band where they no longer con-

tribute to conduction. In other words, as voltage increases, current decreases, a negative resistance condition. The oscillation frequency is determined by the transit time of the conduction electrons, which is inversely related to the thickness of the N^- layer.

The frequency may be controlled to some extent by embedding the gunn diode into a resonant circuit. The lumped circuit equivalent shown in Figure 3.83 is actually a coaxial transmission line or waveguide. Gallium arsenide gunn diodes are available for operation from 10 to 200 gHz at 5 to 65 mw power. Gunn diodes may also serve as amplifiers. [19] [14]

3.12.12 Shockley diode

The *Shockley diode*is a 4-layer thyristor used to trigger larger thyristors. It only conducts in one direction when triggered by a voltage exceeding the *breakover voltage*, about 20 V. See "Thyristors," The Shockley Diode. The bidirectional version is called a *diac*. See "Thyristors," The DIAC.

3.12.13 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. The constant current diode is a two terminal version of a JFET. If we 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 circuit in Figure 3.84(a) and plot diode current against diode voltage, we'd get a graph that rises at first and then levels off at the current regulation point as in Figure 3.84(b).

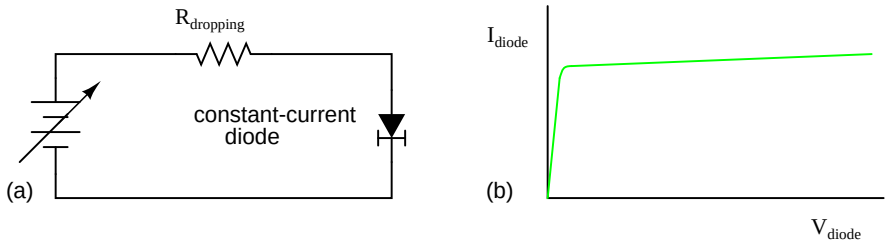


Figure 3.84: *Constant current diode: (a) Test circuit, (b) current vs voltage characteristic.*

One 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 as in Figure 3.85.

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 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.

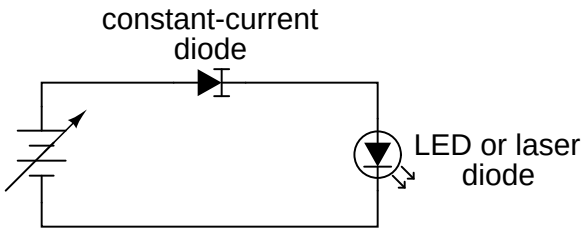


Figure 3.85: *Constant current diode application: driving laser diode.*

3.13 Other diode technologies

3.13.1 SiC diodes

Diodes manufactured from silicon carbide are capable of high temperature operation to 400^{o}C. This could be in a high temperature environment: down hole oil well logging, gas turbine engines, auto engines. Or, operation in a moderate environment at high power dissipation. Nuclear and space applications are promising as SiC is 100 times more resistant to radiation compared with silicon. SiC is a better conductor of heat than any metal. Thus, SiC is better than silicon at conducting away heat. Breakdown voltage is several kV. SiC power devices are expected to reduce electrical energy losses in the power industry by a factor of 100.

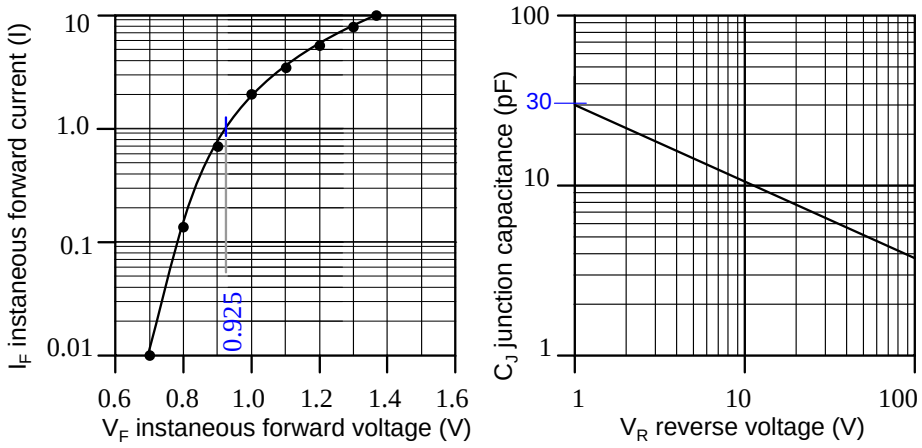
3.13.2 Polymer diode

Diodes based on organic chemicals have been produced using low temperature processes. Hole rich and electron rich conductive polymers may be ink jet printed in layers. Most of the research and development is of the *organic LED* (OLED). However, development of inexpensive printable organic RFID (radio frequency identification) tags is on going. In this effort, a pentacene organic rectifier has been operated at 50 MHz. Rectification to 800 MHz is a development goal. An inexpensive *metal insulator metal* (MIM) diode acting like a back-to-back zener diode clipper has been delveloped. Also, a tunnel diode like device has been fabricated.

3.14 SPICE models

The SPICE circuit simulation program provides for modeling diodes in circuit simulations. The diode model is based on characterization of individual devices as described in a product data sheet and manufacturing process characteristics not listed. Some information has been extracted from a 1N4004 data sheet in Figure 3.86.

The diode statement begins with a diode element name which must begin with “d” plus optional characters. Example diode element names include: d1, d2, dtest, da, db, d101. Two node numbers specify the connection of the anode and cathode, respectively, to other components. The node numbers are followed by a model name, referring to a subsequent “.model” statement.



Max avg rectified current I_O (A)	1	Forward voltage drop V_F (V)	1
Peak repetitive reverse voltage V_{RRM} (V)	400	@ I_F (A)	1
Peak forward surge current I_{FSM} (A)	30	Max reverse current I_R (µA)	5
Total capacitance C_T (pF)	15	@ V_R (V)	400

Figure 3.86: *Data sheet 1N4004 excerpt, after [6].*

The model statement line begins with ".model," followed by the model name matching one or more diode statements. Next, a "d" indicates a diode is being modeled. The remainder of the model statement is a list of optional diode parameters of the form Parameter-Name=ParameterValue. None are used in Example below. Example2 has some parameters defined. For a list of diode parameters, see Table 3.6.

```
 General form:  d[name] [anode] [cathode] [modelname]
                .model ([modelname] d [parmtr1=x] [parmtr2=y] . . .)

 Example:       d1 1 2 mod1
                 .model mod1 d

 Example2:      D2 1 2 Da1N4004
                 .model Da1N4004 D (IS=18.8n RS=0 BV=400 IBV=5.00u CJO=30
M=0.333 N=2)
```

The easiest approach to take for a SPICE model is the same as for a data sheet: consult the manufacturer's web site. Table 3.7 lists the model parameters for some selected diodes. A fallback strategy is to build a SPICE model from those parameters listed on the data sheet. A third strategy, not considered here, is to take measurements of an actual device. Then, calculate, compare and adjust the SPICE parameters to the measurements.

If diode parameters are not specified as in "Example" model above, the parameters take on

Table 3.6: *Diode SPICE parameters*

Symbol	**Name**	**Parameter**	**Units**	**Default**
I_S	IS	Saturation current (diode equation)	A	1E-14
R_S	RS	Parsitic resistance (series resistance)	Ω	0
n	N	Emission coefficient, 1 to 2	-	1
τ_D	TT	Transit time	s	0
$C_D(0)$	CJO	Zero-bias junction capacitance	F	0
ϕ_0	VJ	Junction potential	V	1
m	M	Junction grading coefficient	-	0.5
-	-	0.33 for linearly graded junction	-	-
-	-	0.5 for abrupt junction	-	-
E_g	EG	Activation energy:	eV	1.11
-	-	Si: 1.11	-	-
-	-	Ge: 0.67	-	-
-	-	Schottky: 0.69	-	-
p_i	XTI	IS temperature exponent	-	3.0
-	-	pn junction: 3.0	-	-
-	-	Schottky: 2.0	-	-
k_f	KF	Flicker noise coefficient	-	0
a_f	AF	Flicker noise exponent	-	1
FC	FC	Forward bias depletion capacitance coefficient	-	0.5
BV	BV	Reverse breakdown voltage	V	∞
IBV	IBV	Reverse breakdown current	A	1E-3

the default values listed in Table 3.6 and Table 3.7. These defaults model integrated circuit diodes. These are certainly adequate for preliminary work with discrete devices For more critical work, use SPICE models supplied by the manufacturer [5], SPICE vendors, and other sources. [16]

Table 3.7: *SPICE parameters for selected diodes; sk=schottky Ge=germanium; else silicon.*

Part	**IS**	**RS**	**N**	**TT**	**CJO**	**M**	**VJ**	**EG**	**XTI**	**BV**	**IBV**
Default	1E-14	0	1	0	0	0.5	1	1.11	3	∞	1m
1N5711 sk	315n	2.8	2.03	1.44n	2.00p	0.333	-	0.69	2	70	10u
1N5712 sk	680p	12	1.003	50p	1.0p	0.5	0.6	0.69	2	20	-
1N34 Ge	200p	84m	2.19	144n	4.82p	0.333	0.75	0.67	-	60	15u
1N4148	35p	64m	1.24	5.0n	4.0p	0.285	0.6	-	-	75	-
1N3891	63n	9.6m	2	110n	114p	0.255	0.6	-	-	250	-
10A04 10A	844n	2.06m	2.06	4.32u	277p	0.333	-	-	-	400	10u
1N4004 1A	76.9n	42.2m	1.45	4.32u	39.8p	0.333	-	-	-	400	5u
1N4004 data sheet	18.8n	-	2	-	30p	0.333	-	-	-	400	5u

Otherwise, derive some of the parameters from the data sheet. First select a value for spice parameter N between 1 and 2. It is required for the diode equation (n). Massobrio [1] pp 9, recommends ".. n, the emission coefficient is usually about 2." In Table 3.7, we see that power rectifiers 1N3891 (12 A), and 10A04 (10 A) both use about 2. The first four in the table are not relevant because they are schottky, schottky, germanium, and silicon small signal, respectively. The saturation current, IS, is derived from the diode equation, a value of (V_D, I_D) on the graph in Figure 3.86, and N=2 (n in the diode equation).

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

V_T = 26 mV at 25^{o}C n = 2.0 V_D = 0.925 V at 1 A from graph

1 A = $I_S(e^{(0.925V)/(2)(26mV)} - 1)$

I_S = 18.8E-9

The numerical values of IS=18.8n and N=2 are entered in last line of Table 3.7 for comparison to the manufacturers model for 1N4004, which is considerably different. RS defaults to 0 for now. It will be estimated later. The important DC static parameters are N, IS, and RS.

Rashid [15] suggests that TT, τ_D, the transit time, be approximated from the reverse recovery stored charge Q_{RR}, a data sheet parameter (not available on our data sheet) and I_F, forward current.

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

$$\tau_D = Q_{RR}/I_F$$

We take the TT=0 default for lack of Q_{RR}. Though it would be reasonable to take TT for a similar rectifier like the 10A04 at 4.32u. The 1N3891 TT is not a valid choice because it is a fast recovery rectifier. CJO, the zero bias junction capacitance is estimated from the V_R vs C_J graph in Figure 3.86. The capacitance at the nearest to zero voltage on the graph is 30 pF at 1 V. If simulating high speed transient response, as in switching regulator power supplies, TT and CJO parameters must be provided.

The junction grading coefficient M is related to the doping profile of the junction. This is not a data sheet item. The default is 0.5 for an abrupt junction. We opt for M=0.333 corresponding to a linearly graded junction. The power rectifiers in Table 3.7 use lower values for M than 0.5.

We take the default values for VJ and EG. Many more diodes use VJ=0.6 than shown in Table 3.7. However the 10A04 rectifier uses the default, which we use for our 1N4004 model (Da1N4001 in Table 3.6). Use the default EG=1.11 for silicon diodes and rectifiers. Table 3.6 lists values for schottky and germanium diodes. Take the XTI=3, the default IS temperature coefficient for silicon devices. See Table 3.6 for XTI for schottky diodes.

The abbreviated data sheet, Figure 3.86, lists I_R = 5 μA @ V_R = 400 V, corresponding to IBV=5u and BV=400 respectively. The 1n4004 SPICE parameters derived from the data sheet are listed in the last line of Table 3.7 for comparison to the manufacturer's model listed above it. BV is only necessary if the simulation exceeds the reverse breakdown voltage of the diode, as is the case for zener diodes. IBV, reverse breakdown current, is frequently omitted, but may be entered if provided with BV.

Figure 3.87 shows a circuit to compare the manufacturers model, the model derived from the datasheet, and the default model using default parameters. The three dummy 0 V sources are necessary for diode current measurement. The 1 V source is swept from 0 to 1.4 V in 0.2 mV steps. See .DC statement in the netlist in Table 3.8. DI1N4004 is the manufacturer's diode model, Da1N4004 is our derived diode model.

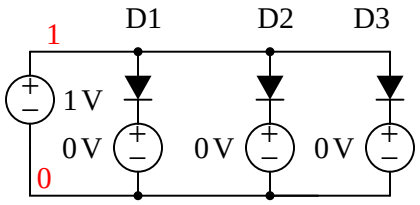


Figure 3.87: *SPICE circuit for comparison of manufacturer model (D1), calculated datasheet model (D2), and default model (D3).*

We compare the three models in Figure 3.88. and to the datasheet graph data in Table 3.9. VD is the diode voltage versus the diode currents for the manufacturer's model, our calculated datasheet model and the default diode model. The last column "1N4004 graph" is from the datasheet voltage versus current curve in Figure 3.86 which we attempt to match. Comparison of the currents for the three model to the last column shows that the default model is good at low currents, the manufacturer's model is good at high currents, and our calculated datasheet model is best of all up to 1 A. Agreement is almost perfect at 1 A because the IS calculation is

Table 3.8: *SPICE netlist parameters: (D1) DI1N4004 manufacturer's model, (D2) Da1N40004 datasheet derived, (D3) default diode model.*

```
*SPICE circuit <03468.eps> from XCircuit v3.20
D1 1 5 DI1N4004
V1 5 0 0
D2 1 3 Da1N4004
V2 3 0 0
D3 1 4 Default
V3 4 0 0
V4 1 0 1
.DC V4 0 1400mV 0.2m
.model Da1N4004 D (IS=18.8n RS=0      BV=400 IBV=5.00u CJO=30
+M=0.333   N=2.0  TT=0)
.MODEL DI1N4004 D (IS=76.9n RS=42.0m  BV=400 IBV=5.00u CJO=39.8p
+M=0.333 N=1.45 TT=4.32u)
.MODEL Default D
.end
```

based on diode voltage at 1 A. Our model grossly over states current above 1 A.

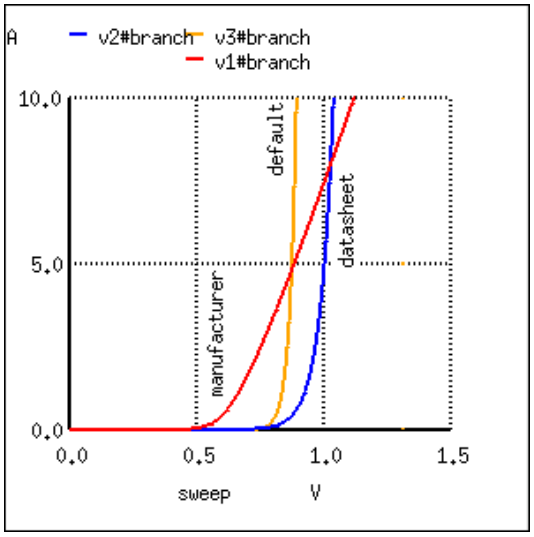


Figure 3.88: *First trial of manufacturer model, calculated datasheet model, and default model.*

The solution is to increase RS from the default RS=0. Changing RS from 0 to 8m in the datasheet model causes the curve to intersect 10 A (not shown) at the same voltage as the manufacturer's model. Increasing RS to 28.6m shifts the curve further to the right as shown in Figure 3.89. This has the effect of more closely matching our datasheet model to the datasheet graph (Figure 3.86). Table 3.10 shows that the current 1.224470e+01 A at 1.4 V matches the

Table 3.9: *Comparison of manufacturer model, calculated datasheet model, and default model to 1N4004 datasheet graph of V vs I.*

index	VD	model manufacturer	model datasheet	model default	1N4004 graph
3500	7.000000e-01	1.612924e+00	1.416211e-02	5.674683e-03	0.01
4001	8.002000e-01	3.346832e+00	9.825960e-02	2.731709e-01	0.13
4500	9.000000e-01	5.310740e+00	6.764928e-01	1.294824e+01	0.7
4625	9.250000e-01	5.823654e+00	1.096870e+00	3.404037e+01	1.0
5000	1.000000e-00	7.395953e+00	4.675526e+00	6.185078e+02	2.0
5500	1.100000e+00	9.548779e+00	3.231452e+01	2.954471e+04	3.3
6000	1.200000e+00	1.174489e+01	2.233392e+02	1.411283e+06	5.3
6500	1.300000e+00	1.397087e+01	1.543591e+03	6.741379e+07	8.0
7000	1.400000e+00	1.621861e+01	1.066840e+04	3.220203e+09	12.0

graph at 12 A. However, the current at 0.925 V has degraded from 1.096870e+00 above to 7.318536e-01.

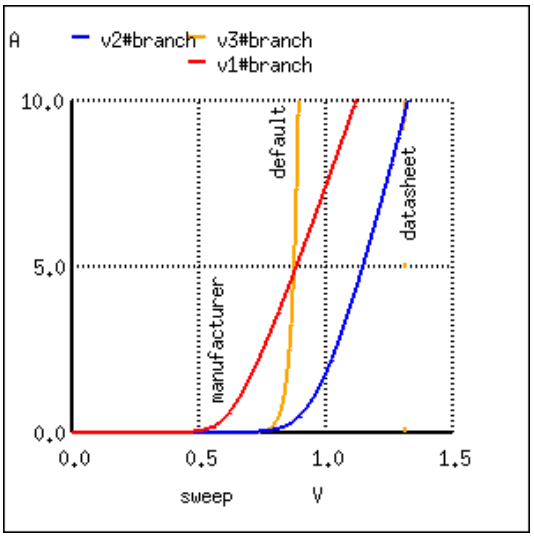


Figure 3.89: *Second trial to improve calculated datasheet model compared with manufacturer model and default model.*

Suggested reader exercise: decrease N so that the current at VD=0.925 V is restored to 1 A. This may increase the current (12.2 A) at VD=1.4 V requiring an increase of RS to decrease current to 12 A.

Zener diode: There are two approaches to modeling a zener diode: set the BV parameter to the zener voltage in the model statement, or model the zener with a subcircuit containing a diode clamper set to the zener voltage. An example of the first approach sets the breakdown

Table 3.10: *Changing Da1N4004 model statement RS=0 to RS=28.6m decreases the current at VD=1.4 V to 12.2 A.*

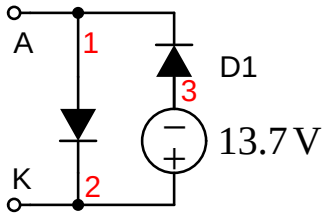
```
.model Da1N4004 D (IS=18.8n RS=28.6m  BV=400 IBV=5.00u CJO=30
+M=0.333   N=2.0  TT=0)
```

index	VD	model manufacturer	model datasheet	1N4001 graph
3505	7.010000e-01	1.628276e+00	1.432463e-02	0.01
4000	8.000000e-01	3.343072e+00	9.297594e-02	0.13
4500	9.000000e-01	5.310740e+00	5.102139e-01	0.7
4625	9.250000e-01	5.823654e+00	7.318536e-01	1.0
5000	1.000000e-00	7.395953e+00	1.763520e+00	2.0
5500	1.100000e+00	9.548779e+00	3.848553e+00	3.3
6000	1.200000e+00	1.174489e+01	6.419621e+00	5.3
6500	1.300000e+00	1.397087e+01	9.254581e+00	8.0
7000	1.400000e+00	1.621861e+01	1.224470e+01	12.

voltage BV to 15 for the 1n4469 15 V zener diode model (IBV optional):

```
.model D1N4469 D ( BV=15 IBV=17m )
```

The second approach models the zener with a subcircuit. Clamper D1 and VZ in Figure 3.90 models the 15 V reverse breakdown voltage of a 1N4477A zener diode. Diode DR accounts for the forward conduction of the zener in the subcircuit.




```
.SUBCKT DI-1N4744A
1 2
* Terminals A K
D1 1 2 DF
DZ 3 1 DR
VZ 2 3 13.7
.MODEL DF D (
IS=27.5p RS=0.620
N=1.10
+ CJO=78.3p
VJ=1.00 M=0.330
TT=50.1n )
.MODEL DR D (
IS=5.49f RS=0.804
N=1.77 )
.ENDS
```

Figure 3.90: *Zener diode subcircuit uses clamper (D1 and VZ) to model zener.*

Tunnel diode: A tunnel diode may be modeled by a pair of field effect transistors (JFET) in a SPICE subcircuit. [11] An oscillator circuit is also shown in this reference.

Gunn diode: A Gunn diode may also be modeled by a pair of JFET's. [12] This reference shows a microwave relaxation oscillator.

- **REVIEW:**
- Diodes are described in SPICE by a diode component statement referring to .model statement. The .model statement contains parameters describing the diode. If parameters are not provided, the model takes on default values.
- Static DC parameters include N, IS, and RS. Reverse breakdown parameters: BV, IBV.
- Accurate dynamic timing requires TT and CJO parameters
- Models provided by the manufacturer are highly recommended.

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.

Bibliography

[1] Paolo Antognetti, Giuseppe Massobrio "Semiconductor Device Modeling with SPICE," ISBN 0-07-002107-4, 1988

[2] ATCO Newsletter, Volume 14 No. 1, January 1997 at http://www.atco.tv/homepage/vol14_1.pdf

[3] D.A. Brunner, et al,, "A Cockcroft-Walton Base for the FEU84-3 Photomultiplier Tube," Department of Physics, Indiana University, Bloomington, Indiana 47405 January 1998, at http://dustbunny.physics.indiana.edu/~paul/cwbase/

[4] Brenton Burnet, "The Basic Physics and Design of III-V Multijunction Solar," NREL, at photochemistry.epfl.ch/EDEY/NREL.pdf

[5] Diodes Incorporated http://www.diodes.com/products/spicemodels/index.php

[6] Diodes Incorporated, "1N4001/L - 1N4007/l, 1.0A rectifier," at http://www.diodes.com/datasheets/ds28002.pdf

[7] "Solar firm gains $30 million in funding," EE Times, 07/12/2007 at http://www.eetimes.com/news/latest/showArticle.jhtml?articleID=201001129

[8] Christiana Honsberg, Stuart Bowden, "Photovoltaics CDROM," at http://www.udel.edu/igert/pvcdrom/

[9] R. R. King, et. al., "40% efficient metamorphic GaInP/GaInAs/Ge multijunction solar cells", Applied Physics Letters, 90, 183516 (2007) , at http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=APPLAB000090000018183516000001&idtype=cvips&gifs=yes

[10] Kim W Mitchell, "Method of making a thin film cadmium telluride solar cell," United States Patent 4734381,http://www.freepatentsonline.com/4734381.html

[11] Karl H. Muller "RF/Microwave Analysis" Intusoft Newsletter #51, November 1997, at http://www.intusoft.com/nlhtm/nl51.htm

[12] "A Gunn Diode Relaxation Oscillator," Intusoft Newsletter #52, February 1998, at http://www.intusoft.com/nlhtm/nl52.htm

[13] OAK Solar., "Technical LED's LED color chart," at http://www.oksolar.com/led/led_color_chart.htm

[14] Ian Poole, "Summary of the Gunn Diode," at http://www.radio-electronics.com/nfo/data/semicond/gunndiode/gunndiode.php

[15] Muhammad H. Rashid, "SPICE for Power Electronics and Electric Power," ISBN 0-13-030420-4, 1993

[16] "SPICE model index," V2.16 30-Nov-05, at http://homepages.which.net/~paul.hills/Circuits/Spice/ModelIndex.html

[17] Neil Thomas, "Advancing CIGS Solar Cell Manufacturing Technology," April 6, 2007 at http://www.renewableenergyaccess.com/rea/news/story?id=48033&src=rss

[18] P.J. Verlinden, Sinton, K. Wickham, R.M. Swanson Crane, "BACKSIDE-CONTACT SILICON SOLAR CELLS WITH IMPROVED EFFICIENCY." at http://www.sunpowercorp.com/techpapers/EPSEC97.pdf

[19] Christian Wolff, "Radar Principles," Radar components, Gunn diodes at at http://www.radartutorial.eu/17.bauteile/bt12.en.htm

[20] L. Yuan, M. R. Melloch, J. A. Cooper, K. J. Webb,"Silicon Carbide IMPATT Oscillators for High-Power Microwave and Millimeter-Wave Generation," IEEE/Cornell Conference on Advanced Concepts in High Speed Semiconductor Devices and Circuits, Ithaca, NY, August 7-9, 2000. at http://www.ecn.purdue.edu/WBG/Device_Research/IMPATT_Diodes/Index.html

[21] Alan Seabaugh, Zhaoming HU, Qingmin LIU, David Rink, Jinli Wang, "Silicon Based Tunnel Diodes and Integrated Circuits," at http://www.nd.edu/~nano/0a1003QFDpaper_v1.pdf

[22] S. M. Sze, G. Gibbons, "Avalanche breakdown voltages of abrupt and linearly graded p-n junctions in Ge, Si, GaAs, and Ga P," Appl. Phys. Lett., 8, 111 (1966).

[23] Lisa Zyga, "40% efficient solar cells to be used for solar electricity", PhysOrgForum, at http://www.physorg.com/news99904887.html

Chapter 4

BIPOLAR JUNCTION TRANSISTORS

Contents

4.1 Introduction . . . 174
4.2 The transistor as a switch . . . 176
4.3 Meter check of a transistor . . . 179
4.4 Active mode operation . . . 183
4.5 The common-emitter amplifier . . . 189
4.6 The common-collector amplifier . . . 202
4.7 The common-base amplifier . . . 210
4.8 The cascode amplifier . . . 218
4.9 Biasing techniques . . . 222
4.10 Biasing calculations . . . 235
4.10.1 Base Bias . . . 235
4.10.2 Collector-feedback bias . . . 236
4.10.3 Emitter-bias . . . 237
4.10.4 Voltage divider bias . . . 243
4.11 Input and output coupling . . . 247
4.12 Feedback . . . 256
4.13 Amplifier impedances . . . 263
4.14 Current mirrors . . . 264
4.15 Transistor ratings and packages . . . 269
4.16 BJT quirks . . . 271
4.16.1 Nonlinearity . . . 272
4.16.2 Temperature drift . . . 272
4.16.3 Thermal runaway . . . 273
4.16.4 Junction capacitance . . . 274
4.16.5 Noise . . . 274

4.16.6 Thermal mismatch (problem with paralleling transistors) 275
4.16.7 High frequency effects 276
Bibliography 277

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 in Figure 4.1(b) or N-P-N at (d). Each layer forming the transistor has a specific name, and each layer is provided with a wire contact for connection to a circuit. The schematic symbols are shown in Figure 4.1(a) and (d).

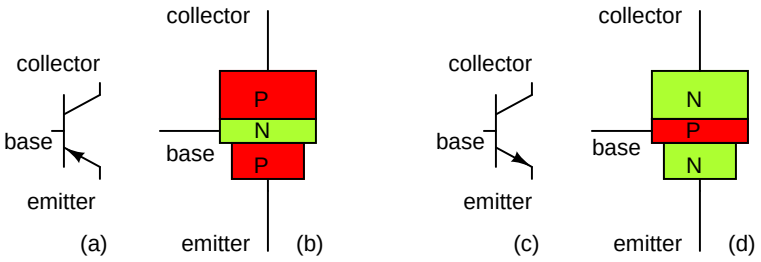


Figure 4.1: *BJT transistor: (a) PNP schematic symbol, (b) physical layout (c) NPN symbol, (d) layout.*

The 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 kind of transistor are exactly opposite each other.

Bipolar transistors work as current-controlled current *regulators*. In other words, transistors restrict the amount of current passed 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 kind of transistor it is (PNP or NPN, respectively). According to the standards of semiconductor symbology, the arrow always points *against* the direction of electron flow. (Figure 4.2)

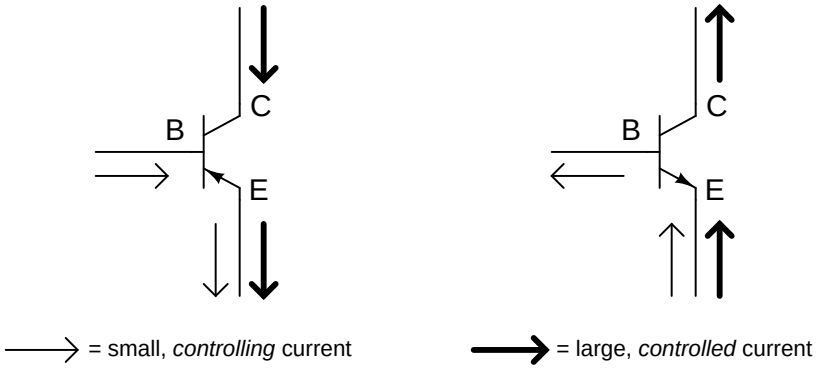


Figure 4.2: *Small Base-Emitter current controls large Collector-Emitter current flowing against emitter arrow.*

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.

No current through the base of the transistor, shuts it off like an open switch and prevents current through the collector. A base current, turns the transistor 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 with a switch. Such a circuit would be extremely simple as in Figure 4.3(a).

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 is 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 were. 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 as in Figure 4.3(b).

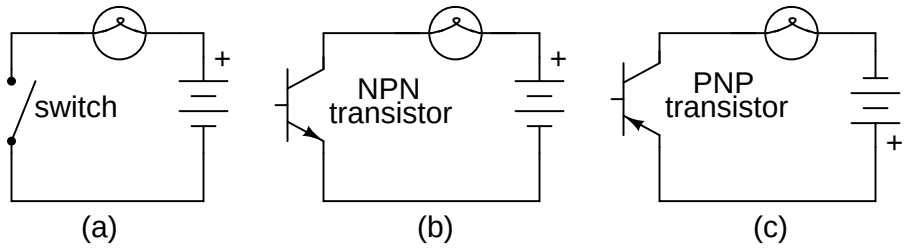


Figure 4.3: *(a) mechanical switch, (b) NPN transistor switch, (c) PNP transistor switch.*

A PNP transistor could also have been chosen for the job. Its application is shown in Figure 4.3(c).

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 as in Figure 4.4 (a).

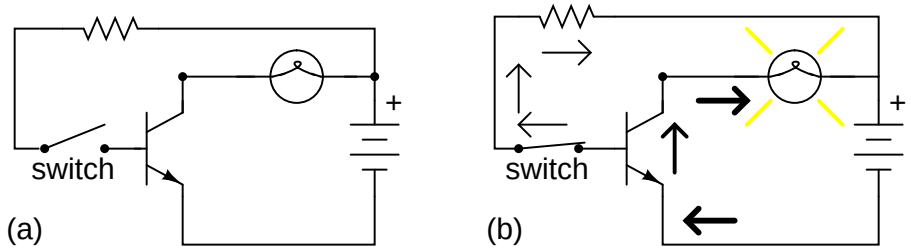


Figure 4.4: *Transistor: (a) cutoff, lamp off; (b) saturated, lamp on.*

If the switch is open as in Figure 4.4 (a), 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 as in Figure 4.4 (b), electrons will be able to flow from the emitter through to the base of the transistor, through the switch, 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?

Two points can 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; the transistor itself handles most 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. More importantly, the current-controlling behavior of the transistor enables us to use something completely different to turn the lamp on or off. Consider Figure 4.5, where a pair of solar cells provides 1 V to overcome the 0.7 V_{BE} of the transistor to cause base current flow, which in turn controls the lamp.

Or, we could use a thermocouple (many connected in series) to provide the necessary base current to turn the transistor on in Figure 4.6.

Even a microphone (Figure 4.7) with enough voltage and current (from an amplifier) output could 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 only need be a fraction of the current

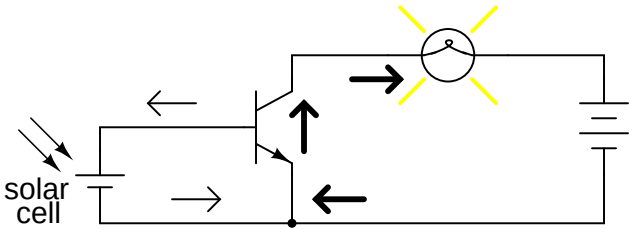


Figure 4.5: *Solar cell serves as light sensor.*

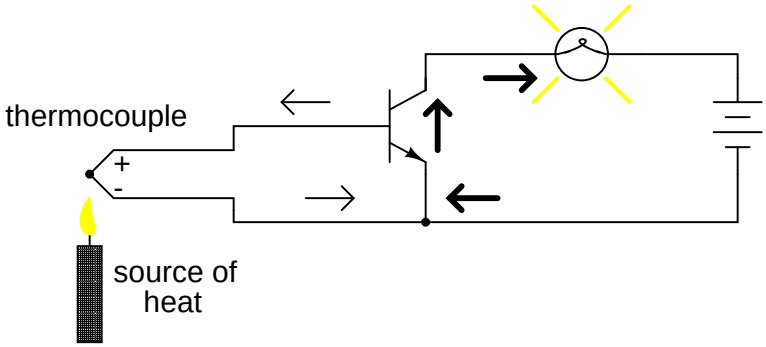


Figure 4.6: *A single thermocouple provides less than 40 mV. Many in series could produce in excess of the 0.7 V transistor V_{BE} to cause base current flow and consequent collector current to the lamp.*

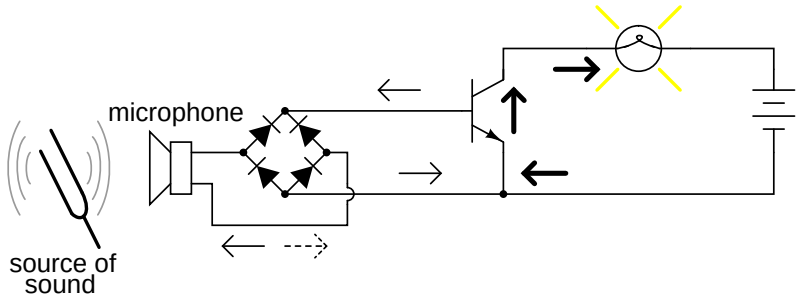


Figure 4.7: *Amplified microphone signal is rectified to DC to bias the base of the transistor providing a larger collector current.*

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; 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, transistors register as two diodes connected back-to-back when tested with a multimeter's "resistance" or "diode check" function as illustrated in Figure below. Low resistance readings on the base with the black negative (-) leads correspond to an N-type material in the base of a PNP transistor. On the symbol, the N-type material is "pointed" to by the arrow of the base-emitter junction, which is the base for this example. The P-type emitter corresponds to the other end of the arrow of the base-emitter junction, the emitter. The collector is very similar to the emitter, and is also a P-type material of the PN junction.

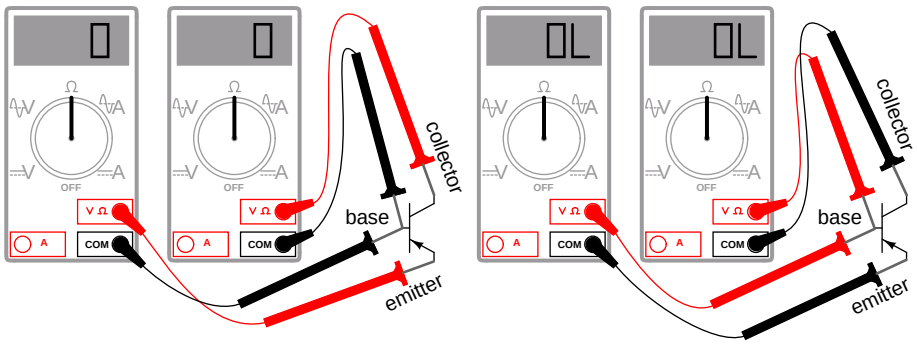


Figure 4.8: *PNP transistor meter check: (a) forward B-E, B-C, resistance is low; (b) reverse B-E, B-C, resistance is* ∞.

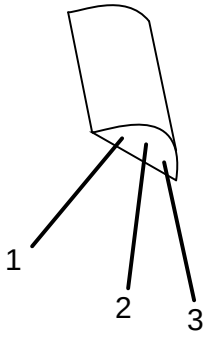
Here I'm assuming the use of a multimeter with only a single continuity range (resistance) function to check the PN junctions. Some multimeters are equipped with two separate continuity check functions: resistance and "diode check," each with its own purpose. If your meter has a designated "diode check" 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. Low resistance readings with the red (+) lead on the base is the "opposite" condition for the NPN transistor.

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 data in Figure 4.9.



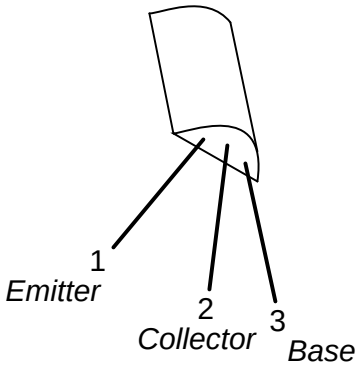
- Meter touching wire 1 (+) and 2 (-): "OL"
- Meter touching wire 1 (-) and 2 (+): "OL"
- Meter touching wire 1 (+) and 3 (-): 0.655 V
- Meter touching wire 1 (-) and 3 (+): "OL"
- Meter touching wire 2 (+) and 3 (-): 0.621 V
- Meter touching wire 2 (-) and 3 (+): "OL"

Figure 4.9: *Unknown bipolar transistor. Which terminals are emitter, base, and collector?* Ω*-meter readings between terminals.*

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 a PNP with base on wire 3, emitter on wire 1 and collector on wire 2 as described in Figure 4.10.



- E and C high R: 1 (+) and 2 (-): "OL"
- C and E high R: 1 (-) and 2 (+): "OL"
- E and B forward: 1 (+) and 3 (-): 0.655 V
- E and B reverse: 1 (-) and 3 (+): "OL"
- C and B forward: 2 (+) and 3 (-): 0.621 V
- C and B reverse: 2 (-) and 3 (+): "OL"

Figure 4.10: *BJT terminals identified by Ω-meter.*

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 in Figure 4.11 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 (up-wards) 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, a saturated transistor shows very little opposition to electrons, all the way from emitter to collector, as evidenced by the lamp's illumination!

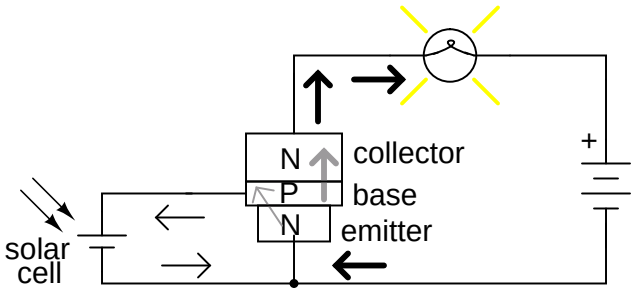


Figure 4.11: *A small base current flowing in the forward biased base-emitter junction allows a large current flow through the reverse biased base-collector junction.*

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, as in Figure 4.12.

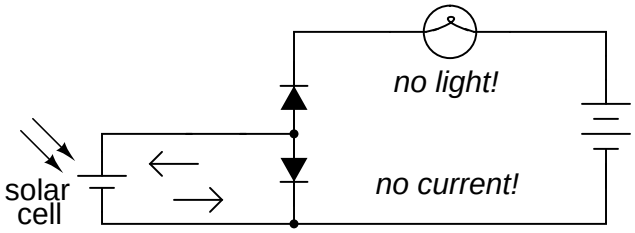


Figure 4.12: *A pair of back-to-back diodes don't act like a transistor!*

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 behaves 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). If 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; 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. See Bipolar junction transistors, Ch 2 for more details.

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), collector and emitter are definitely not identical!

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, a certain, number of electrons pass through the base-collector junction and no more.

In the next section, this current-limiting 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, because of heavier 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 may be thought of as "opening a gate" for a certain, limited amount of current through the collector.

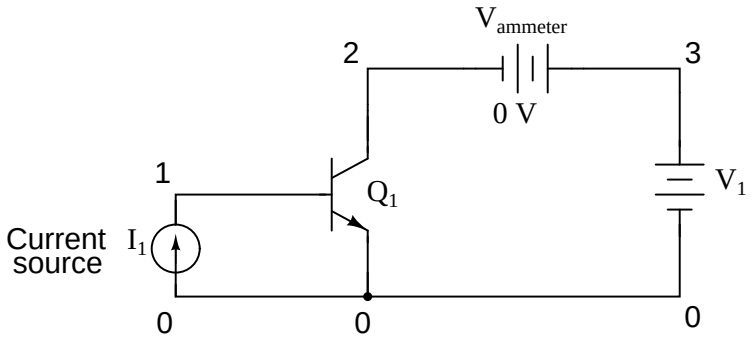
4.4 Active mode operation

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 of 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 the automobile cruising at a constant, controlled speed (constant, controlled collector current) as dictated by the driver. *Saturation* the automobile driving up a steep hill that prevents it from going as fast as the driver wishes. In other words, a "saturated" automobile is one with the accelerator pedal pushed all the way down (base current calling for more collector current than can be provided by the power supply/load circuit).

Let's set up a circuit for SPICE simulation to demonstrate what happens when a transistor is in its active mode of operation. (Figure 4.13)




```
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
```

Figure 4.13: *Circuit for "active mode" SPICE simulation, and netlist.*

"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 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, we'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 ($V_{ammeter}$) in Figure 4.13 with its output of 0 volts serves merely to provide SPICE with a circuit element for current measurement.

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 in (Figure 4.15) 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. (Figure 4.15)

Same result! The collector current in Figure 4.15 holds absolutely steady at 2 mA, although 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.

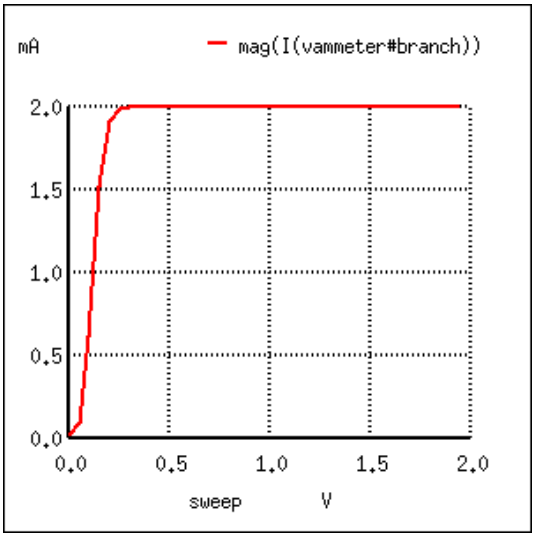
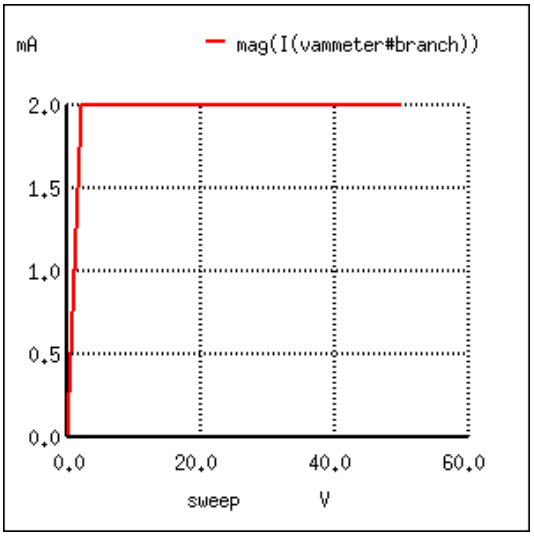


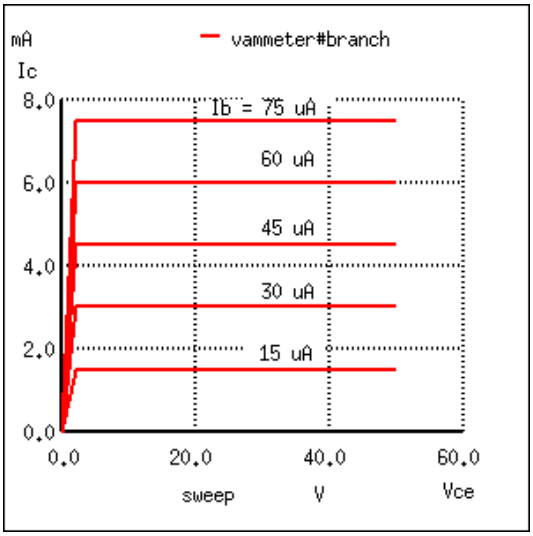
Figure 4.14: *A Sweeping collector voltage 0 to 2 V with base current constant at 20 μA yields constant 2 mA collector current in the saturation region.*




```
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
```

Figure 4.15: *Sweeping collector voltage 0 to 50 V with base current constant at 20 μA yields constant 2 mA collector current.*

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 in Figure 4.16.




```
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 i1
15u 75u 15u
.plot dc
i(vammeter)
.end
```

Figure 4.16: *Sweeping collector voltage 0 to 50 V (.dc v1 0 50 2) with base current constant at 75 μA yields constant 7.5 mA collector current. Other curves are generated by current sweep (i1 15u 75u 15u) in DC analysis statement (.dc v1 0 50 2 i1 15u 75u 15u).*

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 as long as it is 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 as in Figure 4.17. 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):

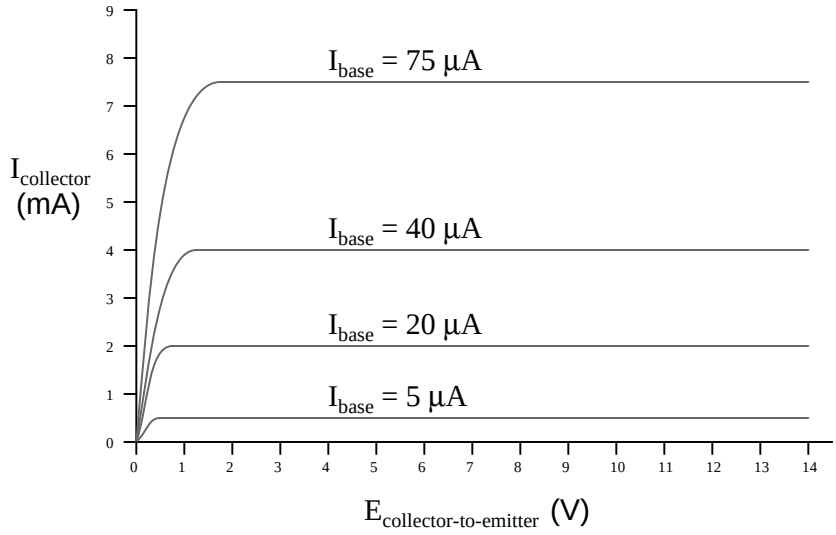


Figure 4.17: *Voltage collector to emitter vs collector current for various base currents.*

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

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 precise predictions of transistor performance with detailed equations. Hybrid parameter variables are many, but each is 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." The β ratio is unitless.

β for any transistor is determined by its design: it cannot be altered after manufacture. It is rare to have two transistors of the same design exactly match because of the physical variables afecting β . 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.

The β of a transistor does not remain stable for all operating conditions. 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; 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 model in Figure 4.18 is used in many

introductory electronics texts.

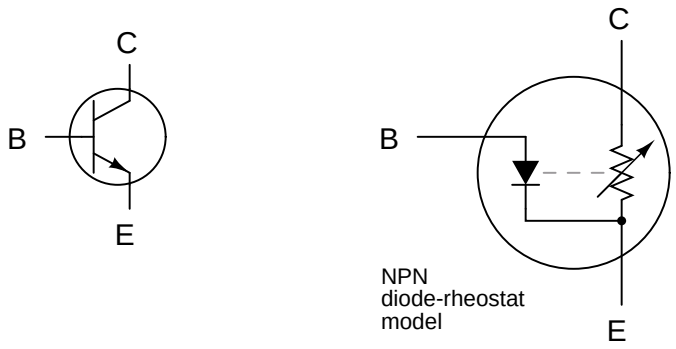


Figure 4.18: *Elementary diode resistor transistor model.*

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 don't like this model because it miscommunicates 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 shown in Figure 4.19.

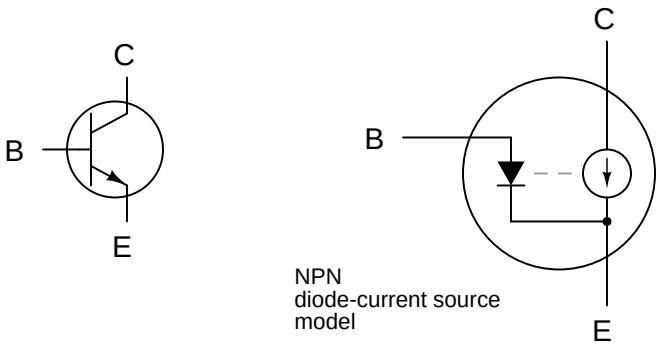


Figure 4.19: *Current source model of transistor.*

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. The current source does not model the fact that its source of energy is a external power supply, similar to an amplifier.

- **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 regulates collector current. By *regulate*, we mean that no more collector current can 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
- β changes 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 previously illustrated the transistor's switching ability. (Figure 4.20)

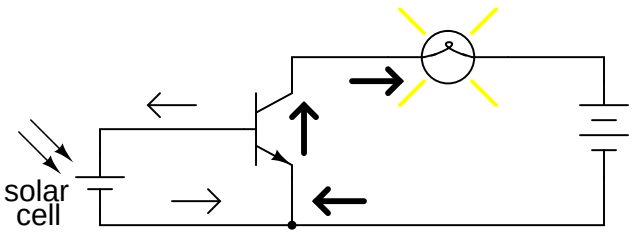


Figure 4.20: *NPN transistor as a simple switch.*

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 shown in Figure 4.21. 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.

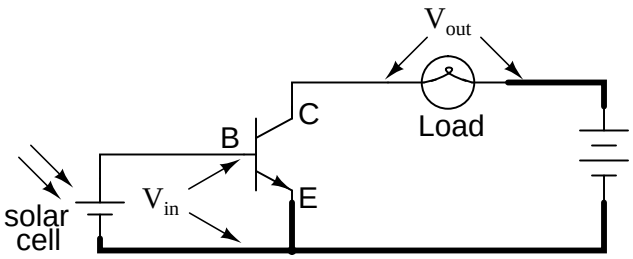


Figure 4.21: *Common-emitter amplifier: The input and output signals both share a connection to the emitter.*

Before, a small solar cell current saturated a transistor, illuminating 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 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 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 (Figure 4.22) for this purpose. In fact, the simplest light-exposure meters for photography work are designed like this.

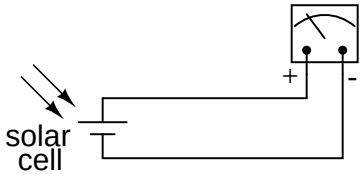


Figure 4.22: *High intensity light directly drives light meter.*

Although 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 (Figure 4.23) to *amplify* the solar cell's current so that more meter deflection may be obtained for less incident light.

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

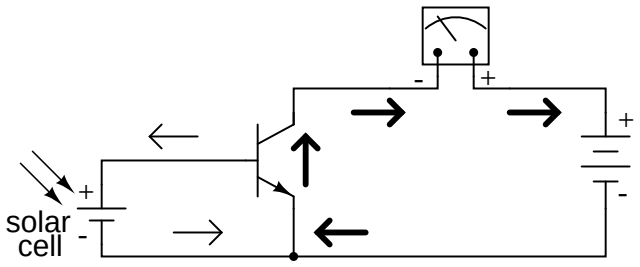


Figure 4.23: *Cell current must be amplified for low intensity light.*

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 current through the movement coil, meter indication in this circuit should depend only on the 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 drive the meter full-scale.

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 in Figure 4.24

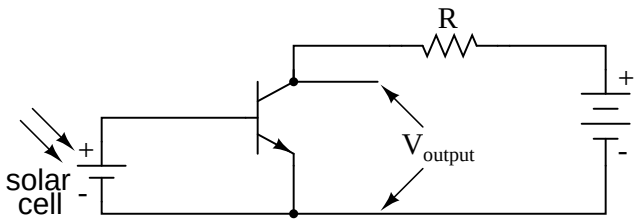


Figure 4.24: *Common emitter amplifier develops voltage output due to current through load resistor.*

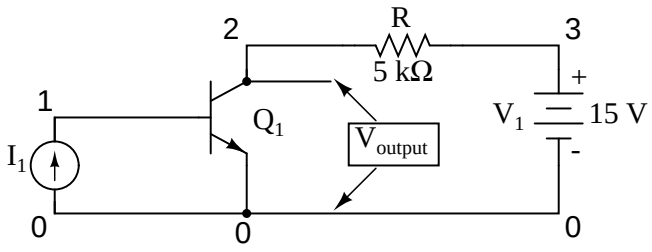
With the solar cell darkened (no current), the transistor will be in cutoff mode and behave as an open switch between collector and emitter. This will produce maximum voltage drop between collector and emitter for maximum V_{output}, equal to the full voltage of the battery.

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 because of 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 *inverted* with respect to the input signal. 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 (Figure 4.26) of the circuit in Figure 4.25 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
```

Figure 4.25: *Common emitter schematic with node numbers and corresponding SPICE netlist.*

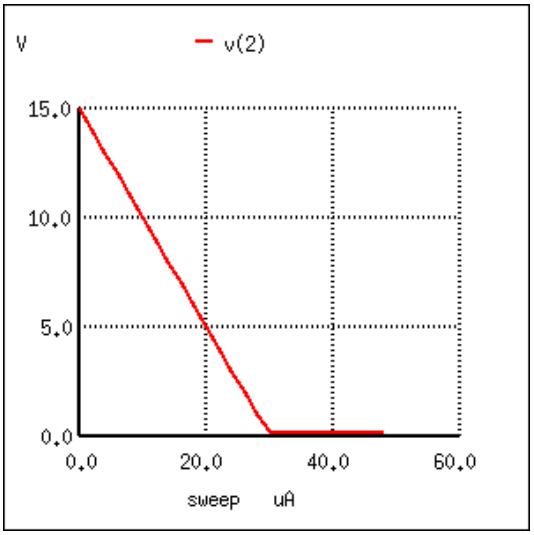


Figure 4.26: *Common emitter: collector voltage output vs base current input.*

At the beginning of the simulation in Figure 4.26 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. Often an *AC* amplifier for amplifying *alternating* current and voltage signals is desired. One common application of this is in audio electronics (radios, televisions, and public-address systems). Earlier, we saw an example of the audio output of a tuning fork activating a transistor switch. (Figure 4.27) Let's see if we can modify that circuit to send power to a speaker rather than to a lamp in Figure 4.28.

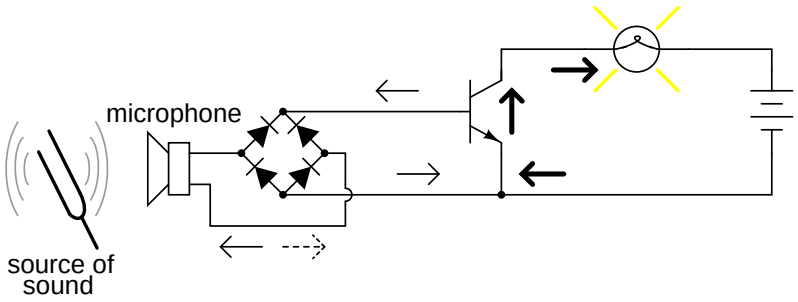


Figure 4.27: *Transistor switch activated by audio.*

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 the circuit in Figure 4.29 with SPICE. The netlist is included in (Figure 4.30)

The simulation plots (Figure 4.30) 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

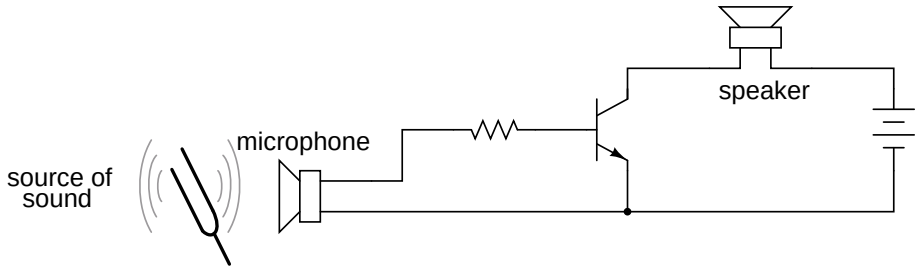


Figure 4.28: *Common emitter amplifier drives speaker with audio frequency signal.*

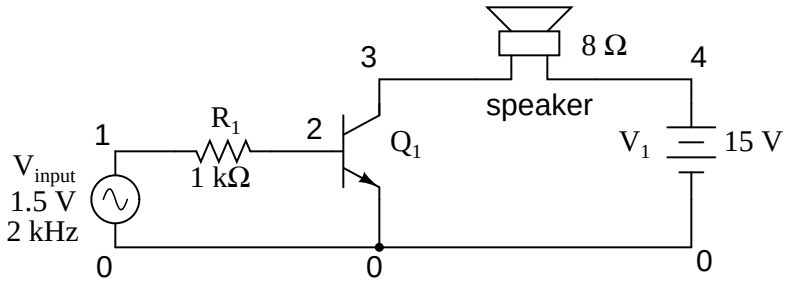
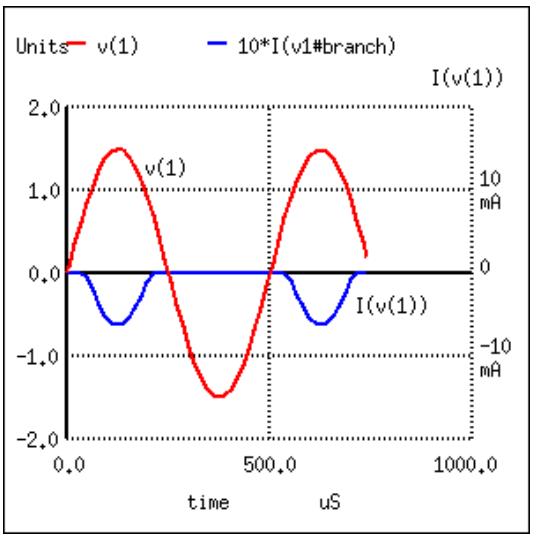


Figure 4.29: *SPICE version of common emitter audio amplifier.*




```
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
```

Figure 4.30: *Signal clipped at collector due to lack of DC base bias.*

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 current source model in Figure 4.31.

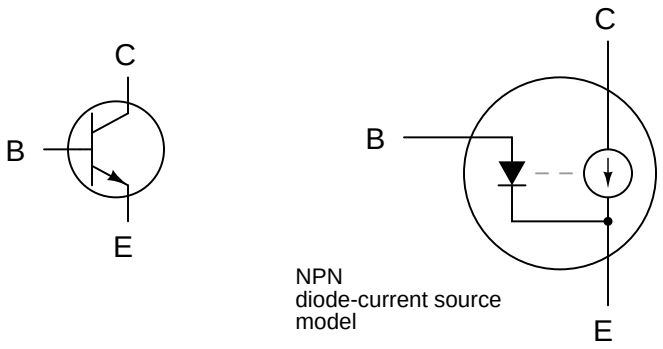


Figure 4.31: *The model shows that base current flow in on direction.*

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. 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 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. (Figure 4.32)

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. (Figure 4.33) 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 illustration in Figure 4.34 is another view of the same circuit, this time with a few

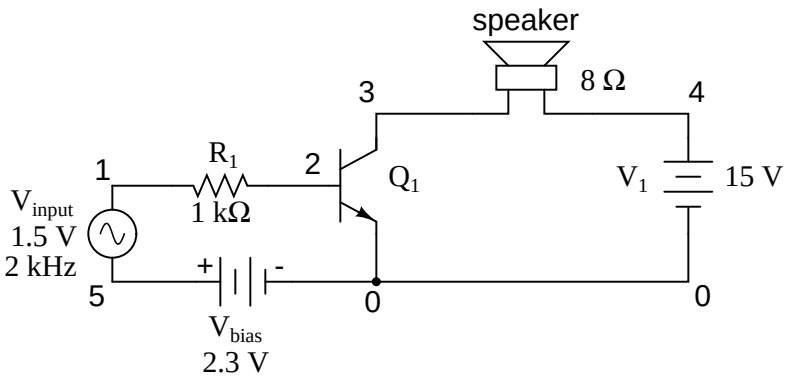
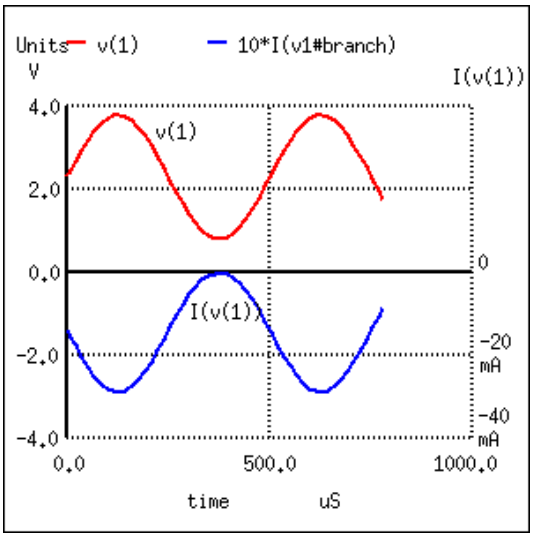


Figure 4.32: *V_{bias} keeps transistor in the active region.*




```
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
```

Figure 4.33: *Undistorted output current I(v(1) due to Vbias*

oscilloscopes (“scopemeters”) connected at crucial points to display all the pertinent signals.

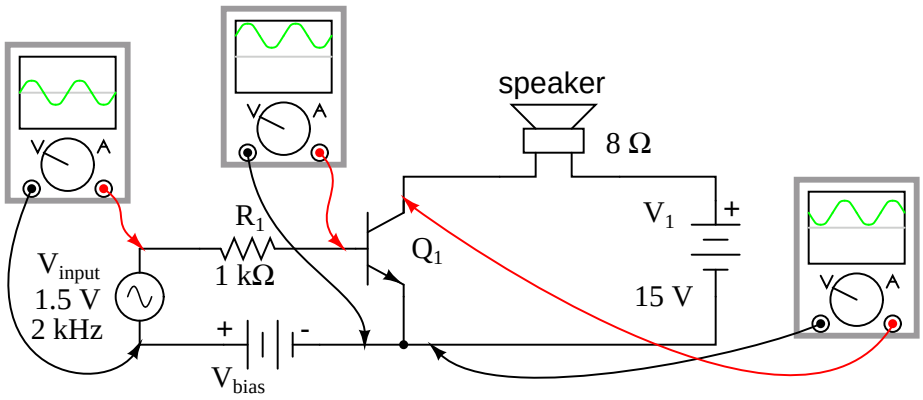


Figure 4.34: *Input is biased upward at base. Output is inverted.*

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 in Table 4.1 (these statistics were cut from the last two analyses for brevity’s sake).

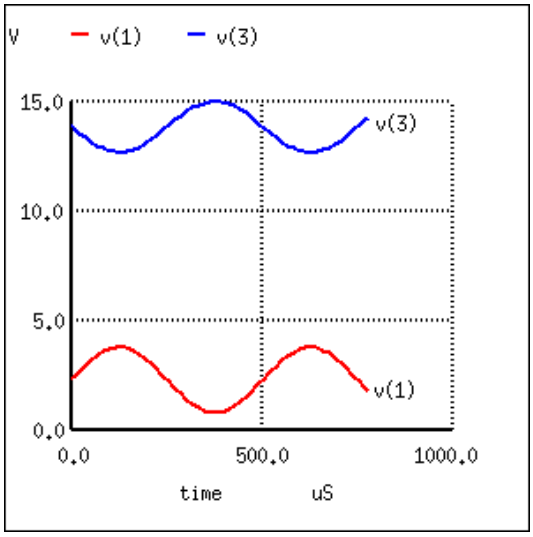
Table 4.1: *BJT SPICE model parameters.*

```
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 modify our last SPICE analysis to plot output voltage rather than output current so we have two voltage plots to compare in Figure 4.35.




```
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(3)
.end
```

Figure 4.35: *V(3), the output voltage across r_{spkr}, compared to the input.*

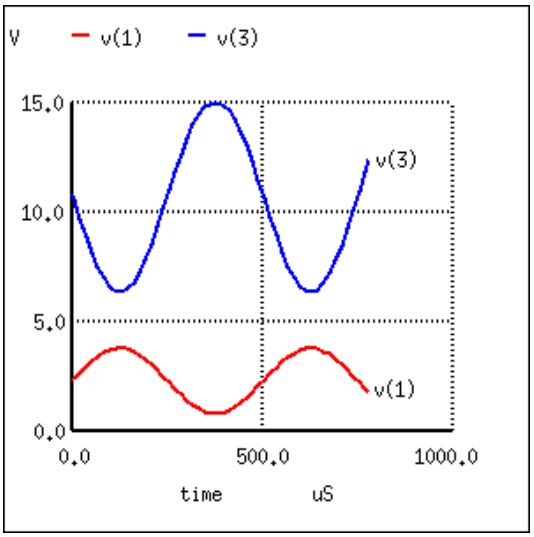
Plotted on the same scale (from 0 to 4 volts), we see that the output waveform in Figure 4.35

To be honest, this low voltage gain is not characteristic to *all* common-emitter amplifiers. 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 in Figure 4.36 load instead of an 8 Ω load.

This time the output voltage waveform in Figure 4.36 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 do another computer analysis of this circuit, 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. (Table 4.2)

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.




```
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(3)
.end
```

Figure 4.36: *Increasing r_{spkr} to 30 Ω increases the output voltage.*

Table 4.2: *SPICE netlist for printing AC input and output voltages.*

```
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
```

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

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

$$A_V = 2.9453$$

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

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

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

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:

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

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

$$A_V = 3$$

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

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

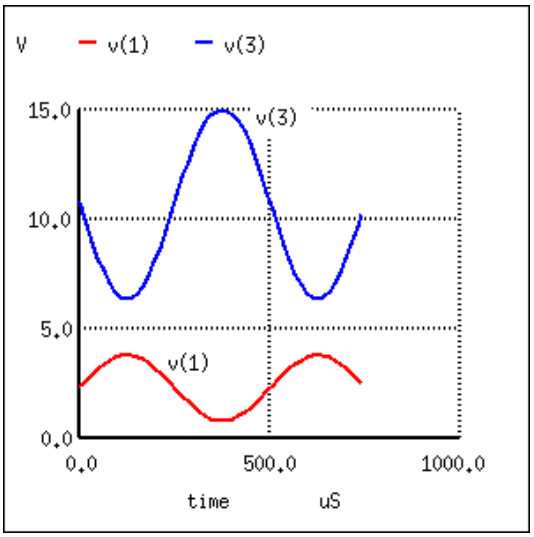
$$A_{V(dB)} = 9.5424\ dB$$

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 in Figure 4.37 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 in Table 4.3 is an AC analysis, to provide simple, peak voltage readings for input and output.

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 tran-




```
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
```

Figure 4.37: *Common-emitter amplifier shows a voltage gain with R_{spkr}=30Ω*

Table 4.3: *SPICE netlist for AC analysis*

```
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
```

sistors are just as valid to use as NPN in *any* amplifier configuration, as long as the proper polarity and current directions are maintained, and the common-emitter amplifier is no exception. The output invertion and gain of a PNP transistor amplifier are the same as its NPN counterpart, just the battery polarities are different. (Figure 4.38)

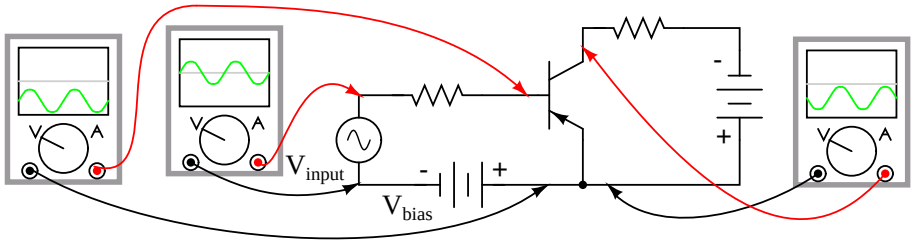


Figure 4.38: *PNP version of common emitter amplifier.*

- **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. 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. Thus, 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 for gain calculations. Called the *common-collector* configuration, its schematic diagram is shown in Figure 4.39.

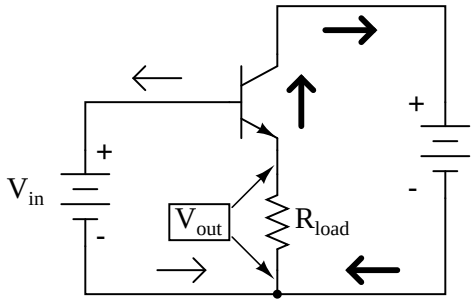


Figure 4.39: *Common collector amplifier has collector common to both input and output.*

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 as in Figure 4.40.

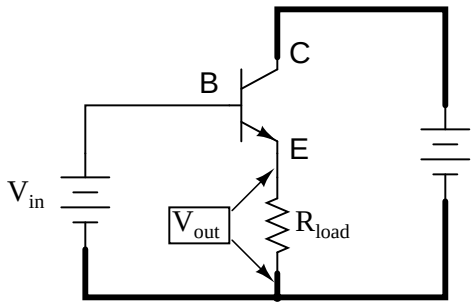


Figure 4.40: *Common collector: Input is applied to base and collector. Output is from emitter-collector circuit.*

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. 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. The circuit is in Figure 4.41. The netlist is in Figure 4.42.

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. See Figure 4.42. As the input voltage increases, so does the output voltage. Moreover, a close examination reveals that the output voltage is nearly *identical* to the input voltage, lagging

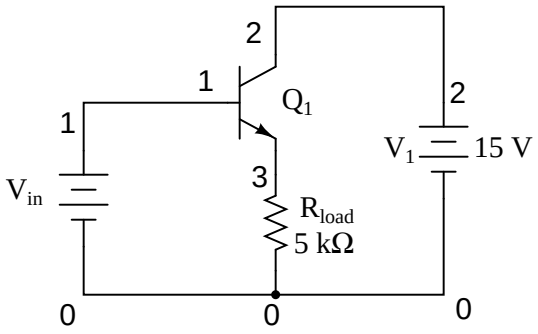
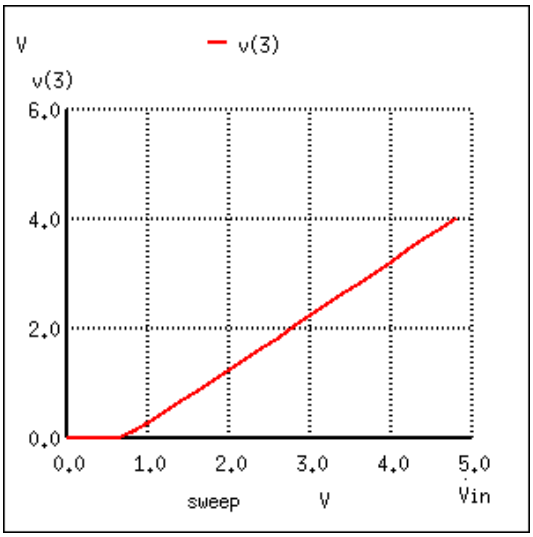


Figure 4.41: *Common collector amplifier for SPICE.*




```
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
```

Figure 4.42: *Common collector: Output equals input less a 0.7 V V_{BE} drop.*

behind by about 0.7 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 to the diode current source transistor model in Figure 4.43, we see that the base current must go through the base-emitter PN junction, which is equivalent to a normal rectifying diode. If 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; thus, we can regard it as being constant:

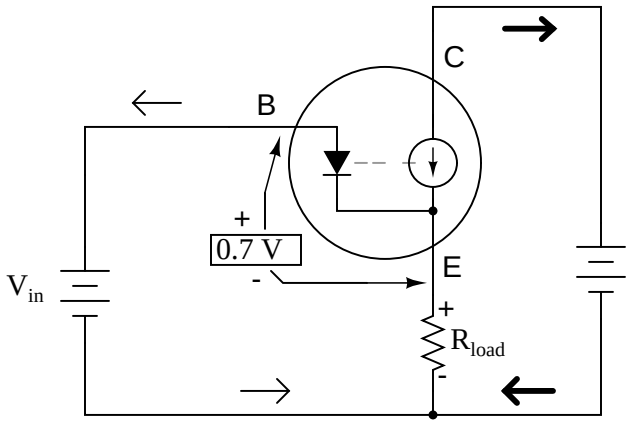


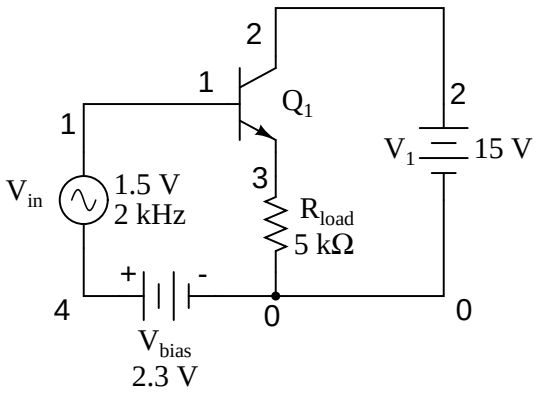
Figure 4.43: *Emitter follower: Emitter voltage follows base voltage (less a 0.7 V V_{BE} drop.)*

Given the voltage polarities across the base-emitter PN junction and the load resistor, we see that these *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, because the emitter load voltages follow the input 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 the non-inverting amplifier in Figure 4.44.

The results of the SPICE simulation in Figure 4.45 show that the output follows the input. The output is the same peak-to-peak amplitude as the input. Though, the DC level is shifted




```
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
```

Figure 4.44: *Common collector (emitter-follower) amplifier.*

downward by one V_{BE} diode drop.

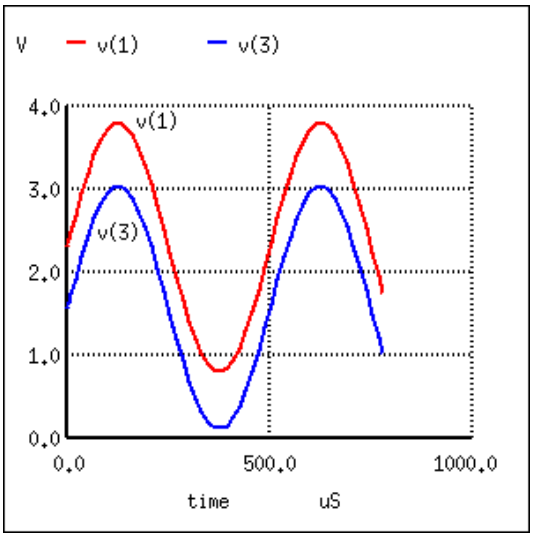


Figure 4.45: *Common collector (emitter-follower): Output V3 follows input V1 less a 0.7 V VBE drop.*

Here's another view of the circuit (Figure 4.46) 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

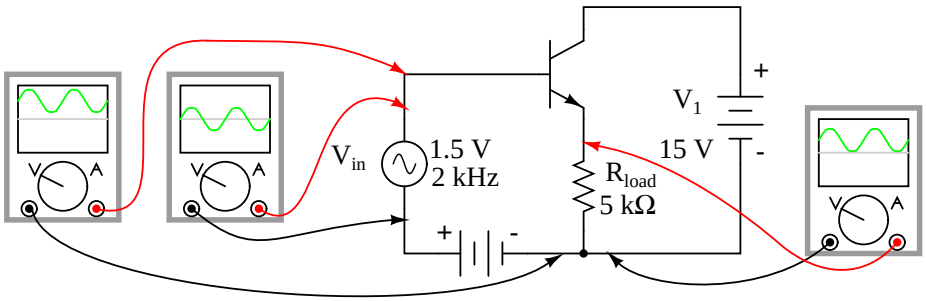


Figure 4.46: *Common collector non-inverting voltage gain is 1.*

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 of the amplified signal. The only difference is in voltage polarities and current directions shown in Figure 4.47.

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 shown in Figure 4.48.

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, a plain zener diode voltage regulator would have to shunt a heavy current through the diode to be effective at regulating load voltage in the event of large load resistance or voltage source changes.

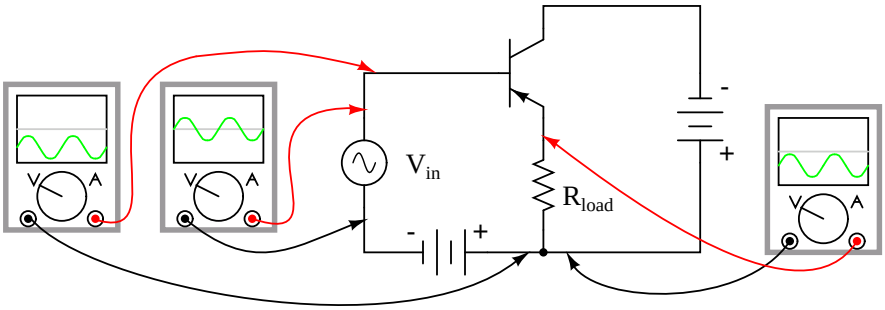


Figure 4.47: *PNP version of the common-collector amplifier.*

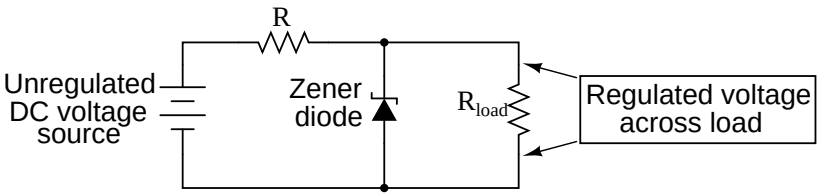


Figure 4.48: *Zener diode voltage regulator.*

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. (Figure 4.49)

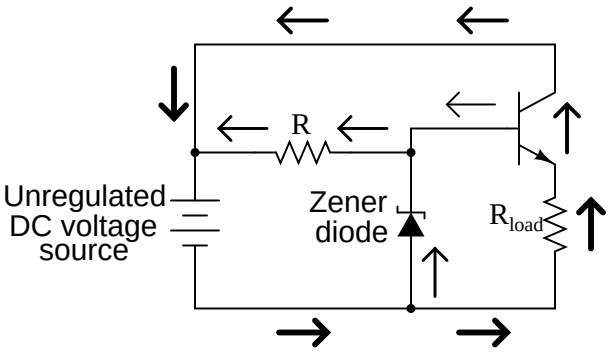


Figure 4.49: *Common collector application: voltage regulator.*

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. 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 shown in Figure 4.50.

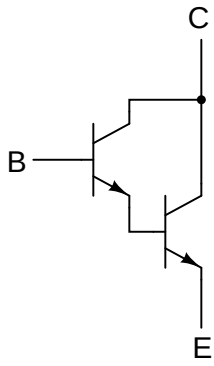


Figure 4.50: *An NPN darlington pair.*

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 shown in Figure 4.51.

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 con-

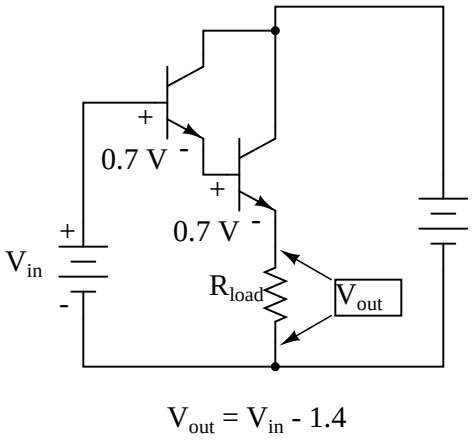


Figure 4.51: *Darlington pair based common-collector amplifier loses two V_{BE} diode drops.*

sidering any power supplies.

- The common-collector amplifier is also known as an emitter-follower.
- 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 (Figure 4.52) 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 shown in Figure 4.53.

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.

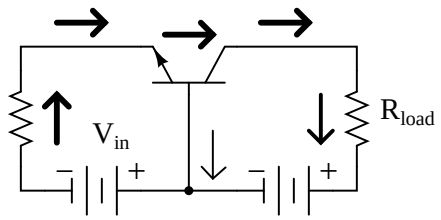


Figure 4.52: *Common-base amplifier*

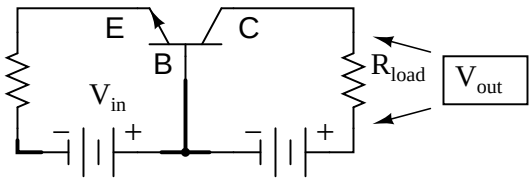


Figure 4.53: *Common-base amplifier: Input between emitter and base, output between collector and base.*

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 of the circuit in Figure 4.54 will vindicate that assumption.

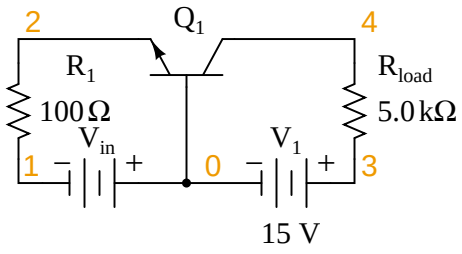
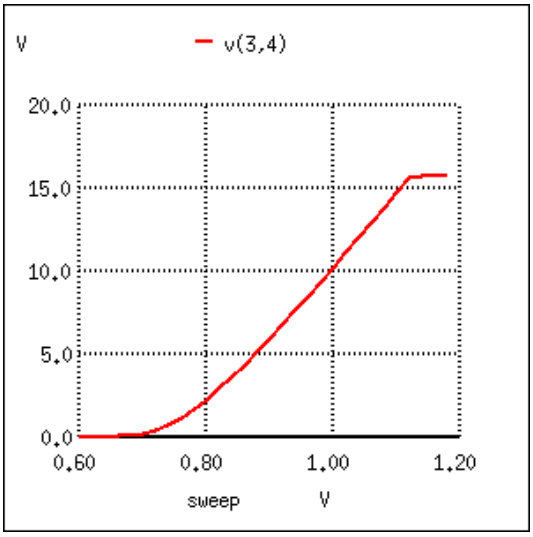


Figure 4.54: *Common-base circuit for DC SPICE analysis.*

Notice in Figure 4.55 that 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




```
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
```

Figure 4.55: *Common-base amplifier DC transfer function.*

(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 input voltage source.

A second set of SPICE analyses (circuit in Figure 4.56) with an AC signal source (and DC bias voltage) tells the same story: a high voltage gain

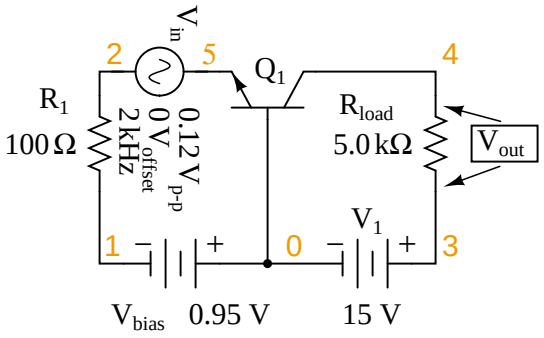
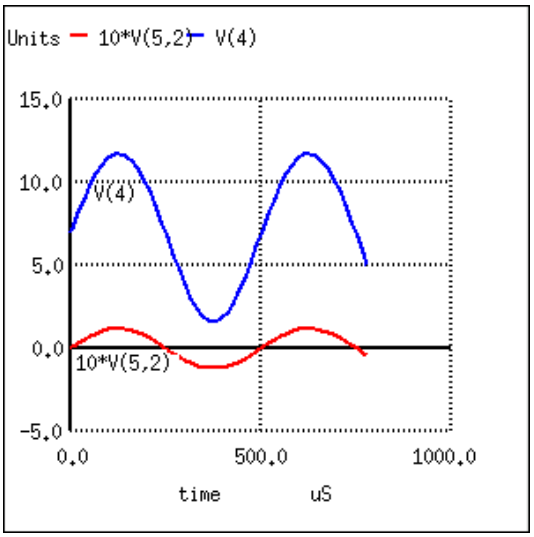


Figure 4.56: *Common-base circuit for SPICE AC analysis.*

As you can see, the input and output waveforms in Figure 4.57 are in phase with each other. This tells us that the common-base amplifier is non-inverting.

The AC SPICE analysis in Table 4.4 at a single frequency of 2 kHz provides input and output voltages for gain calculation.




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

Figure 4.57:

Table 4.4: *Common-base AC analysis at 2 kHz– netlist followed by output.*

```
common-base amplifier
vin 5 2  ac 0.1 sin
vbias 0 1  dc 0.95
r1 2 1 100
q1 4 0 5 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
.ac dec 1 2000 2000
.print ac vm(5,2) vm(4,3)
.end
frequency       mag(v(5,2))     mag(v(4,3))
---------------------------------------------
0.000000e+00    1.000000e-01    4.273864e+00
```

Voltage figures from the second analysis (Table 4.4) show a voltage gain of 42.74 (4.274 V / 0.1 V), or 32.617 dB:

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

$$A_V = \frac{4.274 \text{ V}}{0.10 \text{ V}}$$

$$A_V = 42.74$$

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

$$A_{V(dB)} = 20 \log 42.74$$

$$A_{V(dB)} = 32.62 \text{ dB}$$

Here's another view of the circuit in Figure 4.58, summarizing the phase relations and DC offsets of various signals in the circuit just simulated.

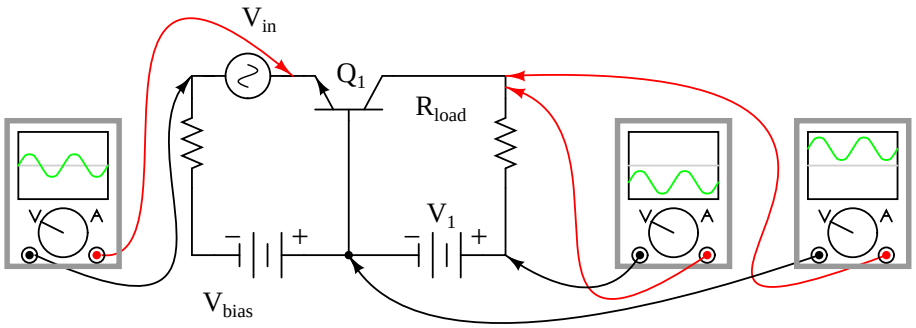


Figure 4.58: *Phase relationships and offsets for NPN common base amplifier.*

. . . and for a PNP transistor: Figure 4.59.

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 (Figure 4.56), changing the DC bias voltage slightly (vbias in Figure 4.60) 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

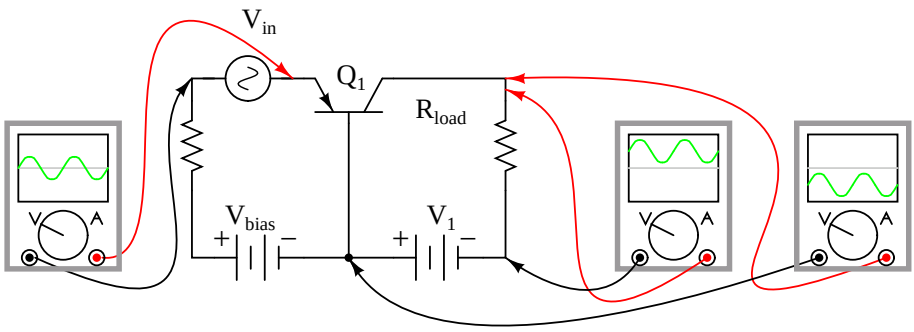


Figure 4.59: *Phase relationships and offsets for PNP common base amplifier.*

amplitudes will be noted.

Although these analyses will all be conducted in the "transfer function" mode, each was 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. See "*.tran 0.02m 0.78m" in Figure 4.60, the "commented out" transient analysis statement. Gain calculations cannot be based on waveforms that are distorted. SPICE can calculate the small signal DC gain for us with the ".tf v(4) vin" statement. The output is **v(4)** and the input as **vin**.

At the command line, **spice -b filename.cir** produces a printed output due to the **.tf** statement: transfer_function, output_impedance, and input_impedance. The abbreviated output listing is from runs with **vbias** at 0.85, 0.90, 0.95, 1.00 V as recorded in Table 4.5.

A trend should be evident in Table 4.5. With increases in DC bias voltage, voltage gain (transfer_function) increases as well. We can see that the voltage gain is increasing because each subsequent simulation (vbias= 0.85, 0.8753, 0.90, 0.95, 1.00 V) produces greater gain (transfer_function= 37.6, 39.4 40.8, 42.7, 44.0), respectively. The changes are largely due to minuscule variations in bias voltage.

The last three lines of Table 4.5(right) show the **I(v1)/Iin** current gain of 0.99. (The last two lines look invalid.) This makes sense for β=100; α= β/(β+1), α=0.99=100/(100-1). The combination of low current gain (always less than 1) and somewhat unpredictable voltage gain conspire against the common-base design, relegating it to few practical applications.

Those few applications include radio frequency amplifiers. The grounded base helps shield the input at the emitter from the collector output, preventing instability in RF amplifiers. The common base configuration is usable at higher frequencies than common emitter or common collector. See "Class C common-base 750 mW RF power amplifier" (page 441). For a more elaborate circuit see "Class A common-base small-signal high gain amplifier" (page 442).

- **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-

```
common-base amp
vbias=0.85V
vin 5 2  sin (0 0.12 2000
0 0)
vbias 0 1  dc 0.85
r1 2 1 100
q1 4 0 5 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
*.tran 0.02m 0.78m
.tf v(4) vin
.end
```

```
common-base amp current gain
Iin 55 5 0A
vin 55 2  sin (0 0.12 2000 0
0)
vbias 0 1  dc 0.8753
r1 2 1 100
q1 4 0 5 mod1
v1 3 0 dc 15
rload 3 4 5k
.model mod1 npn
*.tran 0.02m 0.78m
.tf I(v1) Iin
.end
Transfer function
information:
transfer function =
9.900990e-01
iin input impedance =
9.900923e+11
v1 output impedance =
1.000000e+20
```

Figure 4.60: *SPICE net list: Common-base, transfer function (voltage gain) for various DC bias voltages. SPICE net list: Common-base amp current gain; Note .tf v(4) vin statement. Transfer function for DC current gain I(vin)/Iin; Note .tf I(vin) Iin statement.*

Table 4.5: *SPICE output: Common-base transfer function.*

```
Circuit: common-base amp vbias=0.85V
transfer_function = 3.756565e+01
output_impedance_at_v(4) = 5.000000e+03
vin#input_impedance = 1.317825e+02

Circuit: common-base amp vbias=0.8753V Ic=1 mA
Transfer function information:
transfer_function = 3.942567e+01
output_impedance_at_v(4) = 5.000000e+03
vin#input_impedance = 1.255653e+02

Circuit: common-base amp vbias=0.9V
transfer_function = 4.079542e+01
output_impedance_at_v(4) = 5.000000e+03
vin#input_impedance = 1.213493e+02

Circuit: common-base amp vbias=0.95V
transfer_function = 4.273864e+01
output_impedance_at_v(4) = 5.000000e+03
vin#input_impedance = 1.158318e+02

Circuit: common-base amp vbias=1.00V
transfer_function = 4.401137e+01
output_impedance_at_v(4) = 5.000000e+03
vin#input_impedance = 1.124822e+02
```

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 The cascode amplifier

While the C-B (common-base) amplifier is known for wider bandwidth than the C-E (common-emitter) configuration, the low input impedance (10s of Ω) of C-B is a limitation for many applications. The solution is to precede the C-B stage by a low gain C-E stage which has moderately high input impedance (kΩs). See Figure 4.61. The stages are in a *cascode* configuration, stacked in series, as opposed to cascaded for a standard amplifier chain. See "Capacitor coupled three stage common-emitter amplifier" (page 253) for a cascade example. The cascode amplifier configuration has both wide bandwidth and a moderately high input impedance.

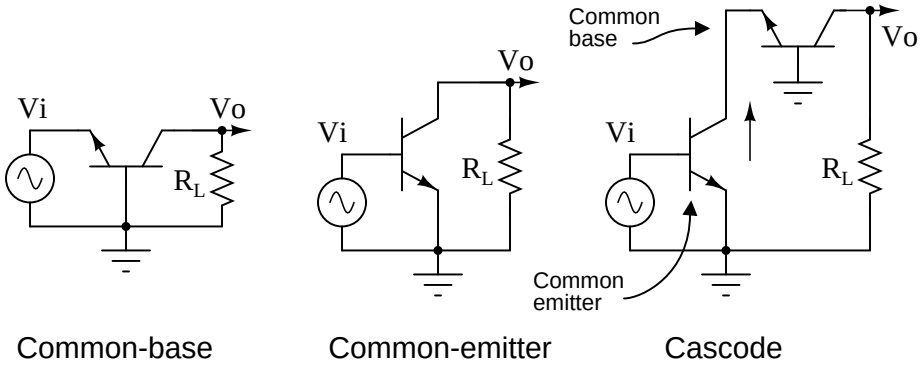


Figure 4.61: *The cascode amplifier is combined common-emitter and common-base. This is an AC circuit equivalent with batteries and capacitors replaced by short circuits.*

The key to understanding the wide bandwidth of the cascode configuration is the *Miller effect*. The (page 277) is the multiplication of the bandwidth robbing collector-base capacitance by voltage gain A_v. This C-B capacitance is smaller than the E-B capacitance. Thus, one would think that the C-B capacitance would have little effect. However, in the C-E configuration, the collector output signal is out of phase with the input at the base. The collector signal capacitively coupled back opposes the base signal. Moreover, the collector feedback is (1-A_v) times larger than the base signal. Keep in mind that A_v is a negative number for the inverting C-E amplifier. Thus, the small C-B capacitance appears (1+$A|_v|$) times larger than its actual value. This capacitive gain reducing feedback increases with frequency, reducing the high frequency response of a C-E amplifier.

The approximate voltage gain of the C-E amplifier in Figure 4.62 is -R_L/r_{EE}. The emitter current is set to 1.0 mA by biasing. R_{EE}= 26mV/I_E = 26mV/1.0ma = 26 Ω. Thus, A_v = -R_L/R_{EE} = -4700/26 = -181. The pn2222 datasheet list C_{cbo} = 8 pF.[5] The miller capacitance is C_{cbo}(1-A_v). Gain A_v = -181, negative since it is inverting gain. C_{miller} = C_{cbo}(1-A_v) = 8pF(1-(-181)=1456pF

A common-base configuration is not subject to the Miller effect because the grounded base shields the collector signal from being fed back to the emitter input. Thus, a C-B amplifier has better high frequency response. To have a moderately high input impedance, the C-E stage is still desirable. The key is to reduce the gain (to about 1) of the C-E stage which reduces the Miller effect C-B feedback to $1{\cdot}C_{CBO}$. The total C-B feedback is the feedback capacitance $1{\cdot}C_{CB}$ plus the actual capacitance C_{CB} for a total of $2{\cdot}C_{CBO}$. This is a considerable reduction from $181{\cdot}C_{CBO}$. The miller capacitance for a gain of -2 C-E stage is $C_{miller} = C_{cbo}(1\text{-}A_v) = C_{miller} = C_{cbo}(1\text{-}(-1)) = C_{cbo}{\cdot}2$.

The way to reduce the common-emitter gain is to reduce the load resistance. The gain of a C-E amplifier is approximately R_C/R_E. The internal emitter resistance r_{EE} at 1mA emitter current is 26Ω. For details on the 26Ω, see "Derivation of R_{EE}", see (page 241). The collector load R_C is the resistance of the emitter of the C-B stage loading the C-E stage, 26Ω again. CE gain amplifier gain is approximately $A_v = R_C/R_E$=26/26=1. This Miller capacitance is $C_{miller} = C_{cbo}(1\text{-}A_v)$ = 8pF(1-(-1)=16pF. We now have a moderately high input impedance C-E stage without suffering the Miller effect, but no C-E dB voltage gain. The C-B stage provides a high voltage gain, A_V = -181. Current gain of cascode is β of the C-E stage, 1 for the C-B, β overall. Thus, the cascode has moderately high input impedance of the C-E, good gain, and good bandwidth of the C-B.

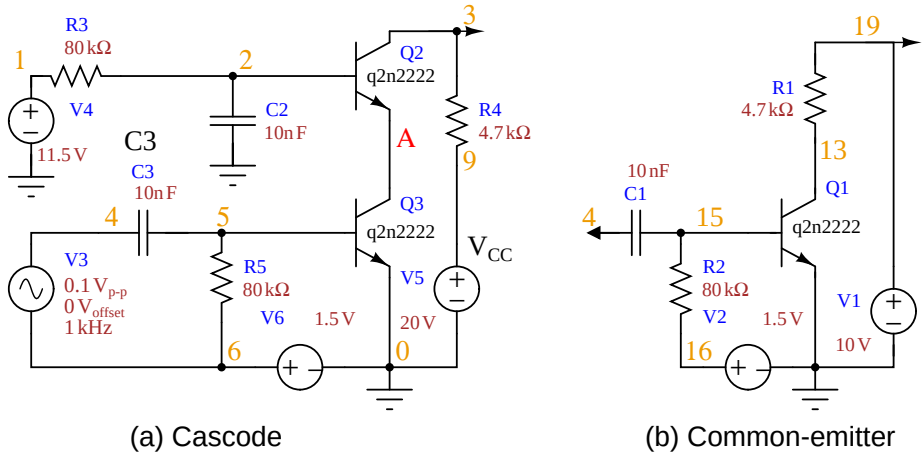


Figure 4.62: *SPICE: Cascode and common-emitter for comparison.*

The SPICE version of both a cascode amplifier, and for comparison, a common-emitter amplifier is shown in Figure 4.62. The netlist is in Table 4.6. The AC source V3 drives both amplifiers via node 4. The bias resistors for this circuit are calculated in an example problem (page 246).

The waveforms in Figure 4.63 show the operation of the cascode stage. The input signal is displayed multiplied by 10 so that it may be shown with the outputs. Note that both the Cascode, Common-emitter, and Va (intermediate point) outputs are inverted from the input. Both the Cascode and Common emitter have large amplitude outputs. The Va point has a DC level of about 10V, about half way between 20V and ground. The signal is larger than can be accounted for by a C-E gain of 1, It is three times larger than expected.

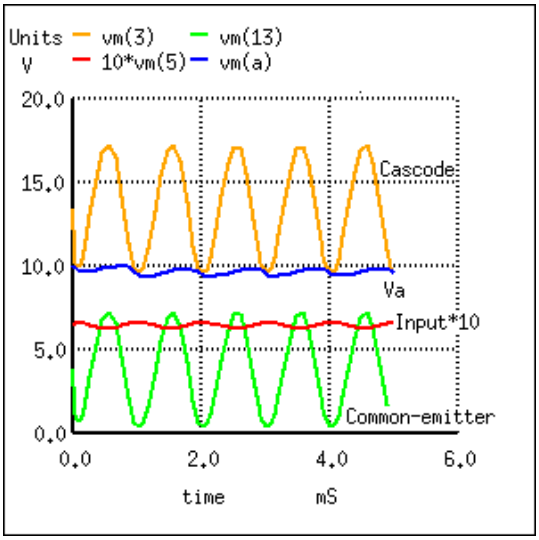


Figure 4.63: *SPICE waveforms. Note that Input is multiplied by 10 for visibility.*

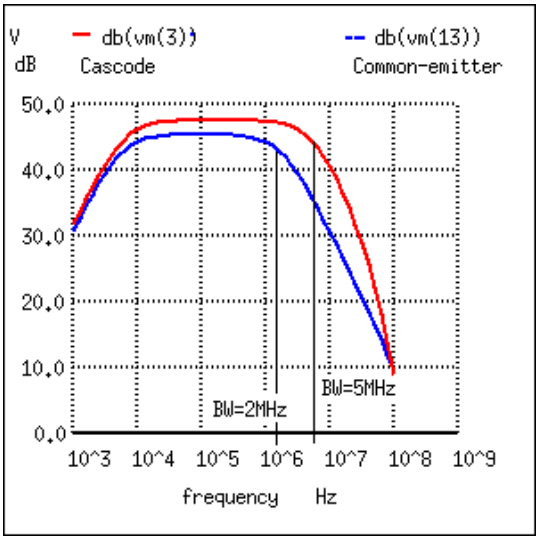


Figure 4.64: *Cascode vs common-emitter banwidth.*

Table 4.6: *SPICE netlist for printing AC input and output voltages.*

```
*SPICE circuit <03502.eps> from XCircuit v3.20
V1 19 0 10
Q1 13 15 0 q2n2222
Q2 3 2 A q2n2222
R1 19 13 4.7k
V2 16 0 1.5
C1 4 15 10n
R2 15 16 80k
Q3 A 5 0 q2n2222
V3 4 6 SIN(0 0.1 1k)  ac 1
R3 1 2 80k
R4 3 9 4.7k
C2 2 0 10n
C3 4 5 10n
R5 5 6 80k
V4 1 0 11.5
V5 9 0 20
V6 6 0 1.5
.model q2n2222 npn (is=19f bf=150
+ vaf=100 ikf=0.18 ise=50p ne=2.5 br=7.5
+ var=6.4 ikr=12m isc=8.7p nc=1.2 rb=50
+ re=0.4 rc=0.3 cje=26p tf=0.5n
+ cjc=11p tr=7n xtb=1.5 kf=0.032f af=1)
.tran 1u 5m
.AC DEC 10 1k 100Meg
.end
```

Figure 4.64 shows the frequency response to both the cascode and common-emitter amplifiers. The SPICE statements responsible for the AC analysis, extracted from the listing:

```
V3 4 6 SIN(0 0.1 1k)  ac 1
.AC DEC 10 1k 100Meg
```

Note the "ac 1" is necessary at the end of the V3 statement. The cascode has marginally better mid-band gain. However, we are primarily looking for the bandwidth measured at the -3dB points, down from the midband gain for each amplifier. This is shown by the vertical solid lines in Figure 4.64. It is also possible to print the data of interest from nutmeg to the screen, the SPICE graphical viewer (command, first line):

```
nutmeg 6 -> print frequency db(vm(3)) db(vm(13))
Index   frequency      db(vm(3))   db(vm(13))
22      0.158MHz       47.54       45.41
33      1.995MHz       46.95       42.06
37      5.012MHz       44.63       36.17
```

Index 22 gives the midband dB gain for Cascode vm(3)=47.5dB and Common-emitter vm(13)=45.4dB. Out of many printed lines, Index 33 was the closest to being 3dB down from 45.4dB at 42.0dB for the Common-emitter circuit. The corresponding Index 33 frequency is approximately 2Mhz, the common-emitter bandwidth. Index 37 vm(3)=44.6db is approximately 3db down from 47.5db. The corresponding Index37 frequency is 5Mhz, the cascode bandwidth. Thus, the cascode amplifier has a wider bandwidth. We are not concerned with the low frequency degradation of gain. It is due to the capacitors, which could be remedied with larger ones.

The 5MHz bandwith of our cascode example, while better than the common-emitter example, is not exemplary for an RF (radio frequency) amplifier. A pair of RF or microwave transistors with lower interelectrode capacitances should be used for higher bandwidth. Before the invention of the RF dual gate MOSFET, the BJT cascode amplifier could have been found in UHF (ultra high frequency) TV tuners.

- REVIEW
- A *cascode* amplifier consists of a common-emitter stage loaded by the emitter of a common-base stage.
- The heavily loaded C-E stage has a low gain of 1, overcoming the *Miller effect*
- A cascode amplifier has a high gain, moderately high input impedance, a high output impedance, and a high bandwidth.

4.9 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*.

A half-wave output is not problematic for some applications. In fact, some applications may *necessitate* this very kind of amplification. Because it is possible to operate an amplifier in modes other than full-wave reproduction and specific applications require 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 with alphabetical letters: A, B, C, and AB.

For Class A operation, 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.

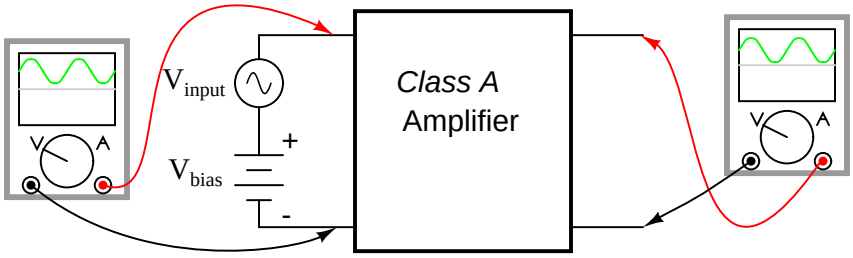


Figure 4.65: *Class A: The amplifier output is a faithful reproduction of the input.*

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.

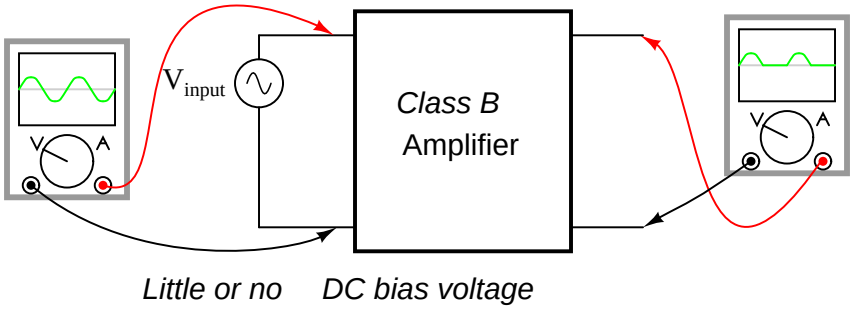


Figure 4.66: *Class B: Bias is such that half (180^o) of the waveform is reproduced.*

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:

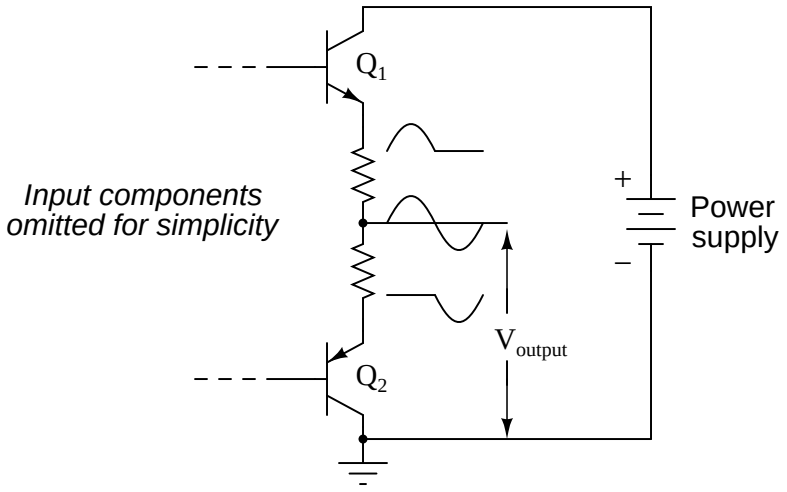


Figure 4.67: *Class B push pull amplifier: Each transistor reproduces half of the waveform. Combining the halves produces a faithful reproduction of the whole wave.*

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, both 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 considerable 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. 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.

Another class of amplifier operation known as *class AB*, 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 most 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

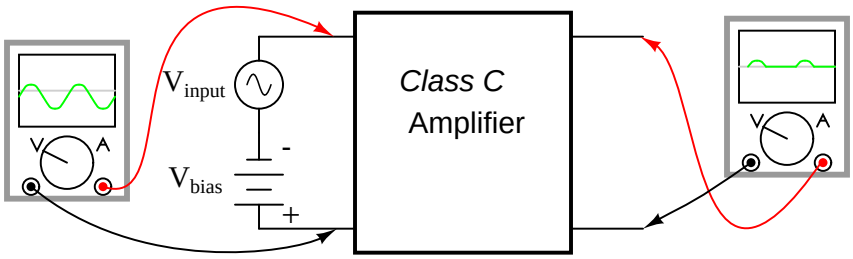


Figure 4.68: *Class C: Conduction is for less than a half cycle (< 180^o).*

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:

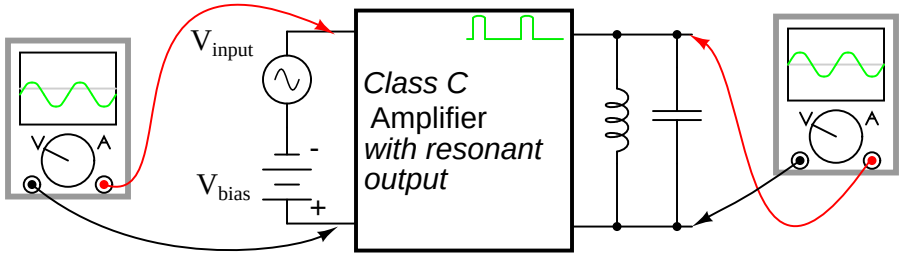


Figure 4.69: *Class C amplifier driving a resonant circuit.*

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 circuit is usable only for amplifying signals of definite, fixed amplitude. A class C amplifier may used in an FM (frequency modulation) radio transmitter. However, the class C amplifier may not directly amplify an AM (amplitude modulated) signal due to distortion.

Another kind 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 is 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 plots in (Figure 4.70 demonstrate this principle.

The greater the instantaneous voltage of the input signal, the greater the duty cycle of

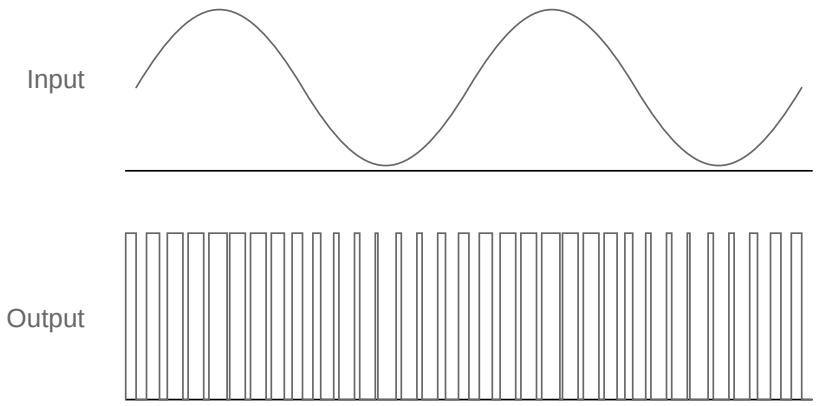


Figure 4.70: *Class D amplifier: Input signal and unfiltered output.*

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, these 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 zero input 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 these 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 shown in Figure 4.71.

A load line is a plot of collector-to-emitter voltage over a range of collector 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

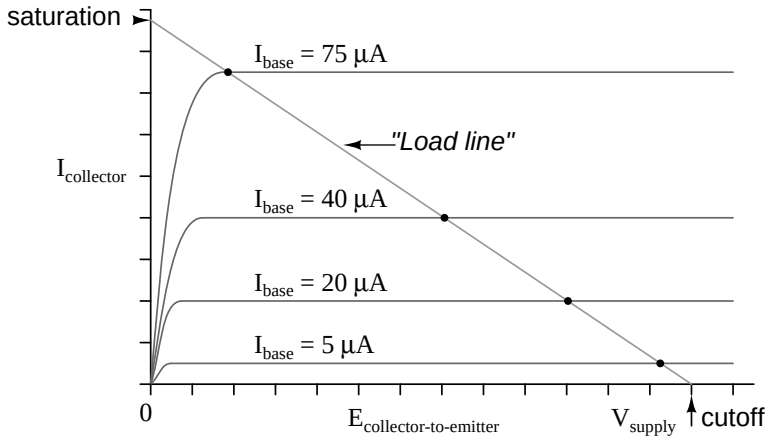


Figure 4.71: *Example load line drawn over transistor characteristic curves from V_{supply} to saturation current.*

given.

Quiescent operating conditions may be shown on this 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 as in (Figure 4.72.

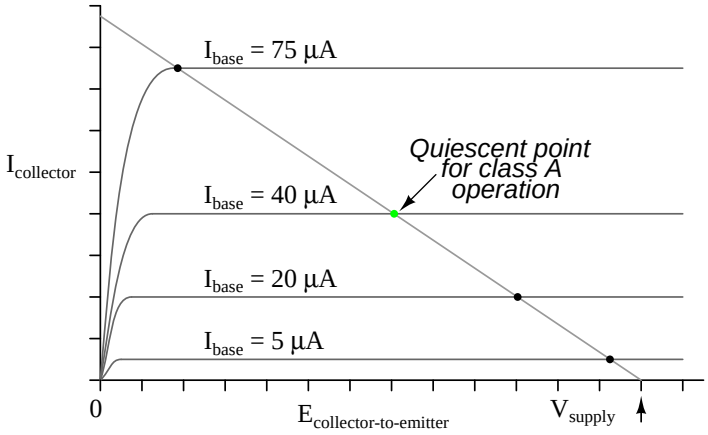


Figure 4.72: *Quiescent point (dot) for class A.*

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 as in (Figure 4.73)

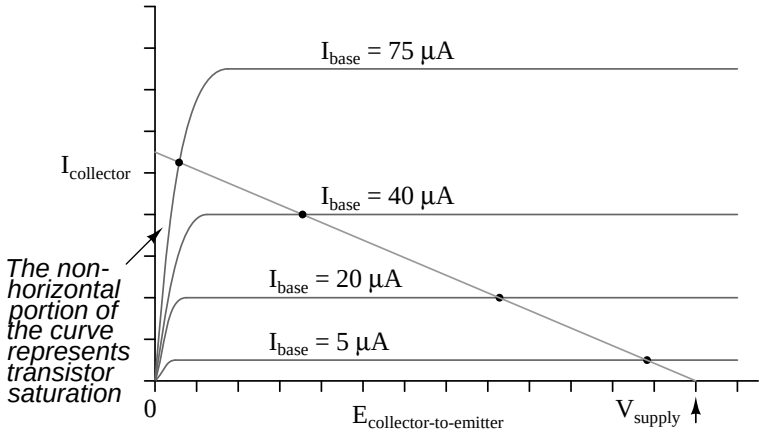


Figure 4.73: *Load line resulting from increased load resistance.*

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. However, collector current increases marginally between base currents 75 μA and 40 μA, because the transistor begins to lose sufficient collector-emitter voltage to continue to regulate collector current.

To maintain linear (no-distortion) operation, transistor amplifiers shouldn't be operated at points where the transistor will saturate; that is, where the load line will not potentially fall on the horizontal portion of a collector current curve. We'd have to add a few more curves to the graph in Figure 4.74 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 shown in Figure 4.75.

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

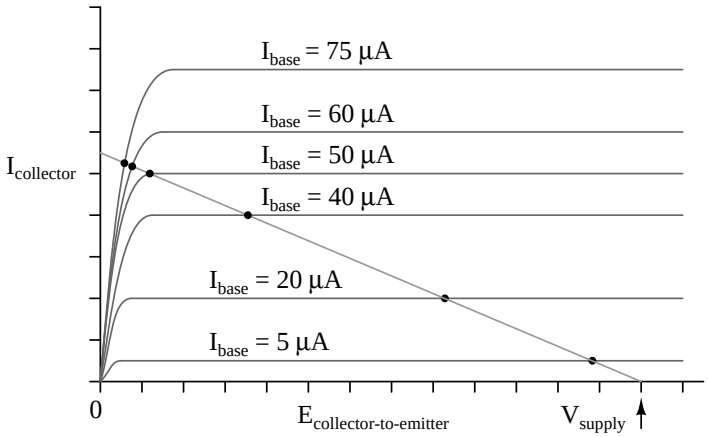


Figure 4.74: *More base current curves shows saturation detail.*

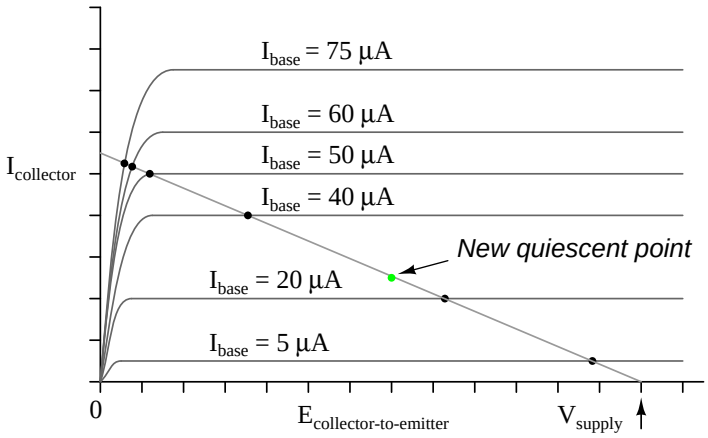


Figure 4.75: *New quiescent point avoids saturation region.*

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 toward class B operation.

Take this circuit, illustrated in the common-emitter section for a SPICE simulation, for instance, in Figure 4.76.

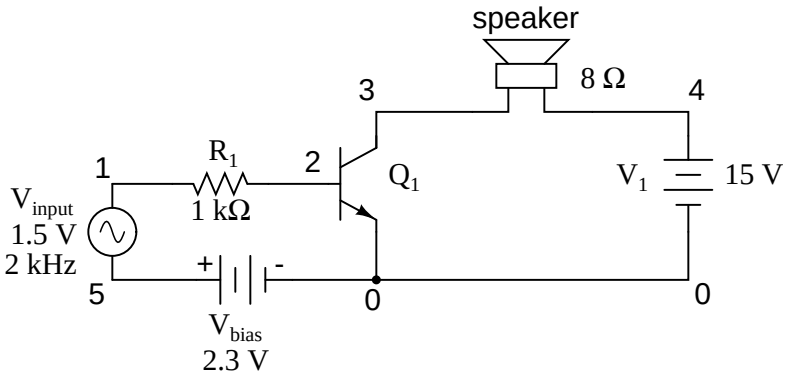


Figure 4.76: *Impractical base battery bias.*

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 easy to design and build. Let's see how this might look in Figure 4.77.

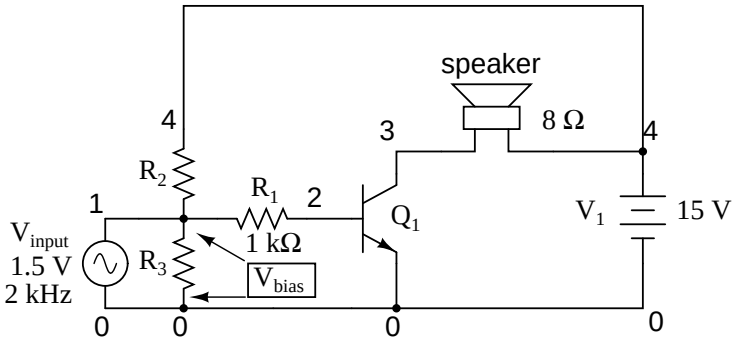


Figure 4.77: *Voltage divider bias.*

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 as in Figure 4.78.

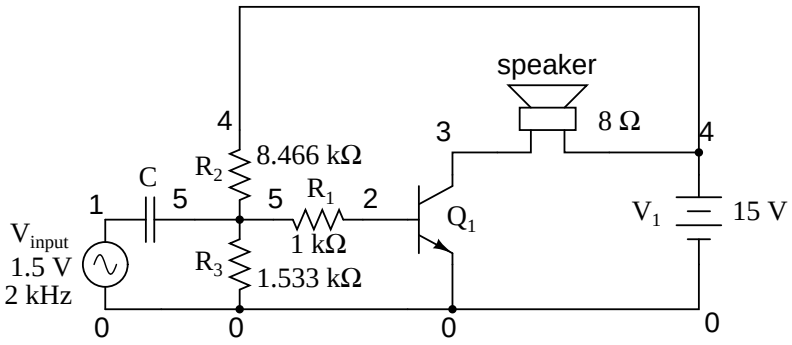


Figure 4.78: *Coupling capacitor prevents voltage divider bias from flowing into signal generator.*

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:

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 in Figure 4.80.

Combining these two separate analyses in Figure 4.81, 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 – its about time for a SPICE simulation of the whole amplifier circuit in Figure 4.82. We will use a capacitor value of 100 μF to obtain an arbitrarily low (0.796 Ω) impedance at 2000 Hz:

Note the substantial distortion in the output waveform in Figure 4.82. 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?

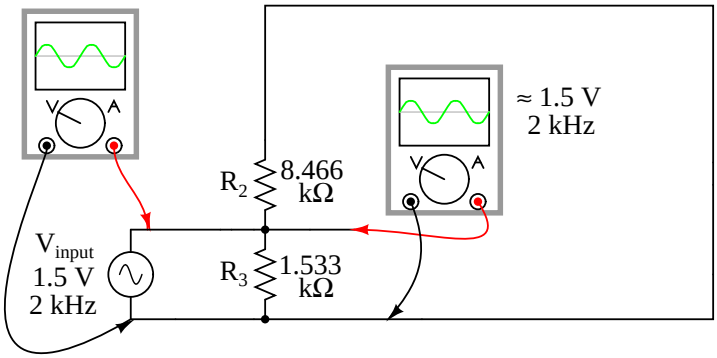


Figure 4.79: *Due to the coupling capacitor's very low impedance at the signal frequency, it behaves much like a piece of wire, thus can be omitted for this step in superposition analysis.*

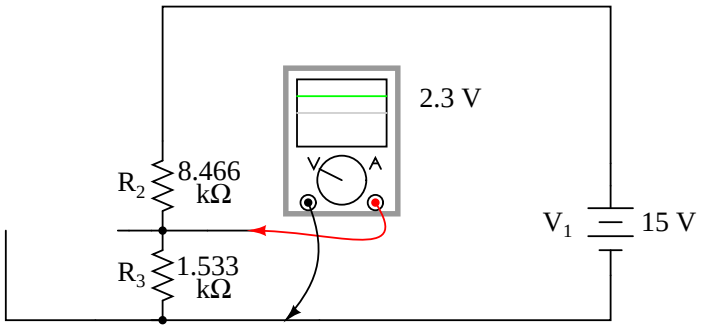


Figure 4.80: *The capacitor appears to be an open circuit as far at the DC analysis is concerned*

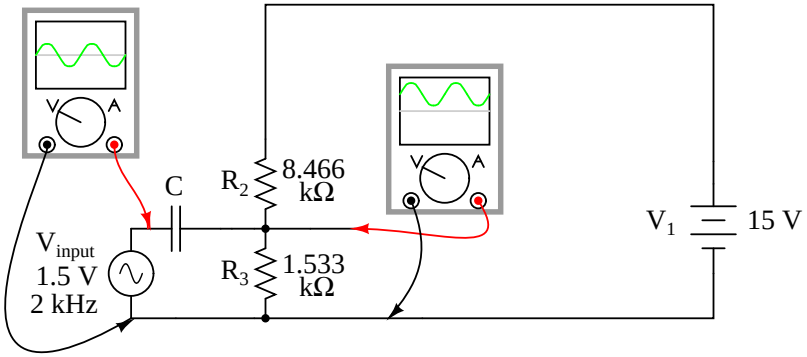
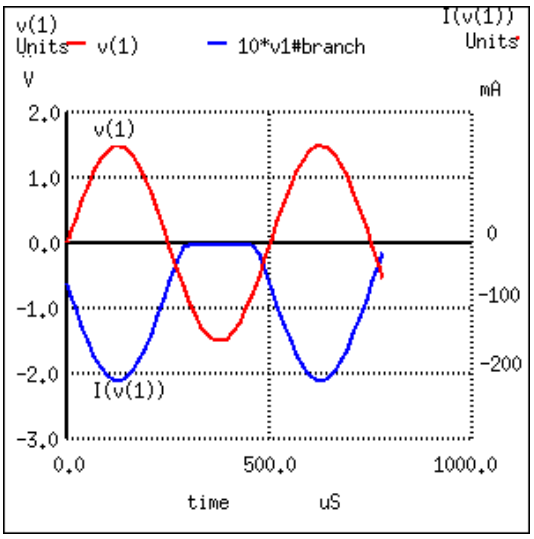


Figure 4.81: *Combined AC and DC circuit.*




```
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
```

Figure 4.82: *SPICE simulation of voltage divider bias.*

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 current source transistor model in Figure 4.83 to illustrate, the bias problem becomes evident.

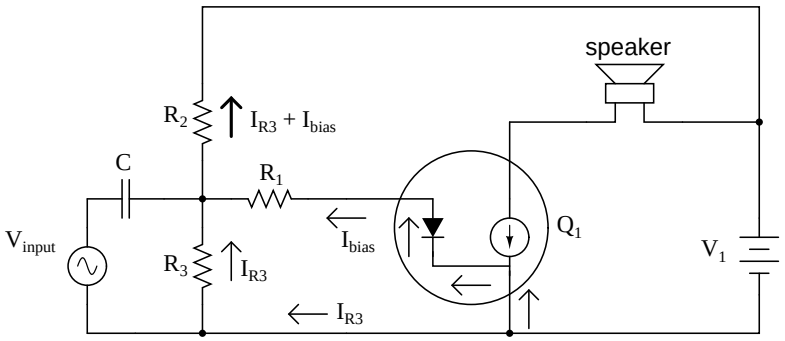
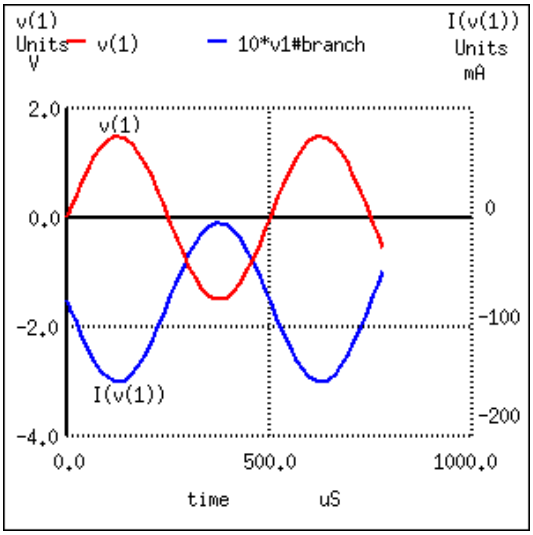


Figure 4.83: *Diode transistor model shows loading of voltage divider.*

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. 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. 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. To *increase* the DC voltage dropped across R_3, 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
r3 5 0 4k  <--- R3
increased to 4 k
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
```

Figure 4.84: *No distortion of the output after adjusting R2 and R3.*

The new resistor values of 6 kΩ and 4 kΩ (R_2 and R_3, respectively) in Figure 4.84 results in class A waveform reproduction, just the way we wanted.

- **REVIEW:**
- *Class A* operation is an amplifier biased to be in the active mode throughout the entire waveform cycle, thus faithfully reproducing the whole waveform.
- *Class B* operation is an amplifier 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 an amplifier is biased at a point somewhere between class A and class B.
- *Class C* is an amplifier biased to amplify only a small portion of the waveform. Most 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 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. Little heat energy dissipated makes energy efficiency 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.

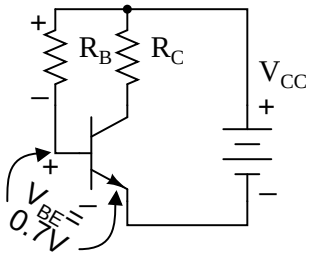
4.10 Biasing calculations

Although transistor switching circuits operate without bias, it is unusual for analog circuits to operate without bias. One of the few examples is "TR One, one transistor radio" (page 435) with an amplified AM (amplitude modulation) detector. Note the lack of a bias resistor at the base in that circuit. In this section we look at a few basic bias circuits which can set a selected emitter current I_E. Given a desired emitter current I_E, what values of bias resistors are required, R_B, R_E, etc?

4.10.1 Base Bias

The simplest biasing applies a *base-bias* resistor between the base and a base battery V_{BB}. It is convenient to use the existing V_{CC} supply instead of a new bias supply. An example of an audio amplifier stage using base-biasing is "Crystal radio with one transistor . . . " (page 435). Note the resistor from the base to the battery terminal. A similar circuit is shown in Figure 4.85.

Write a KVL (Krichhoff's voltage law) equation about the loop containing the battery, R_B, and the V_{BE} diode drop on the transistor in Figure 4.85. Note that we use V_{BB} for the base supply, even though it is actually V_{CC}. If β is large we can make the approximation that I_C =I_E. For silicon transistors $V_{BE} \cong 0.7V$.



$$V_{BB} - I_B R_B - V_{BE} = 0$$

$$V_{BB} - V_{BE} = I_B R_B \qquad \text{(KVL)}$$

$$I_B = \frac{V_{BB} - V_{BE}}{R_B}$$

$$I_E = (\beta+1) I_B \approx \beta I_B$$

$$\mathbf{I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta}} \qquad \text{(IE base-bias)}$$

Figure 4.85: *Base-bias*

Silicon small signal transistors typically have a β in the range of 100-300. Assuming that we have a β=100 transistor, what value of base-bias resistor is required to yield an emitter current of 1mA?

Solving the IE base-bias equation for R_B and substituting β, V_{BB}, V_{BE}, and I_E yields 930kΩ. The closest standard value is 910kΩ.

$\beta = 100 \qquad V_{BB} = 10V \qquad I_C \approx I_E = 1ma$

$$R_B = \frac{V_{BB} - V_{BE}}{I_E/\beta} = \frac{10 - 0.7}{1mA/100} = 930k$$

What is the emitter current with a 910kΩ resistor? What is the emitter current if we randomly get a β=300 transistor?

$\beta = 100 \qquad V_{BB} = 10V \qquad R_B = 910k \qquad V_{BE} = 0.7V$

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta} = \frac{10 - 0.7}{910k/100} = 1.02mA$$

$\beta = 300$

$$I_E = \frac{10 - 0.7}{910k/300} = 3.07mA$$

The emitter current is little changed in using the standard value 910kΩ resistor. However, with a change in β from 100 to 300, the emitter current has tripled. This is not acceptable in a power amplifier if we expect the collector voltage to swing from near V_{CC} to near ground. However, for low level signals from micro-volts to a about a volt, the bias point can be centered for a β of square root of (100·300)=173. The bias point will still drift by a considerable amount . However, low level signals will not be clipped.

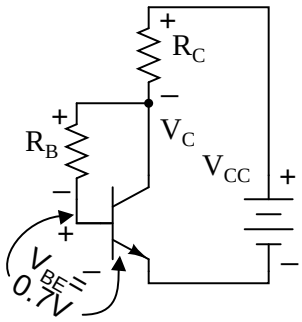
Base-bias by its self is not suitable for high emitter currents, as used in power amplifiers. The base-biased emitter current is not temperature stable. *Thermal run away* is the result of high emitter current causing a temperature increase which causes an increase in emitter current, which further increases temperature.

4.10.2 Collector-feedback bias

Variations in bias due to temperature and beta may be reduced by moving the V_{BB} end of the base-bias resistor to the collector as in Figure 4.86. If the emitter current were to increase, the voltage drop across R_C increases, decreasing V_C, decreasing I_B fed back to the base. This, in turn, decreases the emitter current, correcting the original increase.

Write a KVL equation about the loop containing the battery, R_C , R_B , and the V_{BE} drop. Substitute $I_C \cong I_E$ and $I_B \cong I_E/\beta$. Solving for I_E yields the IE CFB-bias equation. Solving for I_B yields the IB CFB-bias equation.

Find the required collector feedback bias resistor for an emitter current of 1 mA, a 4.7K collector load resistor, and a transistor with β=100 . Find the collector voltage V_C. It should be approximately midway between V_{CC} and ground.



$$I_C = \beta I_B \qquad I_C \approx I_E \qquad I_E \approx \beta I_B$$

$$V_{CC} - I_C R_C - I_B R_B - V_{BE} = 0 \qquad \text{(KVL)}$$

$$V_{CC} - I_E R_C - (I_E/\beta) R_B - V_{BE} = 0$$

$$V_{CC} - V_{BE} = I_E R_C + (I_E/\beta) R_B$$

$$V_{CC} - V_{BE} = I_E((R_B/\beta) + R_C)$$

$$\mathbf{I_E = \frac{V_{CC} - V_{BE}}{R_B/\beta + R_C}} \qquad \text{(IE CFB-bias)}$$

$$\mathbf{R_B = \beta\left[\frac{V_{CC} - V_{BE}}{I_E} - R_C\right]} \qquad \text{(RB CFB-bias)}$$

Figure 4.86: *Collector-feedback bias.*

$$\beta = 100 \qquad V_{CC} = 10V \qquad I_C \approx I_E = 1ma \qquad R_C = 4.7k$$

$$R_B = \beta\left[\frac{V_{CC} - V_{BE}}{I_E} - R_C\right] = 100\left[\frac{10 - 0.7}{1mA} - 4.7k\right] = 460k$$

$$V_C = V_{CC} - I_C R_C = 10 - (1mA)\cdot(4.7k) = 5.3V$$

The closest standard value to the 460k collector feedback bias resistor is 470k. Find the emitter current I_E with the 470 K resistor. Recalculate the emitter current for a transistor with β=100 and β=300.

$$\beta = 100 \qquad V_{CC} = 10V \qquad R_C = 4.7k \qquad R_B = 470k$$

$$I_E = \frac{V_{CC} - V_{BE}}{R_B/\beta + R_C} = \frac{10 - 0.7}{470k/100 + 4.7k} = 0.989mA$$

$$\beta = 300$$

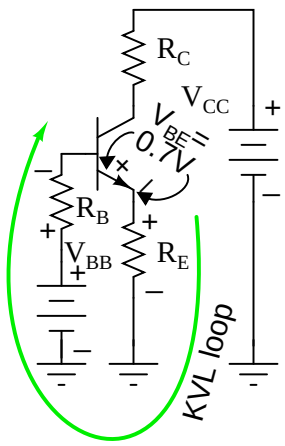
$$I_E = \frac{V_{CC} - V_{BE}}{R_B/\beta + R_C} = \frac{10 - 0.7}{470k/300 + 4.7k} = 1.48mA$$

We see that as beta changes from 100 to 300, the emitter current increases from 0.989mA to 1.48mA. This is an improvement over the previous base-bias circuit which had an increase from 1.02mA to 3.07mA. Collector feedback bias is twice as stable as base-bias with respect to beta variation.

4.10.3 Emitter-bias

Inserting a resistor R_E in the emitter circuit as in Figure 4.87 causes *degeneration*, also known as negative feedback. This opposes a change in emitter current I_E due to temperature changes, resistor tolerances, beta variation, or power supply tolerance. Typical tolerances are as follows: resistor— 5%, beta— 100-300, power supply— 5%. Why might the emitter resistor stabilize a

change in current? The polarity of the voltage drop across R_E is due to the collector battery V_{CC}. The end of the resistor closest to the (-) battery terminal is (-), the end closest to the (+) terminal it (+). Note that the (-) end of R_E is connected via V_{BB} battery and R_B to the base. Any increase in current flow through R_E will increase the magnitude of negative voltage applied to the base circuit, decreasing the base current, decreasing the emitter current. This decreasing emitter current partially compensates the original increase.



$$V_{BB} - I_B R_B - V_{BE} - I_E R_E = 0$$

$$I_E = (\beta+1) I_B \approx \beta I_B$$

$$V_{BB} - (I_E/\beta) R_B - V_{BE} - I_E R_E = 0$$

$$V_{BB} - V_{BE} = I_E ((R_B/\beta) + R_E)$$

$$\mathbf{I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E}} \qquad \text{(IE emitter-bias)}$$

$$R_B/\beta + R_E = \frac{V_{BB} - V_{BE}}{I_E}$$

$$\mathbf{R_B = \beta \left[\frac{V_{BB} - V_{BE}}{I_E} - R_E \right]} \qquad \text{(RB emitter-bias)}$$

Figure 4.87: *Emitter-bias*

Note that base-bias battery V_{BB} is used instead of V_{CC} to bias the base in Figure 4.87. Later we will show that the emitter-bias is more effective with a lower base bias battery. Meanwhile, we write the KVL equation for the loop through the base-emitter circuit, paying attention to the polarities on the components. We substitute $I_B \cong I_E/\beta$ and solve for emitter current I_E. This equation can be solved for R_B , equation: RB emitter-bias, Figure 4.87.

Before applying the equations: RB emitter-bias and IE emitter-bias, Figure 4.87, we need to choose values for R_C and R_E . R_C is related to the collector supply V_{CC} and the desired collector current I_C which we assume is approximately the emitter current I_E. Normally the bias point for V_C is set to half of V_{CC}. Though, it could be set higher to compensate for the voltage drop across the emitter resistor R_E. The collector current is whatever we require or choose. It could range from micro-Amps to Amps depending on the application and transistor rating. We choose I_C = 1mA, typical of a small-signal transistor circuit. We calculate a value for R_C and choose a close standard value. An emitter resistor which is 10-50% of the collector load resistor usually works well.

$V_C = V_{CC}/2 = 10/2 = 5V$

$R_C = Vc/I_C = 5/1mA = 5k$ (4.7k standard value)

$R_E = 0.10 R_C = 0.10(4.7K) = 470\Omega$

Our first example sets the base-bias supply to high at $V_{BB} = V_{CC} = 10V$ to show why a lower voltage is desirable. Determine the required value of base-bias resistor R_B. Choose a standard value resistor. Calculate the emitter current for β=100 and β=300. Compare the stabilization of the current to prior bias circuits.

$$\beta = 100 \quad I_E \approx I_C = 1ma \quad Vcc = V_{BB} = 10V \quad R_E = 470\Omega$$

$$R_B = \beta \left[\frac{V_{BB} - V_{BE}}{I_E} - R_E \right] = 100 \left[\frac{10 - 0.7}{0.001} - 470 \right] = 883k$$

An 883k resistor was calculated for R_B, an 870k chosen. At β=100, I_E is 1.01mA.

$$\beta = 100 \quad R_B = 870k$$

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} = \frac{10 - 0.7}{870K/100 + 470} = 1.01mA$$

$$\beta = 300$$

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} = \frac{10 - 0.7}{870K/300 + 470} = 2.76mA$$

For β=300 the emitter currents are shown in Table 4.7.

Table 4.7: *Emitter current comparison for β=100, β=300.*

Bias circuit	**IC β=100**	**IC β=300**
base-bias	1.02mA	3.07mA
collector feedback bias	0.989mA	1.48mA
emitter-bias, V_{BB}=10V	1.01mA	2.76mA

Table 4.7 shows that for V_{BB} = 10V, emitter-bias does not do a very good job of stabilizing the emitter current. The emitter-bias example is better than the previous base-bias example, but, not by much. The key to effective emitter bias is lowering the base supply V_{BB} nearer to the amount of emitter bias.

How much emitter bias do we Have? Rounding, that is emitter current times emitter resistor: $I_E R_E$ = (1mA)(470) = 0.47V. In addition, we need to overcome the V_{BE} = 0.7V. Thus, we need a $V_{BB} >$(0.47 + 0.7)V or >1.17V. If emitter current deviates, this number will change compared with the fixed base supply V_{BB},causing a correction to base current I_B and emitter current I_E. A good value for $V_B >$1.17V is 2V.

$$\beta = 100 \quad I_E \approx I_C = 1ma \quad Vcc = 10V \quad V_{BB} = 2V \quad R_E = 470\Omega$$

$$R_B = \beta \left[\frac{V_{BB} - V_{BE}}{I_E} - R_E \right] = 100 \left[\frac{2 - 0.7}{0.001} - 470 \right] = 83k$$

The calculated base resistor of 83k is much lower than the previous 883k. We choose 82k from the list of standard values. The emitter currents with the 82k R_B for β=100 and β=300 are:

$\beta = 100 \quad R_B = 82k$

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} = \frac{2 - 0.7}{82K/100 + 470} = 1.01mA$$

$\beta = 300$

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} = \frac{2 - 0.7}{82K/300 + 470} = 1.75mA$$

Comparing the emitter currents for emitter-bias with V_{BB} = 2V at β=100 and β=300 to the previous bias circuit examples in Table 4.8, we see considerable improvement at 1.75mA, though, not as good as the 1.48mA of collector feedback.

Table 4.8: *Emitter current comparison for β=100, β=300.*

Bias circuit	**IC β=100**	**IC β=300**
base-bias	1.02mA	3.07mA
collector feedback bias	0.989mA	1.48mA
emitter-bias, V_{BB}=10V	1.01mA	2.76mA
emitter-bias, V_{BB}=2V	1.01mA	1.75mA

How can we improve the performance of emitter-bias? Either increase the emitter resistor R_E or decrease the base-bias supply V_{BB} or both. As an example, we double the emitter resistor to the nearest standard value of 910Ω.

$\beta = 100 \quad I_E \approx I_C = 1ma \quad Vcc = 10V \quad V_{BB} = 2V \quad R_E = 910\Omega$

$$R_B = \beta \left[\frac{V_{BB} - V_{BE}}{I_E} - R_E \right] = 100 \left[\frac{2 - 0.7}{0.001} - 910 \right] = 39k$$

The calculated R_B = 39k is a standard value resistor. No need to recalculate I_E for β = 100. For β = 300, it is:

$\beta = 300 \quad R_B = 39k$

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} = \frac{2 - 0.7}{39K/300 + 910} = 1.25mA$$

The performance of the emitter-bias circuit with a 910¡Onega¿ emitter resistor is much improved. See Table 4.9.

As an exercise, rework the emitter-bias example with the emitter resistor reverted back to 470Ω, and the base-bias supply reduced to 1.5V.

$\beta = 100 \quad I_E \approx I_C = 1ma \quad Vcc = 10V \quad V_{BB} = 1.5V \quad R_E = 470\Omega$

$$R_B = \beta \left[\frac{V_{BB} - V_{BE}}{I_E} - R_E \right] = 100 \left[\frac{1.5 - 0.7}{0.001} - 470 \right] = 33k$$

Table 4.9: *Emitter current comparison for β=100, β=300.*

Bias circuit	**IC β=100**	**IC β=300**
base-bias	1.02mA	3.07mA
collector feedback bias	0.989mA	1.48mA
emitter-bias, V_{BB}=10V	1.01mA	2.76mA
emitter-bias, V_{BB}=2V, R_E=470	1.01mA	1.75mA
emitter-bias, V_{BB}=2V, R_E=910	1.00mA	1.25mA

The 33k base resistor is a standard value, emitter current at β = 100 is OK. The emitter current at β = 300 is:

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} = \frac{1.5 - 0.7}{33K/300 + 470} = 1.38mA$$

Table 4.10 below compares the exercise results 1mA and 1.38mA to the previous examples.

Table 4.10: *Emitter current comparison for β=100, β=300.*

Bias circuit	**IC β=100**	**IC β=300**
base-bias	1.02mA	3.07mA
collector feedback bias	0.989mA	1.48mA
emitter-bias, V_{BB}=10V	1.01mA	2.76mA
emitter-bias, V_{BB}=2V, R_B=470	1.01mA	1.75mA
emitter-bias, V_{BB}=2V, R_B=910	1.00mA	1.25mA
emitter-bias, V_{BB}=1.5V, R_B=470	1.00mA	1.38mA

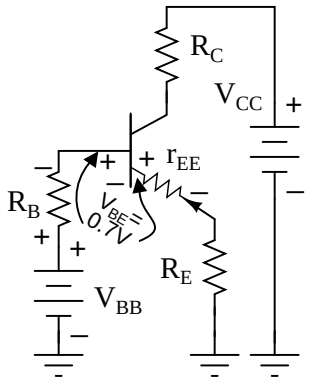
The emitter-bias equations have been repeated in Figure 4.88 with the internal emitter resistance included for better accuracy. The *internal emitter resistance* is the resistance in the emitter circuit contained within the transistor package. This internal resistance r_{EE} is significant when the (external) emitter resistor R_E is small, or even zero. The value of internal resistance R_{EE} is a function of emitter current I_E, Table 4.11.

Table 4.11: *Derivation of r_{EE}*

r_{EE} = KT/I_Em
where:
K=1.38×10^{-23} watt-sec/°C, Boltzman's constant
T= temperature in Kelvins ≅300.
I_E = emitter current
m = varies from 1 to 2 for Silicon
r_{EE} ≅ 0.026V/I_E = 26mV/I_E

For reference the 26mV approximation is listed as equation r_{EE} in Figure 4.88.

The more accurate emitter-bias equations in Figure 4.88 may be derived by writing a KVL equation. Alternatively, start with equations IE emitter-bias and R_B emitter-bias in Figure 4.87, substituting R_E with r_{EE}+R_E. The result is equations IE EB and RB EB, respectively



$$V_{BB} - I_B R_B - V_{BE} - I_E r_{EE} - I_E R_E = 0 \quad \text{(KVL)}$$

$$I_E = (\beta + 1) I_B \approx \beta I_B$$

$$V_{BB} - (I_E/\beta) R_B - V_{BE} - I_E r_{EE} - I_E R_E = 0$$

$$V_{BB} - V_{BE} = (I_E (R_B/\beta) + I_E r_{EE} + I_E R_E)$$

$$\mathbf{I_E} = \frac{\mathbf{V_{BB} - V_{BE}}}{\mathbf{R_B}/\beta + \mathbf{r_{EE}} + \mathbf{R_E}} \quad \text{(IE EB)}$$

$$R_B/\beta + R_{EE} + R_E = \frac{V_{BB} - V_{BE}}{I_E}$$

$$\mathbf{R_B} = \beta \left[\frac{\mathbf{V_{BB} - V_{BE}}}{\mathbf{I_E}} - \mathbf{r_{EE}} - \mathbf{R_E} \right] \quad \text{(RB EB)}$$

$$\mathbf{r_{EE} = 26mV/I_E} \quad \text{(rEE)}$$

Figure 4.88: *Emitter-bias equations with internal emitter resistance r_{EE} included..*

in Figure 4.88.

Redo the R_B calculation in the previous example (page 239) with the inclusion of r_{EE} and compare the results.

$\beta = 100$ $I_E \approx I_C = 1ma$ $Vcc = 10V$ $V_{BB} = 2V$ $R_E = 470\Omega$

$r_{EE} = 26mV/1mA = 26\Omega$

$$R_B = \beta \left[\frac{Vcc - V_{BE}}{I_E} - r_{EE} - R_E \right] = 100 \left[\frac{2.0 - 0.7}{0.001} - 26 - 470 \right] = 80.4k$$

The inclusion of r_{EE} in the calculation results in a lower value of the base resistor R_B a shown in Table 4.12. It falls below the standard value 82k resistor instead of above it.

Table 4.12: *Effect of inclusion of r_{EE} on calculated R_B*

r_{EE}?	**r_{EE} Value**
Without r_{EE}	83k
With r_{EE}	80.4k

Bypass Capacitor for R_E

One problem with emitter bias is that a considerable part of the output signal is dropped across the emitter resistor R_E (Figure 4.89). This voltage drop across the emitter resistor is in series with the base and of opposite polarity compared with the input signal. (This is similar to a common collector configuration having <1 gain.) This degeneration severely reduces the gain from base to collector. The solution for AC signal amplifiers is to bypass the emitter resistor with a capacitor. This restores the AC gain since the capacitor is a short for AC signals. The

DC emitter current still experiences degeneration in the emitter resistor, thus, stabilizing the DC current.

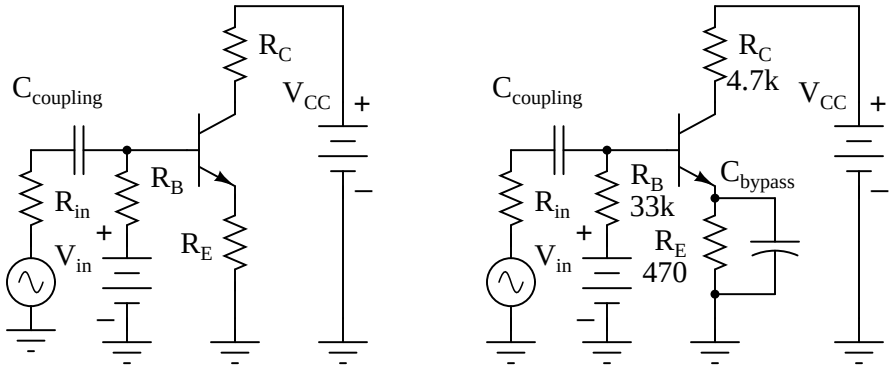


Figure 4.89: *Cbypass is required to prevent AC gain reduction.*

What value should the bypass capacitor be? That depends on the lowest frequency to be amplified. For radio frequencies Cbpass would be small. For an audio amplifier extending down to 20Hz it will be large. A "rule of thumb" for the bypass capacitor is that the reactance should be 1/10 of the emitter resistance or less. The capacitor should be designed to accommodate the lowest frequency being amplified. The capacitor for an audio amplifier covering 20Hz to 20kHz would be:

$$X_C = \frac{1}{2\pi fC}$$

$$C = \frac{1}{2\pi fX_C}$$

$$C = \frac{1}{2\pi 20(470/10)} = 169\mu F$$

Note that the internal emitter resistance r_{EE} is not bypassed by the bypass capacitor.

4.10.4 Voltage divider bias

Stable emitter bias requires a low voltage base bias supply, Figure 4.90. The alternative to a base supply V_{BB} is a voltage divider based on the collector supply V_{CC}.

The design technique is to first work out an emitter-bias design, Then convert it to the voltage divider bias configuration by using Thevenin's Theorem. [4] The steps are shown graphically in Figure 4.91. Draw the voltage divider without assigning values. Break the divider loose from the base. (The base of the transistor is the load.) Apply Thevenin's Theorem to yield a single Thevenin equivalent resistance Rth and voltage source Vth.

The Thevenin equivalent resistance is the resistance from load point (arrow) with the battery (V_{CC}) reduced to 0 (ground). In other words, R1||R2.The Thevenin equivalent voltage is the open circuit voltage (load removed). This calculation is by the voltage divider ratio method.

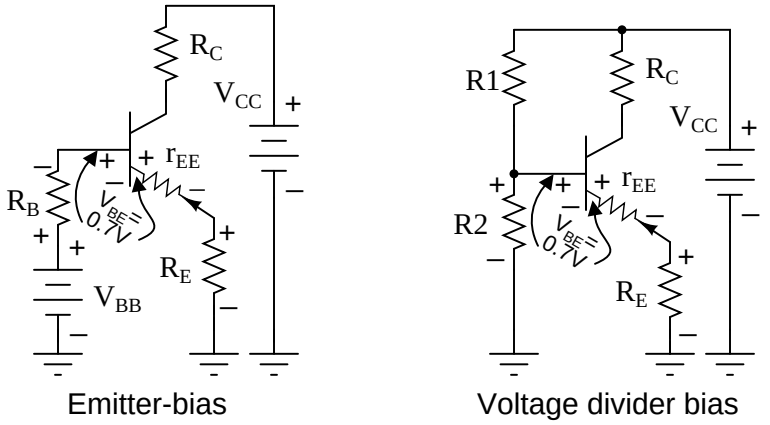


Figure 4.90: *Voltage Divider bias replaces base battery with voltage divider.*

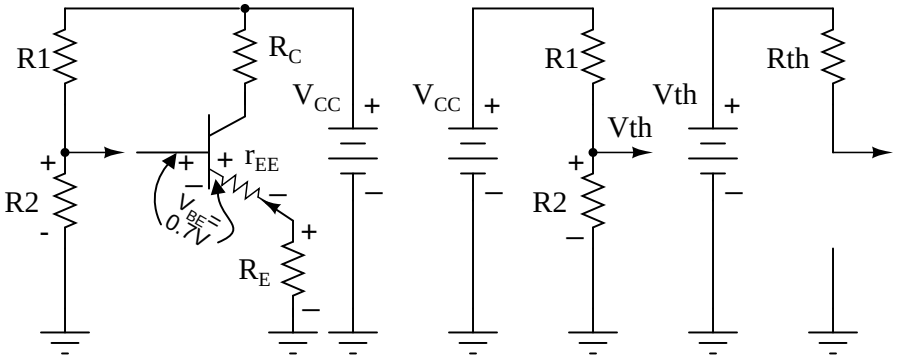


Figure 4.91: *Thevenin's Theorem converts voltage divider to single supply Vth and resistance Rth.*

R1 is obtained by eliminating R2 from the pair of equations for Rth and Vth. The equation of R1 is in terms of known quantities Rth, Vth, Vcc. Note that Rth is R_B , the bias resistor from the emitter-bias design. The equation for R2 is in terms of R1 and Rth.

$$\text{Rth} = \text{R1} \parallel \text{R2} \qquad \text{Vth} = V_{CC}\left[\frac{\text{R2}}{\text{R1}+\text{R2}}\right]$$

$$\frac{1}{\text{Rth}} = \frac{1}{\text{R1}} + \frac{1}{\text{R2}} \qquad f = \frac{\text{Vth}}{V_{CC}} = \left[\frac{\text{R2}}{\text{R1}+\text{R2}}\right]$$

$$\frac{1}{\text{Rth}} = \frac{\text{R2}+\text{R1}}{\text{R1}\cdot\text{R2}} = \frac{1}{\text{R1}}\left[\frac{\text{R2}+\text{R1}}{\text{R2}}\right] = \frac{1}{\text{R1}}\cdot\frac{1}{f}$$

$$\text{R1} = \frac{\text{Rth}}{f} = \text{Rth}\,\frac{V_{CC}}{\text{Vth}} \qquad \frac{1}{\text{R2}} = \frac{1}{\text{Rth}} - \frac{1}{\text{R1}}$$

Convert this previous emitter-bias example to voltage divider bias.

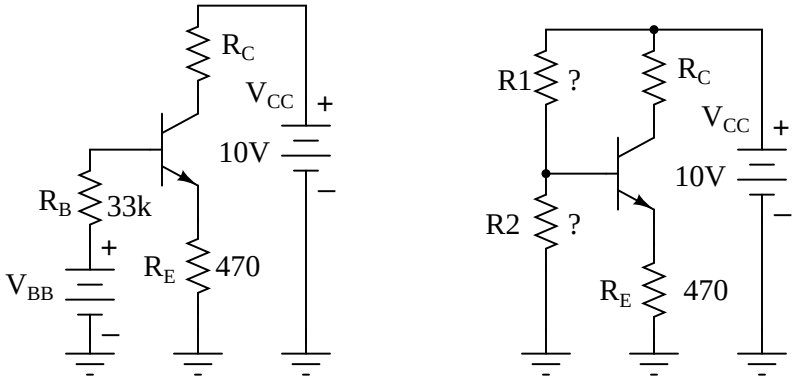


Figure 4.92: *Emitter-bias example converted to voltage divider bias.*

These values were previously selected or calculated for an emitter-bias example

$$\beta = 100 \qquad I_E \approx I_C = 1\text{ma} \qquad \text{Vcc} = 10\text{V} \qquad V_{BB} = 1.5\text{V} \qquad R_E = 470\Omega$$

$$R_B = \beta\left[\frac{V_{BB} - V_{BE}}{I_E} - R_E\right] = 100\left[\frac{1.5 - 0.7}{0.001} - 470\right] = 33\text{k}$$

Substituting V_{CC} , V_{BB} , R_B yields R1 and R2 for the voltage divider bias configuration.

$V_{BB} = Vth = 1.5V$

$R_B = Rth = 33k$

$$R1 = Rth \frac{V_{CC}}{Vth}$$

$$R1 = 33k \frac{10}{1.5} = 220k$$

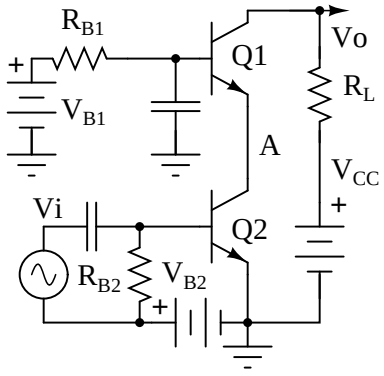
$$\frac{1}{R2} = \frac{1}{Rth} - \frac{1}{R1}$$

$$\frac{1}{R2} = \frac{1}{33k} - \frac{1}{220k}$$

$R2 = 38.8k$

R1 is a standard value of 220K. The closest standard value for R2 corresponding to 38.8k is 39k. This does not change I_E enough for us to calculate it.

Problem: Calculate the bias resistors for the cascode amplifier in Figure 4.93. V_{B2} is the bias voltage for the common emitter stage. V_{B1} is a fairly high voltage at 11.5 because we want the common-base stage to hold the emitter at 11.5-0.7=10.8V, about 11V. (It will be 10V after accounting for the voltage drop across R_{B1} .) That is, the common-base stage is the load, substitute for a resistor, for the common-emitter stage's collector. We desire a 1mA emitter current.



Cascode

$V_{CC} = 20V \quad I_E = 1mA \quad \beta = 100 \quad V_A = 10V \quad R_L = 4.7k$

$V_{BB1} = 11.5V \quad V_{BB2} = 1.5V$

$$I_E = \frac{V_{BB} - V_{BE}}{R_B/\beta} \quad \text{(IE base-bias)}$$

$$R_{B1} = \frac{V_{BB} - V_{BE}}{I_E/\beta} = \frac{(V_{BB1} - V_A) - V_{BE}}{I_E/\beta} = \frac{(11.5\text{-}10) - 0.7}{1mA/100} = 80k$$

$$R_{B2} = \frac{V_{BB2} - V_{BE}}{I_E/\beta} = \frac{(1.5) - 0.7}{1mA/100} = 80k$$

Figure 4.93: *Bias for a cascode amplifier.*

Problem: Convert the base bias resistors for the cascode amplifier to voltage divider bias resistors driven by the V_{CC} of 20V.

$R_{BB1} = 80k$ $V_{CC} = Vth = 20V$ $R_{BB2} = 80k$

$V_{BB1} = 11.5V$ $V_{BB2} = 1.5V$

$V_{BB} = Vth = 11.5V$ $V_{BB} = Vth = 1.5V$

$R_B = Rth = 80k$ $R_B = Rth = 80k$

$R1 = Rth \frac{V_{CC}}{Vth}$ $R3 = Rth \frac{V_{CC}}{Vth}$

$R1 = 80k \frac{20}{11.5} = 139.1k$ $R3 = 80k \frac{20}{1.5} = 1.067Meg$

$\frac{1}{R2} = \frac{1}{Rth} - \frac{1}{R1}$ $\frac{1}{R4} = \frac{1}{Rth} - \frac{1}{R3}$

$\frac{1}{R2} = \frac{1}{80k} - \frac{1}{139.1k}$ $\frac{1}{R4} = \frac{1}{80k} - \frac{1}{1067k}$

$R2 = 210k$ $R4 = 86.5k$

The final circuit diagram is shown in the “Practical Analog Circuits” chapter, “Class A cascode amplifier . . . ” (page 441).

- **REVIEW:**
- See Figure 4.94.
- Select bias circuit configuration
- Select R_C and I_E for the intended application. The values for R_C and I_E should normally set collector voltage V_C to 1/2 of V_{CC}.
- Calculate base resistor R_B to achieve desired emitter current.
- Recalculate emitter current I_E for standard value resistors if necessary.
- For voltage divider bias, perform emitter-bias calculations first, then determine R1 and R2.
- For AC amplifiers, a bypass capacitor in parallel with R_E improves AC gain. Set $X_C \leq 0.10R_E$ for lowest frequency.

4.11 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.

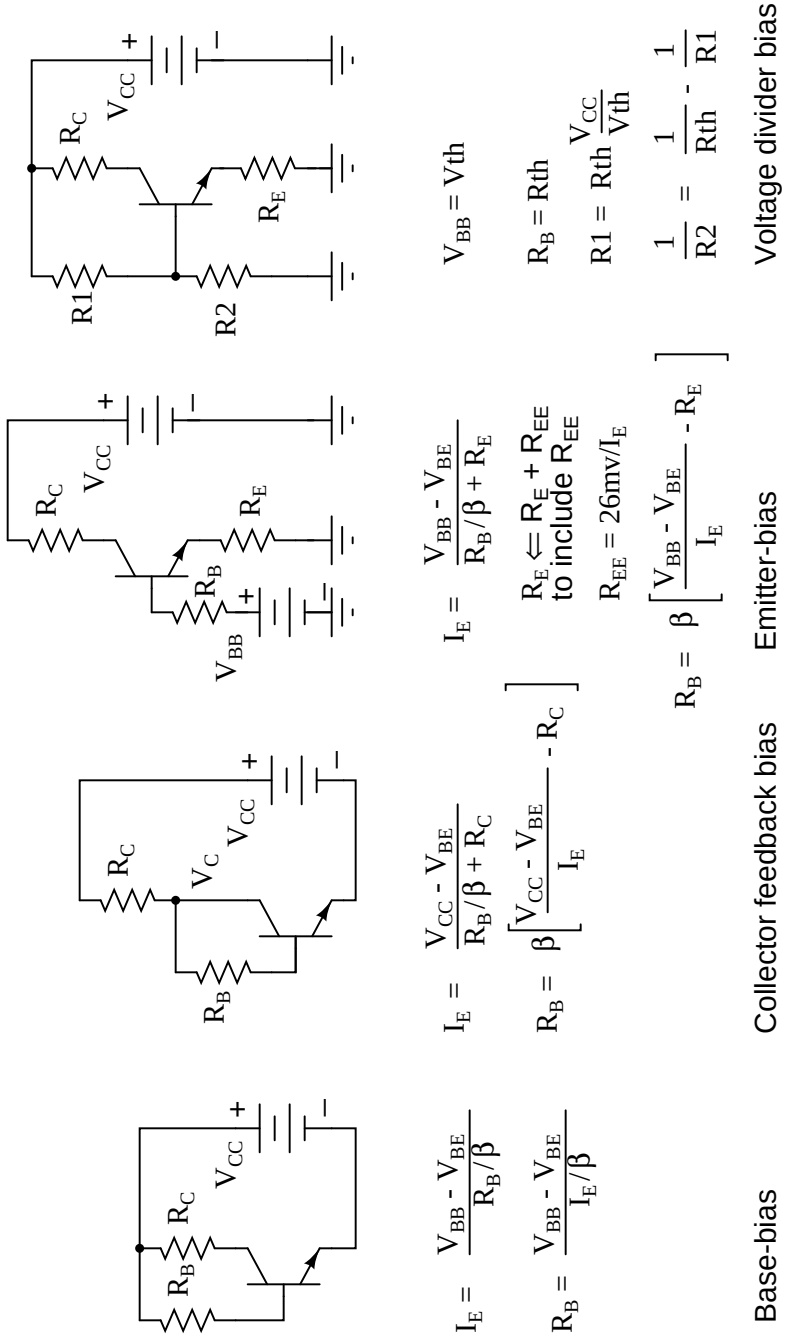


Figure 4.94: *Biasing equations summary.*

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 in Figure 4.95.

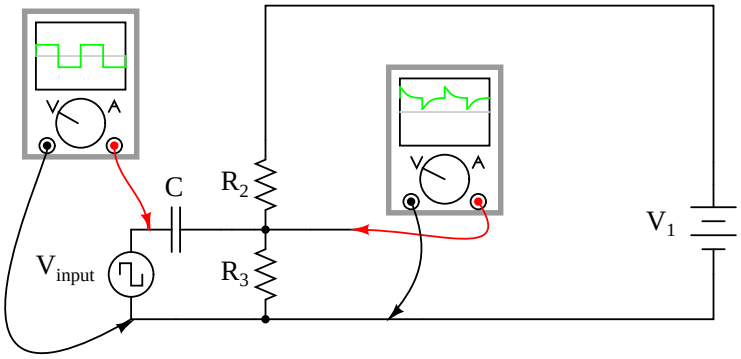


Figure 4.95: *Capacitively coupled low frequency square-wave shows distortion.*

Incidentally, this same problem occurs when oscilloscope inputs are set to the "AC coupling" mode as in Figure 4.97. 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, or contains considerable levels of harmonics over a wide frequency range, the oscilloscope's display of the waveform will not be accurate. (Figure 4.97) Low frequency signals may be viewed by setting the oscilloscope to "DC coupling" in Figure 4.96.

In applications where the limitations of capacitive coupling (Figure 4.95) 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 is shown in Figure 4.98.

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

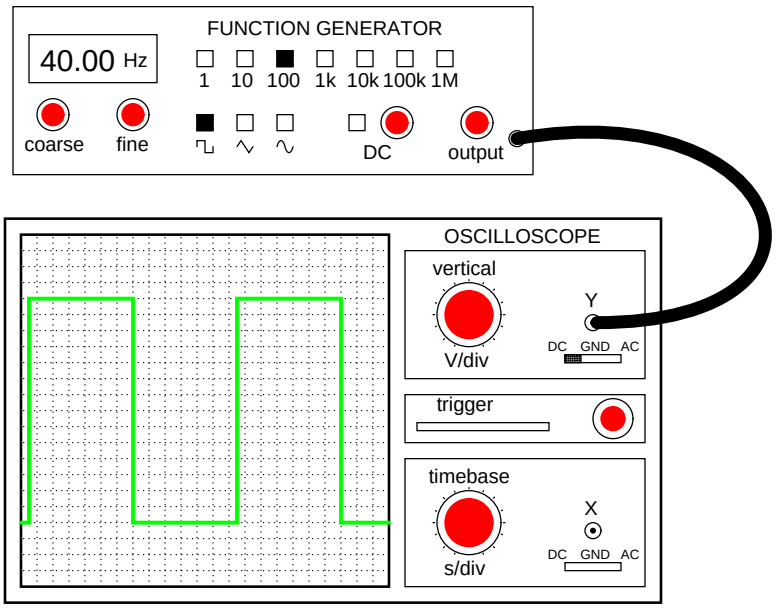


Figure 4.96: *With DC coupling, the oscilloscope properly indicates the shape of the square wave coming from the signal generator.*

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 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 offsets the cone from its natural center position, and this limits the back-and-forth motion it can sustain from the applied AC voltage without overtraveling. However, our example circuit (Figure 4.98) applies a varying voltage of only *one* polarity across the speaker,

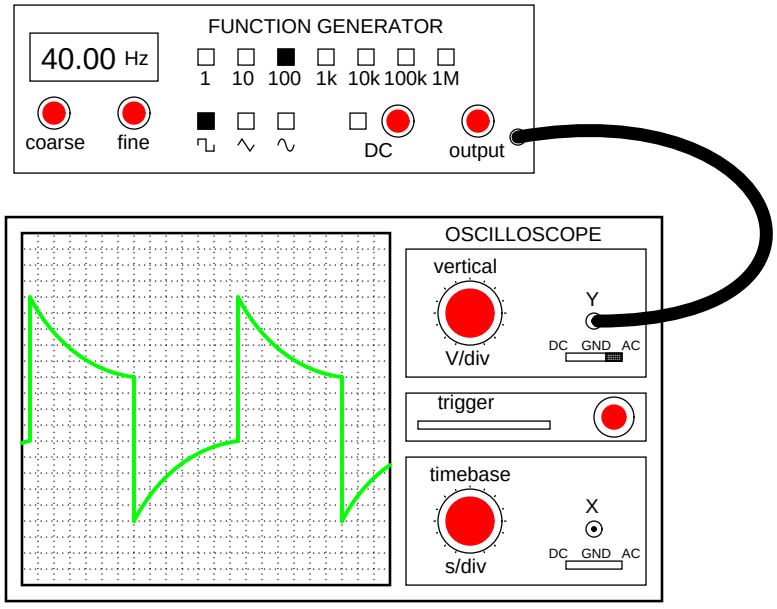


Figure 4.97: *Low frequency: With AC coupling, the high-pass filtering of the coupling capacitor distorts the square wave's shape so that what is seen is not an accurate representation of the real signal.*

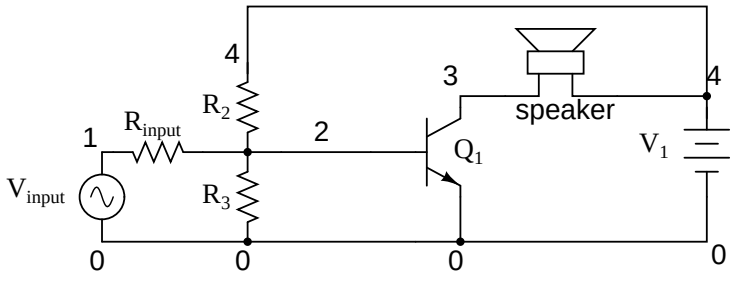


Figure 4.98: *Direct coupled amplifier: direct coupling to speaker.*

because the speaker is connected in series with the transistor which can only conduct current one way. This would be unacceptable for 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 in Figure 4.99)

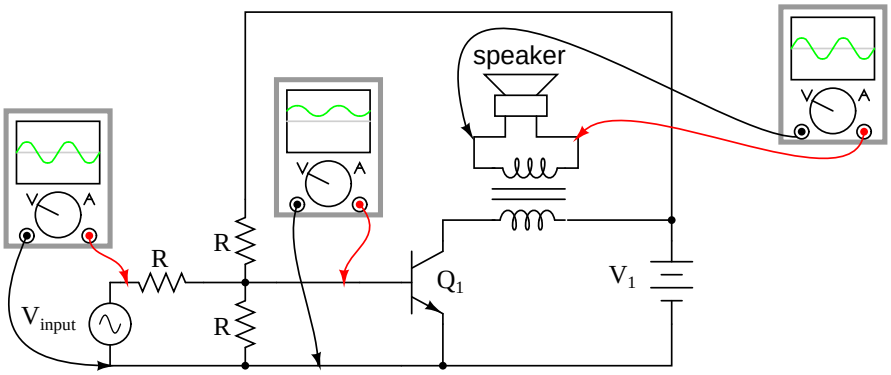


Figure 4.99: *Transformer coupling isolates DC from the load (speaker).*

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. See the coupling transformer between Q4 and the speaker, (page 435) as an example of transformer coupling.

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 (Figure 4.100) to the amplifier.

This circuit in Figure 4.100 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

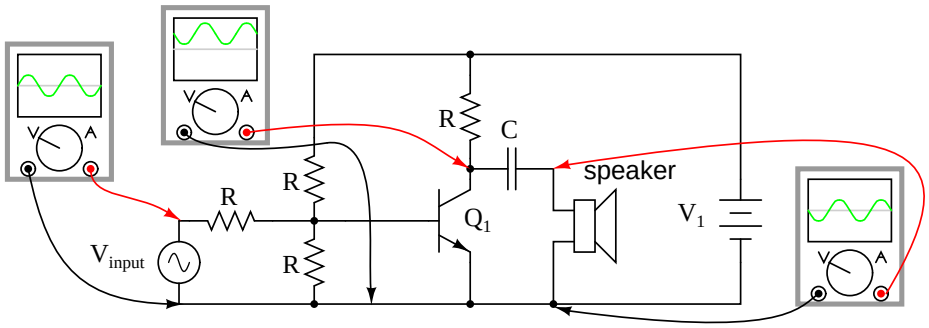


Figure 4.100: *Capacitor coupling isolates DC from the load.*

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 as in Figure 4.101.

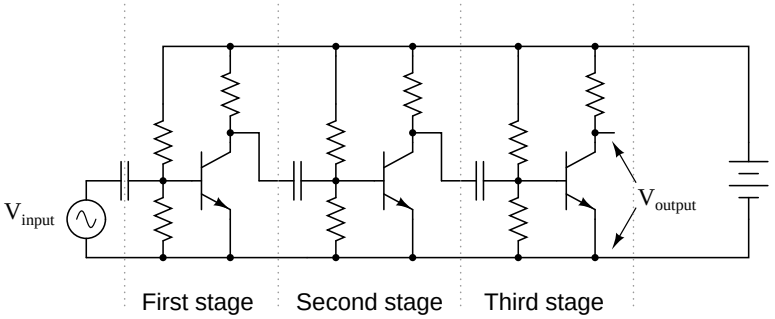


Figure 4.101: *Capacitor coupled three stage common-emitter amplifier.*

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 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 radio-frequency amplifiers (Figure 4.102) with small coupling transformers, having air cores (making them immune to saturation effects), that are part of a resonant circuit to block unwanted harmonic frequencies from passing on to subsequent stages. The use of

resonant circuits assumes that the signal frequency remains constant, which is typical of radio circuitry. Also, the "flywheel" effect of LC tank circuits allows for class C operation for high efficiency.

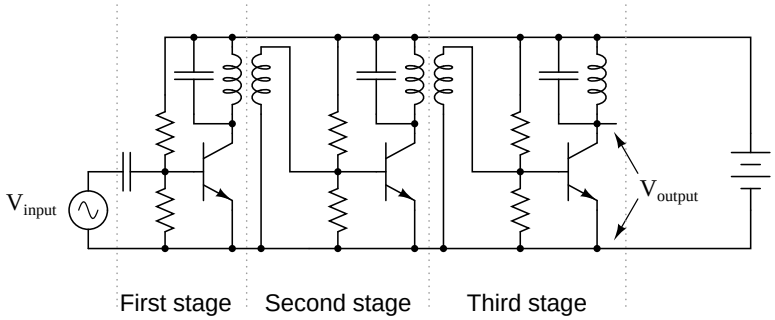


Figure 4.102: *Three stage tuned RF amplifier illustrates transformer coupling.*

Note the transformer coupling between transistors Q1, Q2, Q3, and Q4, (page 435). The three intermediate frequency (IF) transformers within the dashed boxes couple the IF signal from collector to base of following transistor IF amplifiers. The *intermediate freqency* ampliers are RF amplifiers, though, at a different frequency than the antenna RF input.

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.

The trend of electronics to more widespread use of integrated circuits has encouraged the use of direct coupling over transformer or capacitor coupling. The only easily manufactured integrated circuit component is the transistor. Moderate quality resistors can also be produced. Though, transistors are favored. Integrated capacitors to only a few 10's of pF are possible. Large capacitors are not integrable. If necessary, these can be external components. The same is true of transformers. Since integrated transistors are inexpensive, as many transistors as possible are substituted for the offending capacitors and transformers. As much direct coupled gain as possible is designed into ICs between the external coupling components. While external capacitors and transformers are used, these are even being designed out if possible. The result is that a modern IC radio (See "IC radio", (page 438)) looks nothing like the original 4-transistor radio (page 435).

Even discrete transistors are inexpensive compared with transformers. Bulky audio transformers can be replaced by transistors. For example, a common-collector (emitter follower) configuration can impedance match a low output impedance like a speaker. It is also possible to replace large coupling capacitors with transistor circuitry.

We still like to illustrate texts with transformer coupled audio amplifiers. The circuits are simple. The component count is low. And, these are good introductory circuits— easy to understand.

The circuit in Figure 4.103 (a) is a simplified transformer coupled *push-pull* audio amplifier. In push-pull, pair of transistors alternately amplify the positive and negative portions of the input signal. Neither transistor nor the other conducts for no signal input. A positive input

signal will be positive at the top of the transformer secondary causing the top transistor to conduct. A negative input will yield a positive signal at the bottom of the secondary, driving the bottom transistor into conduction. Thus the transistors amplify alternate halves of a signal. As drawn, neither transistor in Figure 4.103 (a) will conduct for an input below 0.7 Vpeak. A practical circuit connects the secondary center tap to a 0.7 V (or greater) resistor divider instead of ground to bias both transistor for true class B.

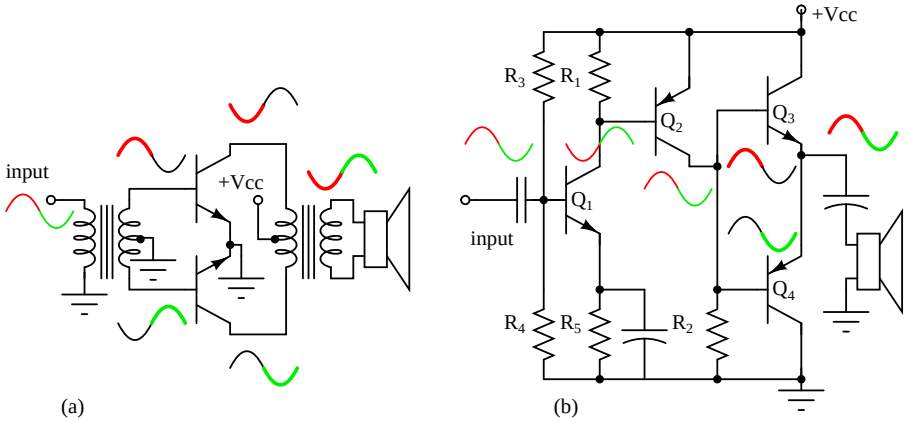


Figure 4.103: *(a) Transformer coupled push-pull amplifier. (b) Direct coupled complementary-pair amplifier replaces transformers with transistors.*

The circuit in Figure 4.103 (b) is the modern version which replaces the transformer functions with transistors. Transistors Q_1 and Q_2 are common emitter amplifiers, inverting the signal with gain from base to collector. Transistors Q_3 and Q_4 are known as a *complementary pair* because these NPN and PNP transistors amplify alternate halves (positive and negative, respectively) of the waveform. The parallel connection the bases allows phase splitting without an input transformer at (a). The speaker is the emitter load for Q_3 and Q_4. Parallel connection of the emitters of the NPN and PNP transistors eliminates the center-tapped output transformer at (a) The low output impedance of the emitter follower serves to match the low 8 Ω impedance of the speaker to the preceding common emitter stage. Thus, inexpensive transistors replace transformers. For the complete circuit see“ Direct coupled complementary symmetry 3 w audio amplifier,” (page 433)

- **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.

4.12 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. See "Phase shift oscillator" (page 433) for a practical application of positive feedback.

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 distorts the input waveform less 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 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, starting with Figure 4.104.

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

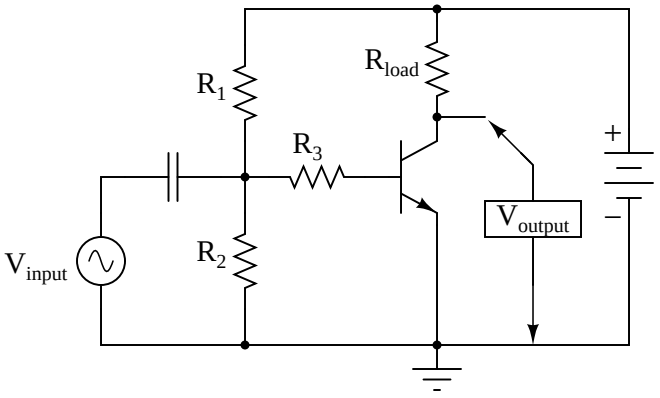


Figure 4.104: *Common-emitter amplifier without feedback.*

the purpose of controlling voltage gain. We could omit it 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 move toward negative, and vice versa. The oscilloscope waveforms are shown in Figure 4.105.

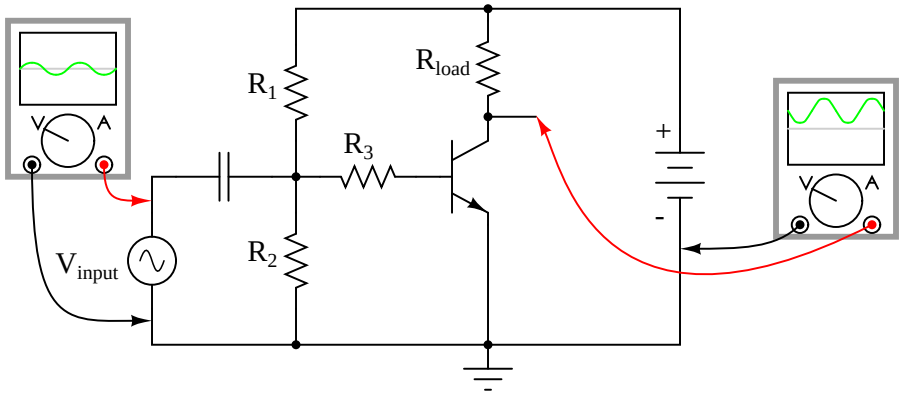


Figure 4.105: *Common-emitter amplifier, no feedback, with reference waveforms for comparison.*

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 in Figure 4.106 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 (with respect to ground) is a weighted average of the input voltage and the feedback voltage, resulting in signal of reduced amplitude

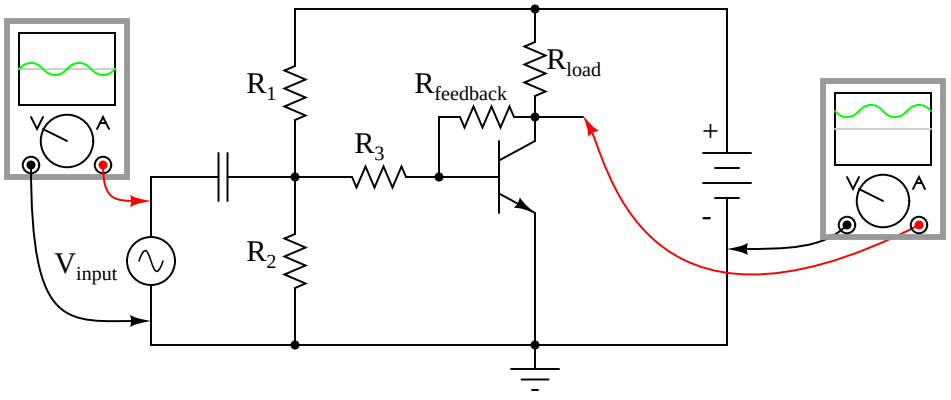


Figure 4.106: *Negative feedback, collector feedback, decreases the output signal.*

going into the transistor. So, the amplifier circuit in Figure 4.106 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 in Figure 4.107.

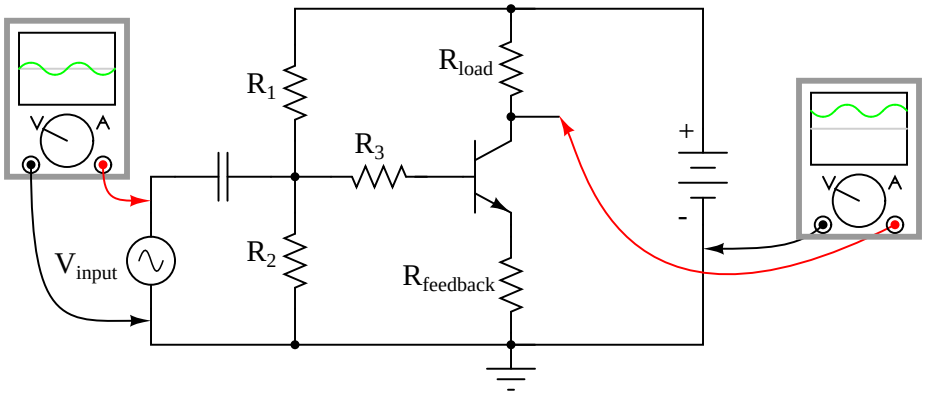


Figure 4.107: *Emitter feedback: A different method of introducing negative feedback into a circuit.*

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 in Figure 4.108.

With no feedback resistor connecting the emitter to ground in Figure 4.108 (a) , 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} 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 in Figure 4.108 (b).

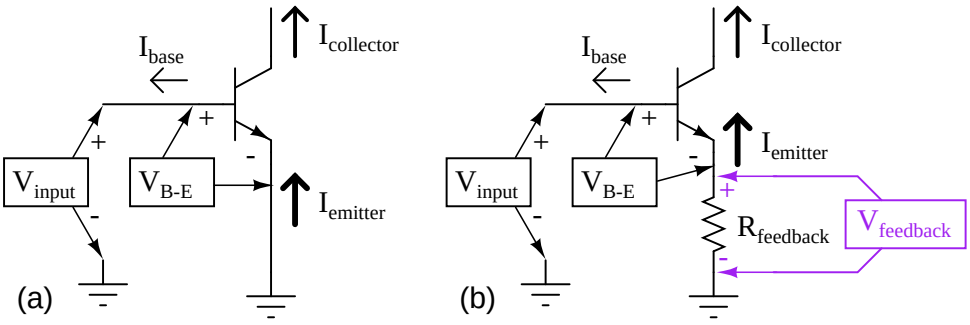


Figure 4.108: *(a) No feedback vs (b) emitter feedback. A waveform at the collector is inverted with respect to the base. At (b) the emitter waveform is in-phase (emitter follower) with base, out of phase with collector. Therefore, the emitter signal subtracts from the collector output signal.*

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; its 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. 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, (Figure 4.109) however, do not suffer from thermal runaway. Why is this? The answer has everything to do with negative feedback.

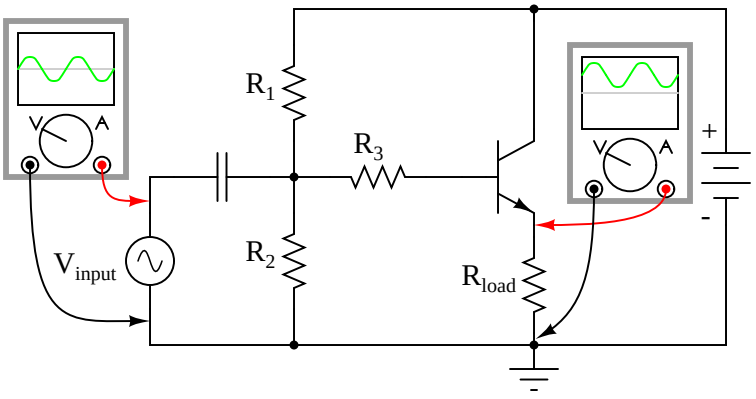


Figure 4.109: *Common collector (emitter follower) amplifier.*

Note that the common-collector amplifier (Figure 4.109) has its load resistor placed in exactly the same spot as we had the $R_{feedback}$ resistor in the last circuit in Figure 4.108 (b): 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. Is it possibe 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 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: its 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 only function with DC. 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 in Figure 4.110, we create the very situation we need: a path from emitter to ground that is easier for AC than it is for DC.

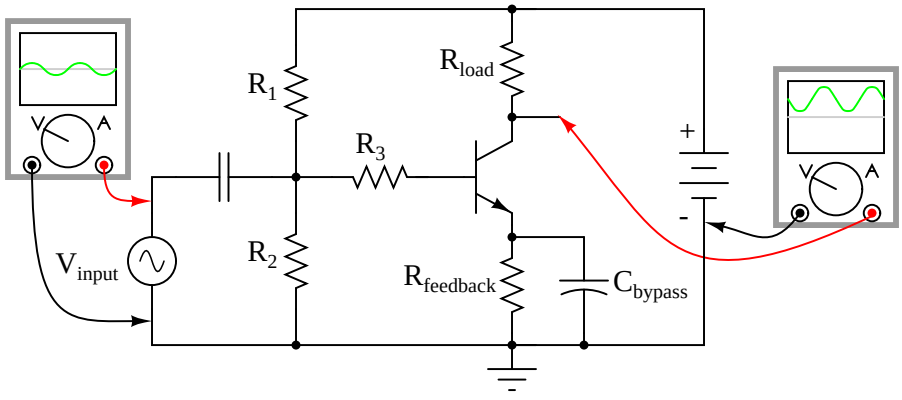


Figure 4.110: *High AC voltage gain reestablished by adding C_{bypass} in parallel with $R_{feedback}$*

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. An example circuit showing negative feedback in a three-stage common-emitter amplifier is Figure 4.111.

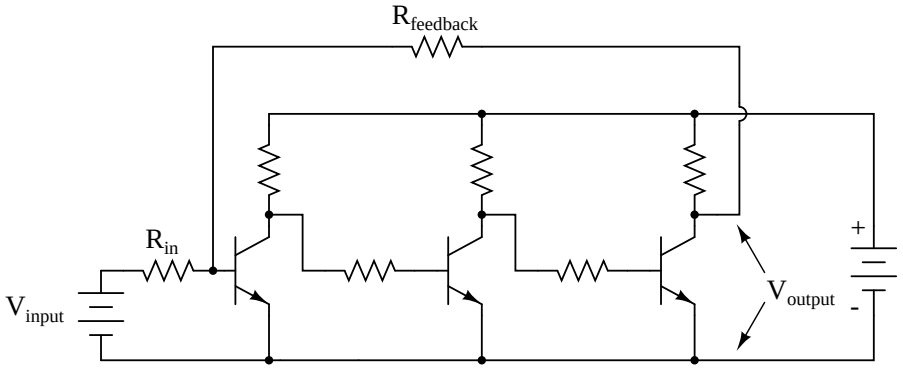


Figure 4.111: *Feedback around an "odd" number of direct coupled stages produce negative feedback.*

The feedback path from the final output to the input is through a single resistor, $R_{feedback}$. Since each stage is a common-emitter amplifier (thus inverting), the odd number of stages from input to output will invert the output signal; the feedback will be negative (degenerative). Relatively large amounts of feedback may be used without sacrificing voltage gain, because the three amplifier stages provide 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 is an advantage for mass-production of electronic circuitry: if amplifiers of predictable gain may be constructed using transistors of widely varied β values, it eases the selection and replacement of components. 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 much 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 little effect on voltage gain with feedback in operation, resulting in an amplifier that is stable and easy to design.

4.13 Amplifier impedances

Input impedance varies considerably with the circuit configuration shown in Figure 4.112. It also varies with biasing. Not considered here, the input impedance is complex and varies with frequency. For the common-emitter and common-collector it is base resistance times β. The base resistance can be both internal and external to the transistor. For the common-collector:

$$R_{in} = \beta R_E$$

It is a bit more complicated for the common-emitter circuit. We need to know the internal emitter resistance r_{EE}. This is given by:

r_{EE} = KT/I_Em
where:
K=1.38×10^{-23} watt-sec/^oC, Boltzman's constant
T= temperature in Kelvins ≅300.
I_E = emitter current
m = varies from 1 to 2 for Silicon
R_E ≅ 0.026V/I_E = 26mV/I_E

Thus, for the common-emitter circuit Rin is

Rin = βr_{EE}

As an example the input resistance of a, β = 100, CE configuration biased at 1 mA is:

r_{EE} = 26mV/1mA = 26Ω

Rin = βr_{EE} = 100(26) = 2600Ω

Moreover, a more accurate Rin for the common-collector should have included r_{EE}

Rin = β(R_E + r_{EE})

This equation (above) is also applicable to a common-emitter configuration with an emitter resistor.

Input impedance for the common-base configuration is Rin = r_{EE}.

The high input impedance of the common-collector configuration matches high impedance sources. A crystal or ceramic microphone is one such high impedance source. The common-base arrangement is sometimes used in RF (radio frequency) circuits to match a low impedance source, for example, a 50 Ω coaxial cable feed. For moderate impedance sources, the common-emitter is a good match. An example is a dynamic microphone.

The output impedances of the three basic configurations are listed in Figure 4.112. The moderate output impedance of the common-emitter configuration helps make it a popular choice for general use. The Low output impedance of the common-collector is put to good use in impedance matching, for example, tranformerless matching to a 4 Ohm speaker. There do not appear to be any simple formulas for the output impedances. However, R. Victor Jones develops expressions for output resistance. [3]

- **REVIEW:**
- See Figure 4.112.

4.14 Current mirrors

An 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

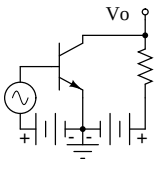
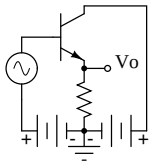
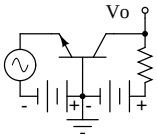
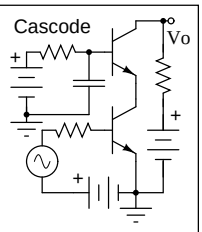
Basic circuit	Common emitter	Common collector	Common base	Cascode
				
Voltage gain	high	less than unity	high, same as CE	high, same as CB
Current gain	high	high	less than unity	high, same as CE
Power gain	high	moderate	moderate	highest
Phase inversion	yes	no	no	yes
Input impedance	moderate ≈ 1 k	highest ≈ 300 k	low ≈ 50 Ω	same as CE, ≈ 1 k
Output impedance	moderate ≈ 50 k	low ≈ 300 Ω	highest ≈ 1 Meg	same as CB, ≈ 1 Meg

Figure 4.112: *Amplifier characteristics, adapted from GE Transistor Manual, Figure 1.21.[2]*

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 Kelvins

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

transistor constant, then its emitter current will be constant, given a constant temperature. (Figure 4.113)

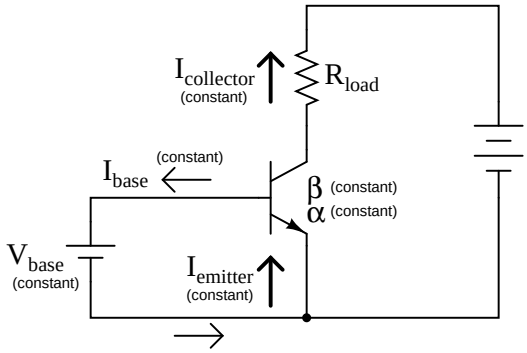


Figure 4.113: *Constant V_{BE} gives constant I_B, constant I_E, and constant I_C.*

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

To maintain a constant voltage across the transistor's base-emitter junction 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 as in Figure 4.114.

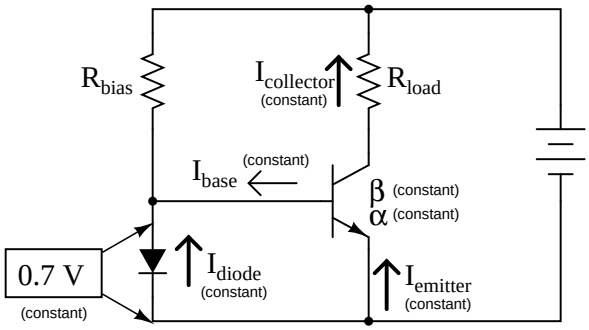


Figure 4.114: *Diode junction 0.7 V maintains constant base voltage, and constant base current.*

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}. 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, as in Figure 4.115 (a).

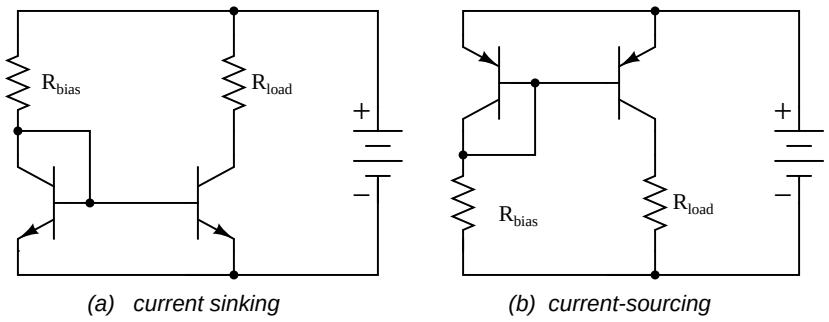


Figure 4.115: *Current mirror circuits.*

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 close to one another to facilitate heat transfer between them.

The current mirror circuit shown with two NPN transistors in Figure 4.115 (a) 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 may use PNP transistors like Figure 4.115 (b).

While resistors can be manufactured in ICs, it is easier to fabricate transistors. IC designers avoid some resistors by replacing load resistors with current sources. A circuit like an operational amplifier built from discrete components will have a few transistors and many resistors.

An integrated circuit version will have many transistors and a few resistors. In Figure 4.116 One voltage reference, Q1, drives multiple current sources: Q2, Q3, and Q4. If Q2 and Q3 are equal area transistors the load currents I_{load} will be equal. If we need a $2 \cdot I_{load}$, parallel Q2 and Q3. Better yet fabricate one transistor, say Q3 with twice the area of Q2. Current I3 will then be twice I2. In other words, load current scales with transistor area.

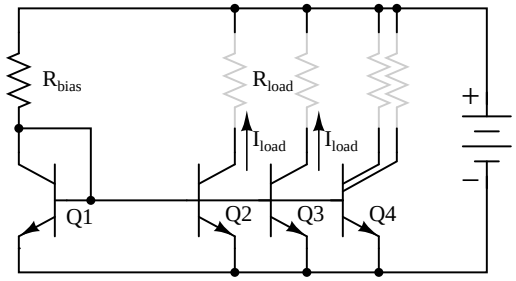


Figure 4.116: *Multiple current mirrors may be slaved from a single (Q1 - R_{bias}) voltage source.*

Note that it is customary to draw the base voltage line right through the transistor symbols for multiple current mirrors! Or in the case of Q4 in Figure 4.116, two current sources are associated with a single transistor symbol. The load resistors are drawn almost invisible to emphasize the fact that these do not exist in most cases. The load is often another (multiple) transistor circuit, say a pair of emitters of a differential amplifier, for example Q3 and Q4 in "A simple operational amplifier" (page 408). Often, the collector load of a transistor is not a resistor but a current mirror. For example the collector load of Q4 collector (page 408) is a current mirror (Q2).

For an example of a current mirror with multiple collector outputs see Q13 in the model 741 op-amp (page 408). The Q13 current mirror outputs substitute for resistors as collector loads for Q15 and Q17. We see from these examples that current mirrors are preferred as loads over resistors in integrated circuitry.

- **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 do 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.

4.15 Transistor ratings and packages

Like all electrical and electronic components, transistors are limited in the amounts of voltage and current each one can handle without sustaining damage. Since transistors are more complex than some of the other components you're used to seeing at this point, these 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 for how many watts each 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 referenced to the temperature of ambient (surrounding) air. When transistors are to be used in hotter environments ($>25_o$, 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 emitter-base junction V_{EB} , collector-base junction V_{CB} , and also from collector to emitter V_{CE} .

V_{EB} , the maximum reverse voltage from emitter to base is approximately 7 V for some small signal transistors. Some circuit designers use discrete BJTs as 7 V zener diodes with a series current limiting resistor. Transistor inputs to analog integrated circuits also have a V_{EB} rating, which if exceeded will cause damage, no zenering of the inputs is allowed.

The rating for maximum collector-emitter voltage V_{CE} can be thought of as 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. A typical value for a small signal transistor is 60 to 80 V. In power transistors, this could range to 1000 V, for example, a horizontal deflection transistor in a cathode ray tube display.

Collector current: A maximum value for collector current I_C will be given by the manufacturer in amps. Typical values for small signal transistors are 10s to 100s of mA, 10s of A for power transistors. 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. 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. Low voltage transistors, low V_{CE} , show lower saturation voltages. The saturation voltage is also lower for higher base drive current.

Base-emitter forward voltage drop, kV_{BE} , is similar to that of an equivalent diode, $\cong$0.7 V, which should come as no surprise.

Beta: The ratio of collector current to base current, β is the fundamental parameter char-

acterizing 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. h_{fe} is small signal AC gain; h_{FE} is large AC signal gain or DC gain.

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

Bipolar transistors come in a wide variety of physical packages. Package type is primarily dependent upon the required power dissipation of the transistor, much like resistors: the greater the maximum power dissipation, the larger the device has to be to stay cool. Figure 4.117 shows several standardized package types for three-terminal semiconductor devices, any of which may be used to house a bipolar transistor. There are many other semiconductor devices other than bipolar transistors which have three connection points. Note that the pinouts of plastic transistors can vary within a single package type, e.g. TO-92 in Figure 4.117. It is *impossible* to positively identify a three-terminal semiconductor device without referencing the part number printed on it, or subjecting it to a set of electrical tests.

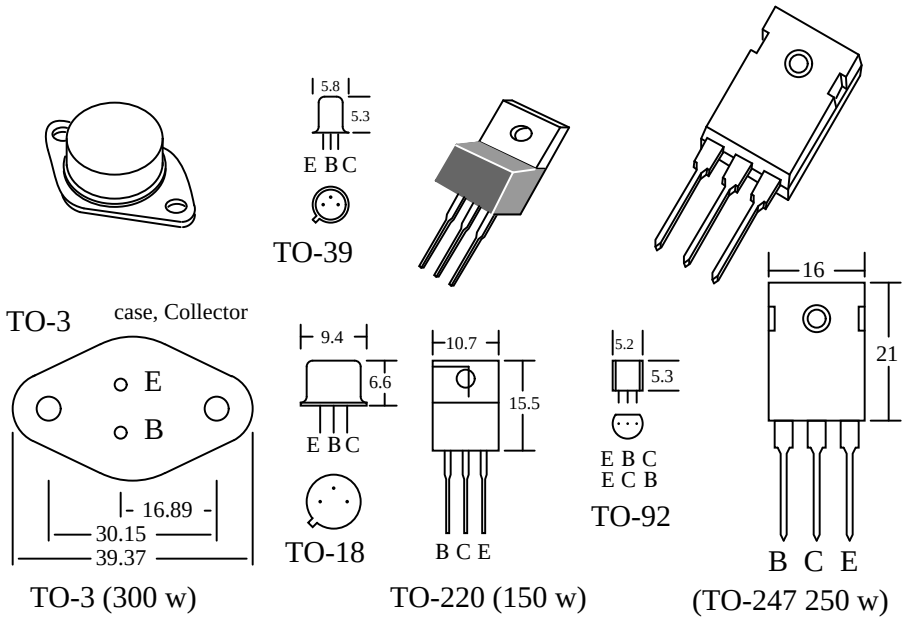


Figure 4.117: *Transistor packages, dimensions in mm.*

Small plastic transistor packages like the TO-92 can dissipate a few hundred milliwatts. The metal cans, TO-18 and TO-39 can dissipate more power, several hundred milliwatts. Plas-

tic power transistor packages like the TO-220 and TO-247 dissipate well over 100 watts, approaching the dissipation of the all metal TO-3. The dissipation ratings listed in Figure 4.117 are the maximum ever encountered by the author for high powered devices. Most power transistors are rated at half or less than the listed wattage. Consult specific device datasheets for actual ratings. The semiconductor die in the TO-220 and TO-247 plastic packages is mounted to a heat conductive metal slug which transfers heat from the back of the package to a metal *heatsink*, not shown. A thin coating of thermally conductive grease is applied to the metal before mounting the transistor to the heatsink. Since the TO-220 and TO-247 slugs, and the TO-3 case are connected to the collector, it is sometimes necessary to electrically isolate these from a grounded heatsink by an interposed mica or polymer washer. The datasheet ratings for the power packages are only valid when mounted to a heatsink. Without a heatsink, a TO-220 dissipates approximately 1 watt safely in free air.

Datasheet maximum power disipation ratings are difficult to acheive in practice. The maximum power dissipation is based on a heatsink maintaining the transistor case at no more than 25^{o}C. This is difficult with an air cooled heatsink. The allowable power dissipation decreases with increasing temperature. This is known as derating. Many power device datasheets include a dissipation versus case termperaure graph.

- **REVIEW:**
- *Power dissipation*: maximum allowable power dissipation on a sustained basis.
- *Reverse voltages*: maximum allowable V_{CE} , V_{CB} , V_{EB} .
- *Collector current*: the maximum allowable collector current.
- *Saturation voltage* is the V_{CE} voltage drop in a saturated (fully conducting) transistor.
- *Beta*: β=I_C/I_B
- *Alpha*: α=I_C/I_E α= β/(β+1)
- Transistor*Packages* are a major factor in power dissipation. Larger packages dissipate more power.

4.16 BJT quirks

An ideal transistor would show 0% distortion in amplifying a signal. Its gain would extend to all frequencies. It would control hundreds of amperes of current, at hundreds of degrees C. In practice, available devices show distortion. Amplification is limited at the high frequency end of the spectrum. Real parts only handle tens of amperes with precautions. Care must be taken when paralleling transistors for higher current. Operation at elevated temperatures can destroy transistors if precautions are not taken.

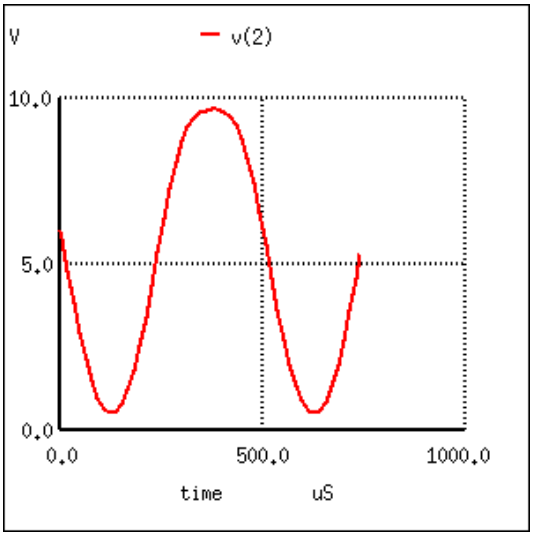


Figure 4.118: *Distortion in large signal common-emitter amplifier.*

4.16.1 Nonlinearity

The class A common-emitter amplifier (similar to Figure 4.34)is driven almost to clipping in Figure 4.118 . Note that the positive peak is flatter than the negative peaks. This distortion is unacceptable in many applications like high-fidelity audio.

Small signal amplifiers are relatively linear because they use a small linear section of the transistor characteristics. Large signal amplifiers are not 100% linear because transistor characteristics like β are not constant, but vary with collector current. β is high at low collector current, and low at very low current or high current. Though, we primarily encounter decreasing β with increasing collector current.

The SPICE listing in Table 4.119 illustrates how to quantify the amount of distortion. The ".fourier 2000 v(2)" command tells SPICE to perm a fourier analysis at 2000 Hz on the output v(2). At the command line "spice -b circuitname.cir" produces the Fourier analysis output in Table 4.119. It shows THD (total harmonic distortion) of over 10%, and the contribution of the individual harmonics.

A partial solution to this distortion is to decrease the collector current or operate the amplifier over a smaller portion of the load line. The ultimate solution is to apply negative feedback. See (page 256).

4.16.2 Temperature drift

Temperature affects the AC and DC characteristics of transistors. The two aspects to this problem are environmental temperature variation and self-heating. Some applications, like military and automotive, require operation over an extended temperature range. Circuits in a benign environment are subject to self-heating, in particular high power circuits.

```
common-emitter amplifier
Vbias 4 0 0.74
Vsig  5 4 sin (0 125m 2000 0 0)
rbias 6 5 2k
q1 2 6 0 q2n2222
r 3 2 1000
v1 3 0 dc 10
.model q2n2222 npn (is=19f bf=150
+ vaf=100 ikf=0.18 ise=50p ne=2.5
+ br=7.5
+ var=6.4 ikr=12m isc=8.7p nc=1.2
+ rb=50
+ re=0.4 rc=0.3 cje=26p tf=0.5n
+ cjc=11p tr=7n xtb=1.5 kf=0.032f
+ af=1)
.fourier 2000 v(2)
.tran 0.02m 0.74m
.end
```

```
spice -b ce.cir
Fourier analysis
v(2):
THD: 10.4688
Har Freq  Norm Mag
--- ----  --------
0   0     0
1   2000  1
2   4000  0.097992
3   6000  0.036541
4   8000  0.004387
5  10000  0.001158
6  12000  0.000893
7  14000  0.000211
8  16000  3.81e-05
9  18000  3.37e-05
```

Figure 4.119: *SPICE net list: for transient and fourier analyses. Fourier analysis shows 10% total harmonic distortion (THD).*

Leakage current I_{CO} and β increase with temperature. The DC β h_{FE} increases exponentially. The AC β h_{fe} increases, but not as rapidly. It doubles over the range of -55^o to 85^o C. As temperature increases, the increase in h_{fe} will yield a larger common-emitter output, which could be clipped in extreme cases. The increase in h_{FE} shifts the bias point, possibly clipping one peak. The shift in bias point is amplified in multi-stage direct-coupled amplifiers. The solution is some form of negative feedback to stabilize the bias point. This also stabilizes AC gain.

Increasing temperature in Figure 4.120 (a) will decrease V_{BE} from the nominal 0.7V for silicon transistors. Decreasing V_{BE} increases collector current in a common-emitter amplifier, further shifting the bias point. The cure for shifting V_{BE} is a pair of transistors configured as a differential amplifier. If both transistors in Figure 4.120 (b) are at the same temperature, the V_{BE} will track with changing temperature and cancel.

The maximum recommended junction temperature for silicon devices is frequently 125^o C. Though, this should be derated for higher reliability. Transistor action ceases beyond 150^o C. Silicon carbide and diamond transistors will operate considerably higher.

4.16.3 Thermal runaway

The problem with increasing temperature causing increasing collector current is that more current increase the power dissipated by the transistor which, in turn, increases its temperature. This self-reinforcing cycle is known as *thermal run away*, which may destroy the transistor. Again, the solution is a bias scheme with some form of negative feedback to stabilize the bias point.

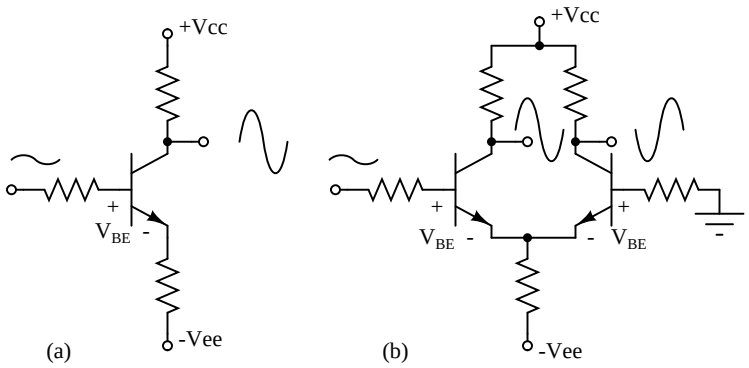


Figure 4.120: *(a) single ended CE amplifier vs (b) differential amplifier with V_{BE} cancellation.*

4.16.4 Junction capacitance

Capacitance exists between the terminals of a transistor. The collector-base capacitance C_{CB} and emitter-base capacitance C_{EB} decrease the gain of a common emitter circuit at higher frequencies.

In a common emitter amplifier, the capacitive feedback from collector to base effectively multiplies C_{CB} by β. The amount of negative gain-reducing feedback is related to both current gain, and amount of collector-base capacitance. This is known as the *Miller effect*, (page 277).

4.16.5 Noise

The ultimate sensitivity of small signal amplifiers is limited by noise due to random variations in current flow. The two major sources of noise in transistors are *shot noise* due to current flow of carriers in the base and *thermal noise*. The source of thermal noise is device resistance and increases with temperature:

$V_n = \sqrt{4kTRB_n}$

where

k = boltzman's conatant ($1.38 \bullet 10^{-23}$ watt-sec/K)
T = resistor tempeature in kelvins
R = resistance in Ohms
B_n = noise bandwidth in Hz

Noise in a transistor amplifier is defined in terms of *excess noise* generated by the amplifier, not that noise amplified from input to output, but that generated within the amplifier. This is determined by measuring the *signal to noise ratio* (S/N) at the amplifier input and output. The AC voltage output of an amplifier with a small signal input corresponds to S+N, signal plus noise. The AC voltage with no signal in corresponds to noise N. The *noise figure* F is defined in terms of S/N of amplifier input and output:

$$F = \frac{(S/N)_i}{(S/N)_o}$$

$$F_{dB} = 10 \log F$$

The noise figure F for RF (radio frequency) transistors is usually listed on transistor data sheets in decibels, F_{dB}. A good VHF (very high frequency, 30 MHz to 300 Mhz) noise figure is < 1 dB. The noise figure above VHF increases considerable, 20 dB per decade as shown in Figure 4.121.

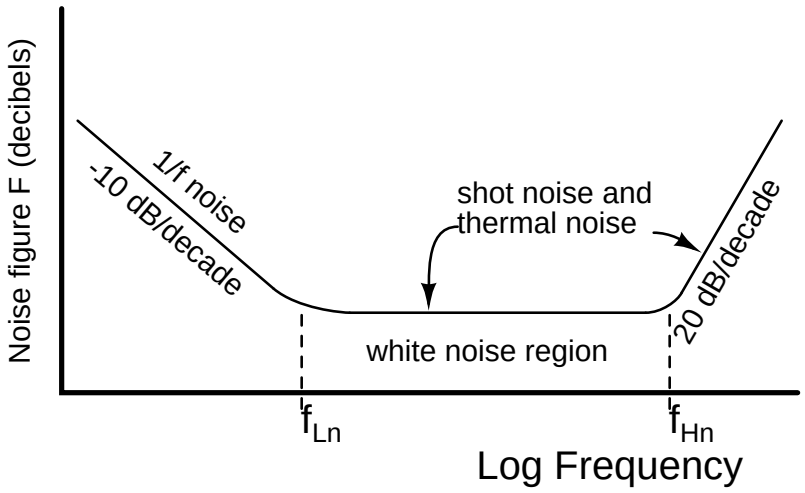


Figure 4.121: *Small signal transistor noise figure vs Frequency. After Thiele, Figure 11.147 [1]*

Figure 4.121 also shows that noise at low frequencies increases at 10 dB per decade with decreasing frequency. This noise is known as *1/f noise*.

Noise figure varies with the transistor type (part number). Small signal RF transistors used at the antenna input of a radio receiver are specifically designed for low noise figure. Noise figure varies with bias current and impedance matching. The best noise figure for a transistor is achieved at lower bias current, and possibly with an impedance mismatch.

4.16.6 Thermal mismatch (problem with paralleling transistors)

If two identical power transistors were paralleled for higher current, one would expect them to share current equally. Because of differences in characteristerics, transistors do not share current equally.

It is not practical to select identical transistors. The β for small signal transistors typically has a range of 100-300, power transistors: 20-50. If each one could be matched, one still might run hotter than the other due to environmental conditions. The hotter transistor draws more current resulting in thermal runaway. The solution when paralleling bipolar transistors is to insert emitter resistors known as *ballast resistors* of less than an ohm. If the hotter transistor draws more current, the voltage drop across the ballast resistor increases— negative feedback.

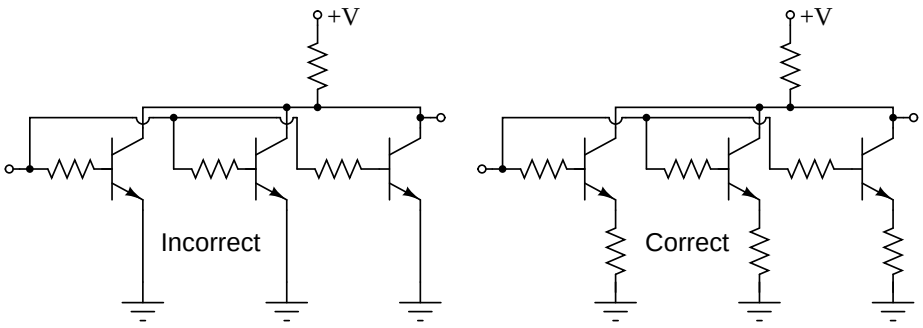


Figure 4.122: *Transistors paralleled for increased power require emitter ballast resistors*

This decreases the current. Mounting all transistors on the same heatsink helps equalize current too.

4.16.7 High frequency effects

The performance of a transistor amplifier is relatively constant, up to a point, as shown by the small signal common-emitter current gain with increasing frequency in Figure 4.123. Beyond that point the performance of a transistor degrades as frequency increases.

Beta cutoff frequency, f_T is the frequency at which common-emitter small signal current gain (h_{fe}) falls to unity. (Figure 4.123) A practical amplifier must have a gain >1. Thus, a transistor cannot be used in a practical amplifier at f_T. A more usable limit for a transistor is $0.1 \cdot f_T$.

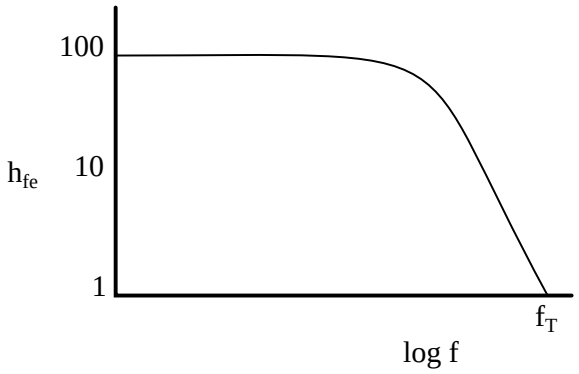


Figure 4.123: *Common-emitter small signal current gain (h_{fe}) vs frequency.*

Some RF silicon bipolar transistors are usable as amplifers up to a few GHz. Silicon-germanium devices extend the upper range to 10 GHz.

Alpha cutoff frequency, f_{alpha} is the frequency at which the α falls to 0.707 of low frequency $\alpha_{,0}$ α=0.707α_0. Alpha cutoff and beta cutoff are nearly equal: $f_{alpha} \cong f_T$ Beta cutoff f_T

is the preferred figure of merit of high frequency performance.

f$_{max}$ is the highest frequency of oscillation possible under the most favorable conditions of bias and impedance matching. It is the frequency at which the power gain is unity. All of the output is fed back to the input to sustain oscillations. f$_{max}$ is an upper limit for frequency of operation of a transistor as an active device. Though, a practical amplifier would not be usable at f$_{max}$.

Miller effect: The high frequency limit for a transistor is related to the junction capacitances. For example a PN2222A has an input capacitance C$_{obo}$=9pF and an output capacitance C$_{ibo}$=25pF from C-B and E-B respectively. [5] Although the C-E capacitance of 25 pF seems large, it is less of a factor than the C-B (9pF) capacitance. because of the *Miller effect*, the C-B capacitance has an effect on the base equivalent to beta times the capacitance in the common-emitter amplifier. Why might this be? A common-emitter amplifier inverts the signal from base to collector. The inverted collector signal fed back to the base opposes the input on the base. The collector signal is beta times larger than the input. For the PN2222A, β=50–300. Thus, the 9pF C-E capacitance looks like 9·50=450pF to 9·300=2700pF.

The solution to the junction capacitance problem is to select a high frequency transistor for wide bandwidth applications— RF (radio frequency) or microwave transistor. The bandwidth can be extended further by using the common-base instead of the common-emitter configuration. The grounded base shields the emitter input from capacitive collector feedback. A two-transistor cascode arrangement will yield the same bandwidth as the common-base, with the higher input impedance of the common-emitter.

- **REVIEW:**
- Transistor amplifiers exhibit distortion because of β variation with collector current.
- I$_c$, V$_{BE}$, β and junction capacitance vary with temperature.
- An increase in temperature can cause an increase in I$_C$, causing an increase in temperature, a vicious cycle known as *thermal runaway.*
- Junction capacitance limits high frequency gain of a transistor. The *Miller effect* makes C$_{cb}$ look β times larger at the base of a CE amplifier.
- Transistor noise limits the ability to amplify small signals. *Noise figure* is a figure of merit concerning transistor noise.
- When paralleling power transistors for increased current, insert *ballast resistors* in series with the emitters to equalize current.
- F$_T$ is the absolute upper frequency limit for a CE amplifier, small signal current gain falls to unity, h$_{fe}$=1.
- F$_{max}$ is the upper frequency limit for an oscillator under the most ideal conditions.

Bibliography

[1] A. G. Thiele in Loyd P. Hunter, “Handbook of Semiconductor Electronics,” Low Frequency Amplifiers, ISBN -07-031305-9, 1970

[2] "GE Transistor Manual", General Electric, 1964.

[3] R. Victor Jones, "Basic BJT Amplifier Configurations", November 7, 2001. at http://people.seas.harvard.edu/~jones/es154/lectures/lecture_3/bjt_amps/bjt_amps.html

[4] Tony Kuphaldt,"Lessons in Electric Circuits", Vol. 1, DC, DC Network Analysis, Thevenin's Theorem, at http://www.openbookproject.net/electricCircuits/DC/DC_10.html#xtocid102679

[5] "PN2222 Datasheet",Fairchild Semiconductor Corporation, 2007 at http://www.fairchildsemi.com/ds/PN/PN2222A.pdf

Chapter 5

JUNCTION FIELD-EFFECT TRANSISTORS

Contents

5.1 Introduction 279
5.2 The transistor as a switch 281
5.3 Meter check of a transistor 284
5.4 Active-mode operation 286
5.5 The common-source amplifier – PENDING 295
5.6 The common-drain amplifier – PENDING 296
5.7 The common-gate amplifier – PENDING 296
5.8 Biasing techniques – PENDING 296
5.9 Transistor ratings and packages – PENDING 297
5.10 JFET quirks – PENDING 297

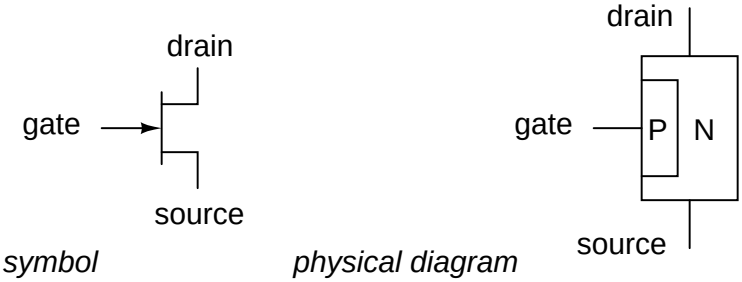
***** INCOMPLETE *****

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.

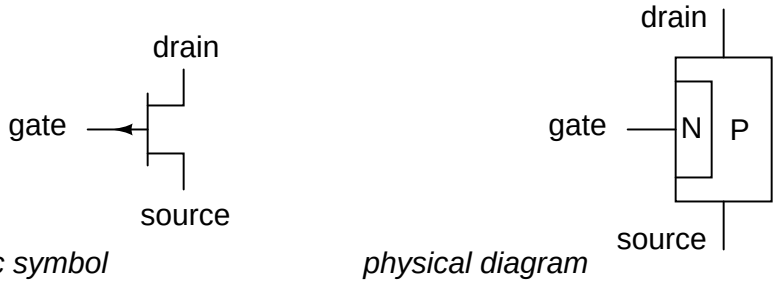
All field-effect transistors are *unipolar* rather than *bipolar* devices. That is, the main current through them is comprised either of electrons through an N-type semiconductor or holes

through a P-type semiconductor. This becomes more evident when a physical diagram of the device is seen:



schematic symbol *physical diagram*

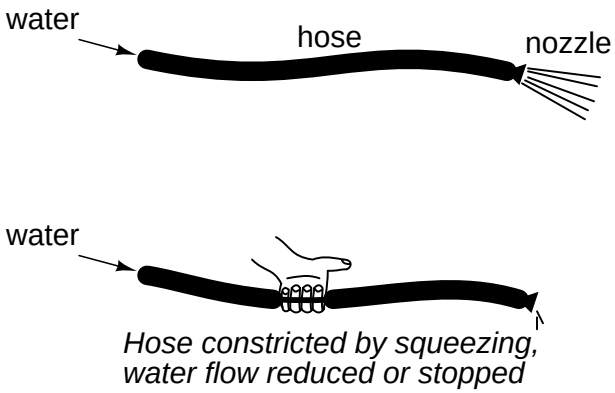
In a junction field-effect transistor, or JFET, the controlled current passes from source to drain, or from drain to source as the case may be. The controlling voltage is applied between the gate and source. Note how the current does not have to cross through a PN junction on its way between source and 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:



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.



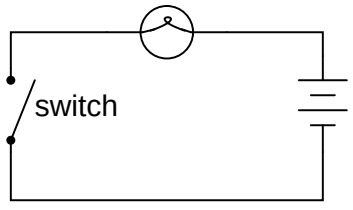
Note how this operational behavior is exactly opposite of the bipolar junction transistor. Bipolar transistors are *normally-off* devices: no current through the base, no current through the collector or the emitter. JFETs, on the other hand, are *normally-on* devices: no voltage applied to the gate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

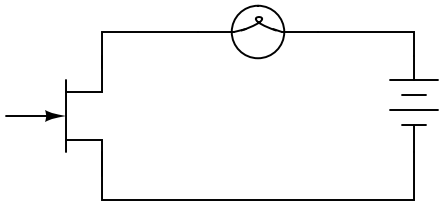
This is a very condensed overview of JFET operation. In the next section, we’ll explore the use of the JFET as a switching device.

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:

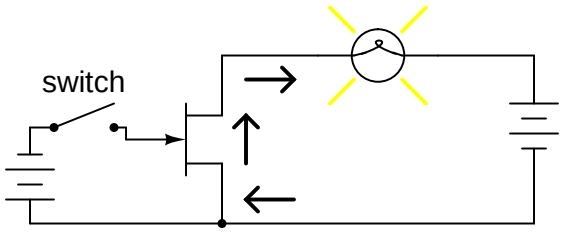


Remembering that the *controlled* current in a JFET flows between source and drain, we substitute the source and drain connections of a JFET for the two ends of the switch in the above circuit:

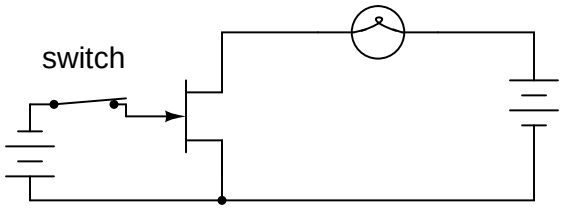


If you haven't noticed by now, the source and drain connections on a JFET look identical on the schematic symbol. Unlike the bipolar junction transistor where the emitter is clearly distinguished from the collector by the arrowhead, a JFET's source and drain lines both run perpendicular into the bar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

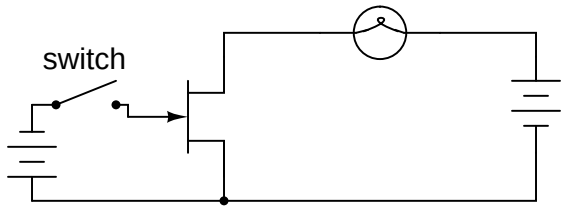


Closing this switch will "pinch off" the JFET's channel, thus forcing it into cutoff and turning the lamp off:

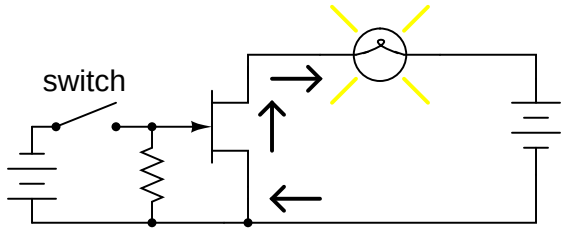


Note that there is no current going through the gate. As a reverse-biased PN junction, it firmly opposes the flow of any electrons through it. As a voltage-controlled device, the JFET requires negligible input current. This is an advantageous trait of the JFET over the bipolar transistor: there is virtually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



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:



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:**

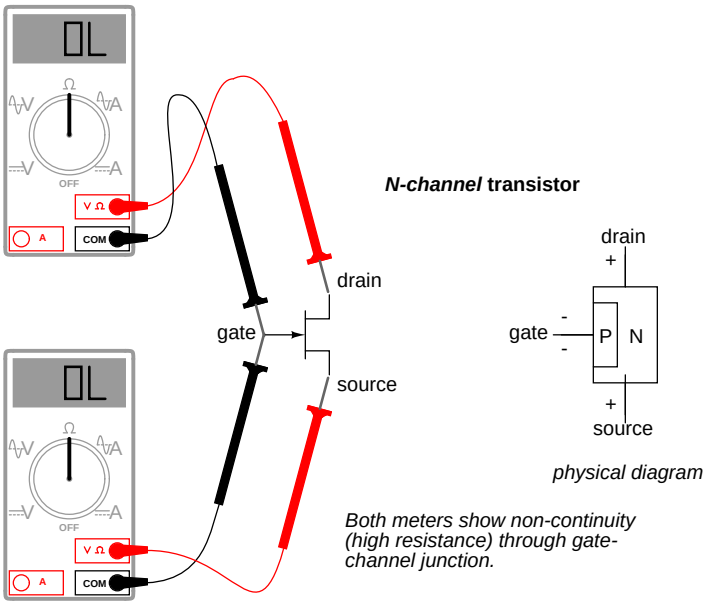
- Field-effect transistors control the current between source and drain connections by a voltage applied between the gate and source. In a *junction* field-effect transistor (JFET), there is a PN junction between the gate and source which is normally reverse-biased for control of source-drain current.

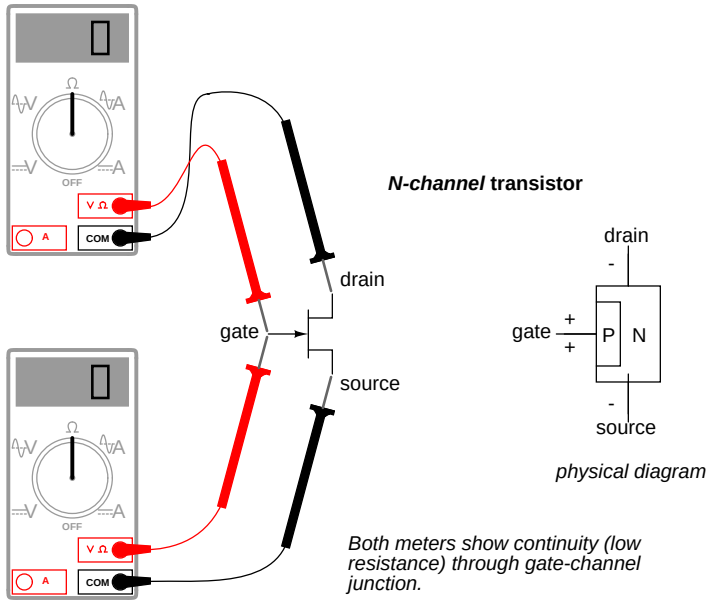
- JFETs are normally-on (normally-saturated) devices. The application of a reverse-biasing voltage between gate and source causes the depletion region of that junction to expand, thereby "pinching off" the channel between source and drain through which the controlled current travels.

- 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.





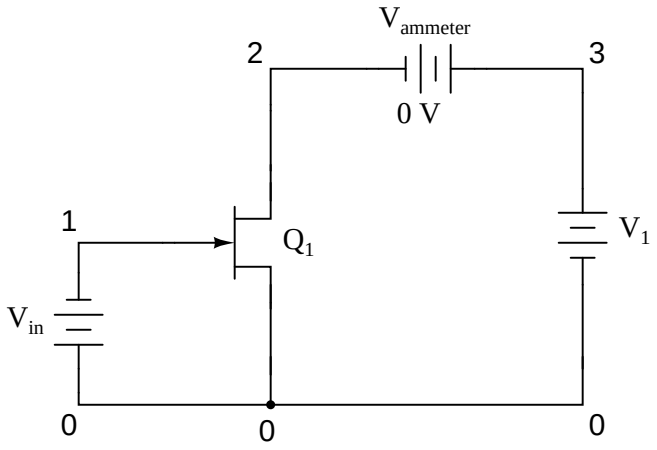
Testing continuity through the drain-source channel is another matter, though. Remember from the last section how a stored charge across the capacitance of the gate-channel PN junction could hold the JFET in a pinched-off state without any external voltage being applied across it? This can occur even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

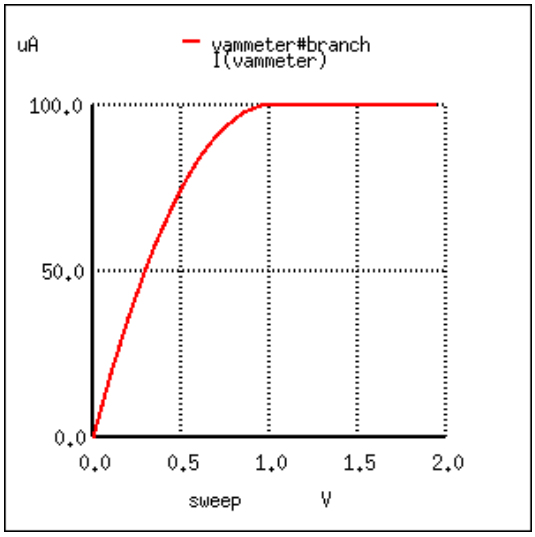
5.4 Active-mode operation

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
```

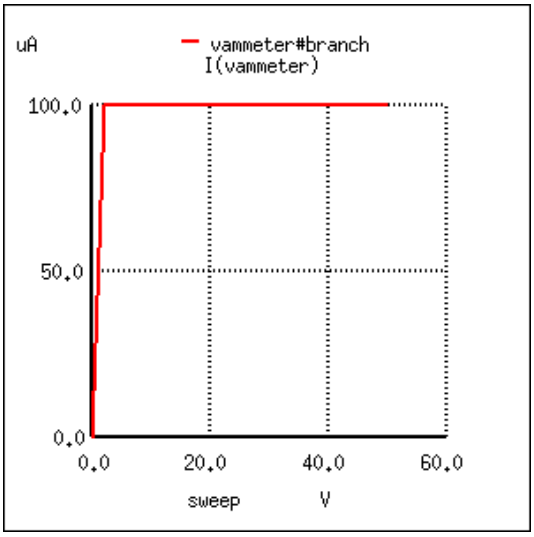
Note that the transistor labeled "Q_1" in the schematic is represented in the SPICE netlist as `j1`. Although all transistor types are commonly referred to as "Q" devices in circuit schematics – just as resistors are referred to by "R" designations, and capacitors by "C" – SPICE needs to be told what 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:

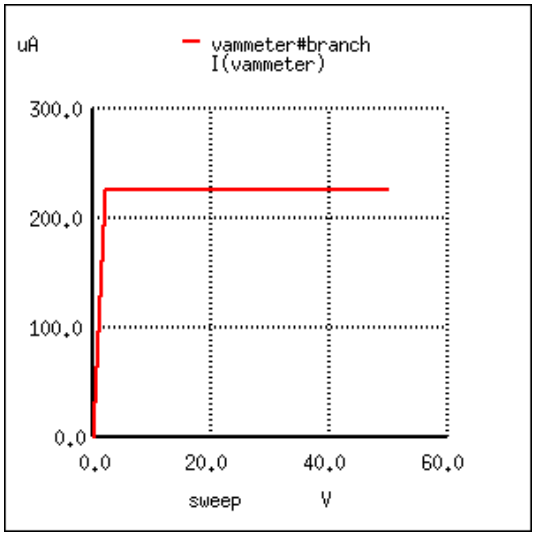
```
jfet simulation
vin 0 1 dc 1
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 njf
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



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:

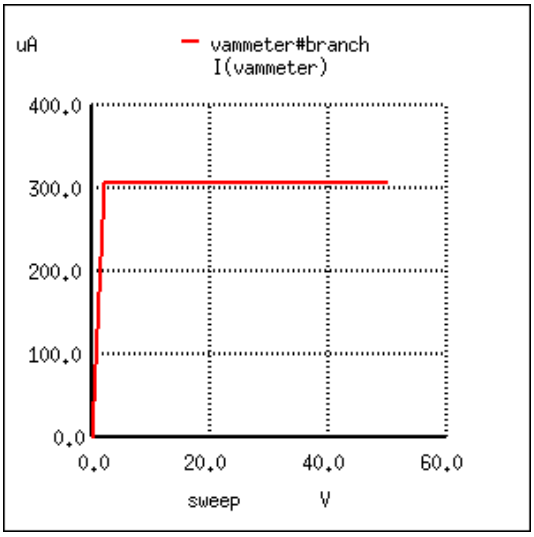
```
jfet simulation
vin 0 1 dc 0.5
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 njf
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



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:

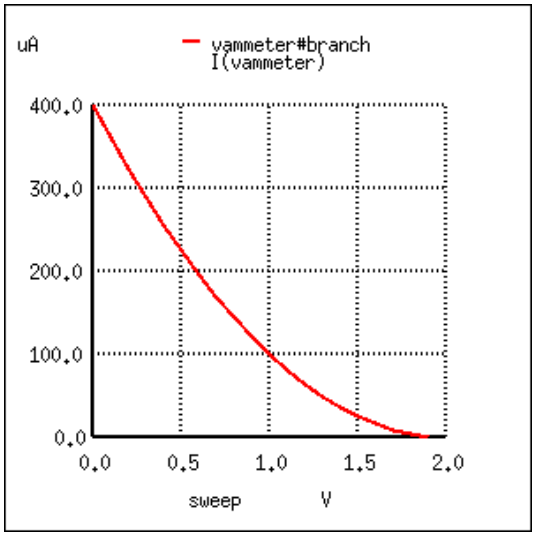
```
jfet simulation
vin 0 1 dc 0.25
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc
.model mod1 njf
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



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:

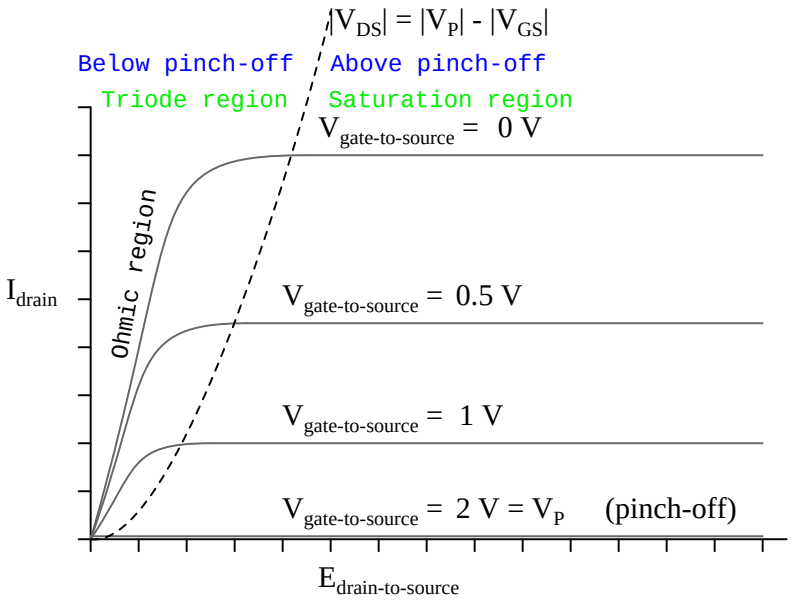
```
jfet simulation
vin 0 1 dc
j1 2 1 0 mod1
vammeter 3 2 dc 0
v1 3 0 dc 25
.model mod1 njf
.dc vin 0 2 0.1
.plot dc i(vammeter)
.end
```



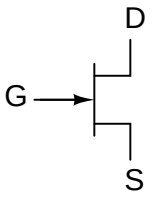
This simulation directly reveals an important characteristic of the junction field-effect transistor: the control effect of gate voltage over drain current is *nonlinear*. Notice how the drain current does not decrease linearly as the gate-source voltage is increased. With the bipolar junction transistor, 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 its being operated at the extreme limits of cutoff and saturation (off and on, respectively, like a switch).

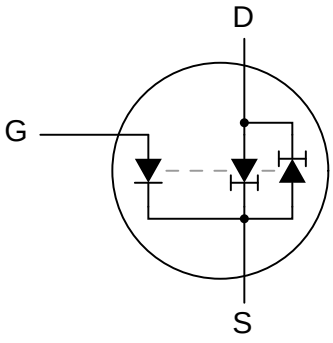
A JFET's characteristic curves display the same current-regulating behavior as for a BJT, and the nonlinearity between gate-to-source voltage and drain current is evident in the disproportionate vertical spacings between the curves:



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:



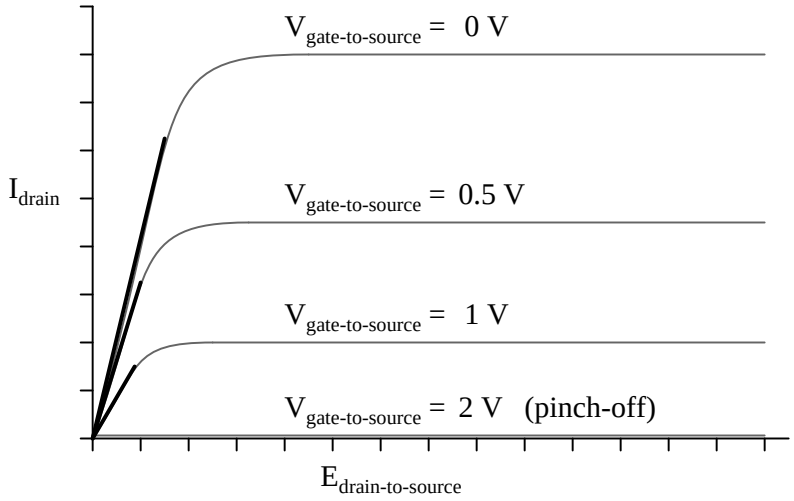
N-channel JFET diode-regulating diode model



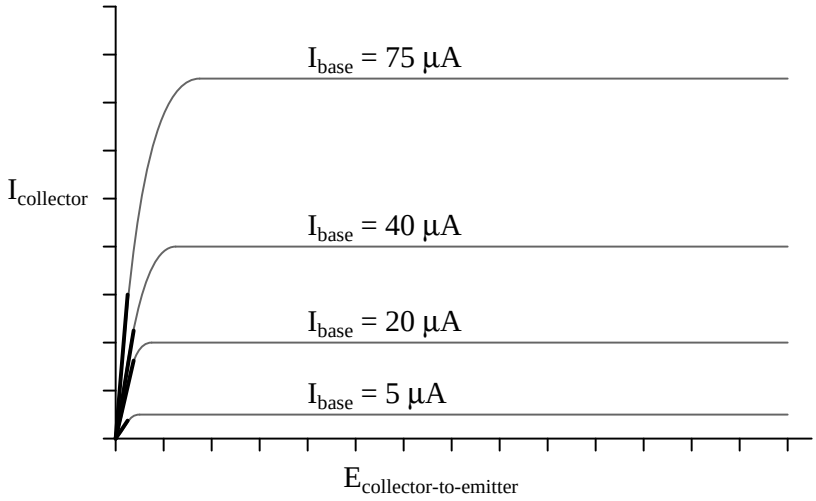
In the case of the JFET, it is the *voltage* across the reverse-biased gate-source diode which sets the current regulation point for the pair of constant-current diodes. A pair of opposing constant-current diodes is included in the model to facilitate current in either direction be-

tween 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 non-horizontal area is surprisingly long compared to the respective portions of a BJT's characteristic curves:

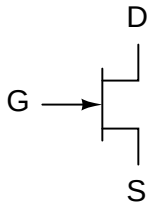


"Ohmic regions"

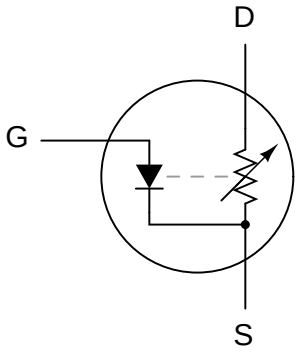


A JFET transistor operated in the *triode region* tends to act very much like a plain resistor as measured from drain to source. Like all simple resistances, its current/voltage graph is a straight line. For this reason, the triode region (non-horizontal) portion of a JFET's characteristic curve is sometimes 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), 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:

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

Where,

I_D = Drain current
I_{DSS} = Drain current with gate shorted to source
V_{GS} = Gate-to-source voltage
V_P = 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 triode region, 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

6.1 Introduction 299
6.2 Depletion-type IGFETs 300
6.3 Enhancement-type IGFETs – PENDING 309
6.4 Active-mode operation – PENDING 309
6.5 The common-source amplifier – PENDING 310
6.6 The common-drain amplifier – PENDING 310
6.7 The common-gate amplifier – PENDING 310
6.8 Biasing techniques – PENDING 310
6.9 Transistor ratings and packages – PENDING 310
6.10 IGFET quirks – PENDING 311
6.11 MESFETs – PENDING 311
6.12 IGBTs 311

***** 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.

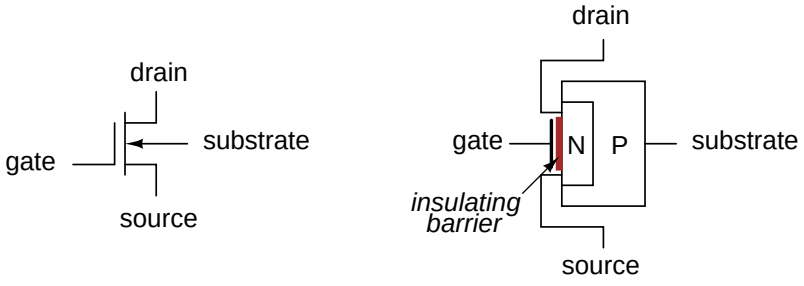
In addition to a choice of N-channel versus P-channel design, IGFETs come in two major types: *enhancement* and *depletion*. The depletion type is more closely related to the JFET, so we will begin our study of IGFETs with it.

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:

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

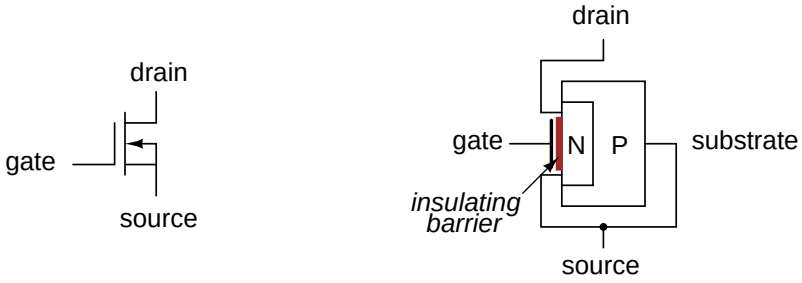


schematic symbol *physical diagram*

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**

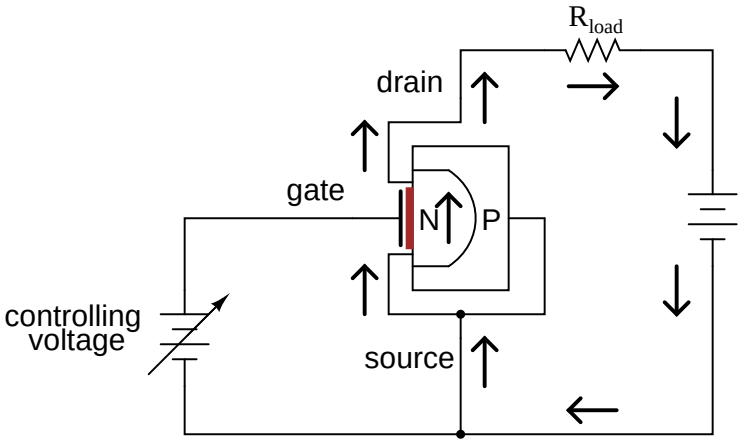


schematic symbol *physical diagram*

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.

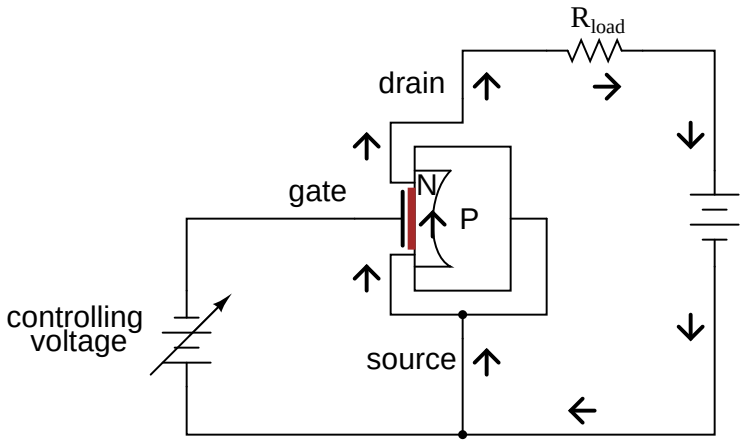
Indeed it is, for when a controlling voltage is applied between gate and source, the conductivity of the channel is changed as a result of the depletion region *moving* closer to or further away from the gate. In other words, the channel's effective width changes just as with the JFET, but this change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

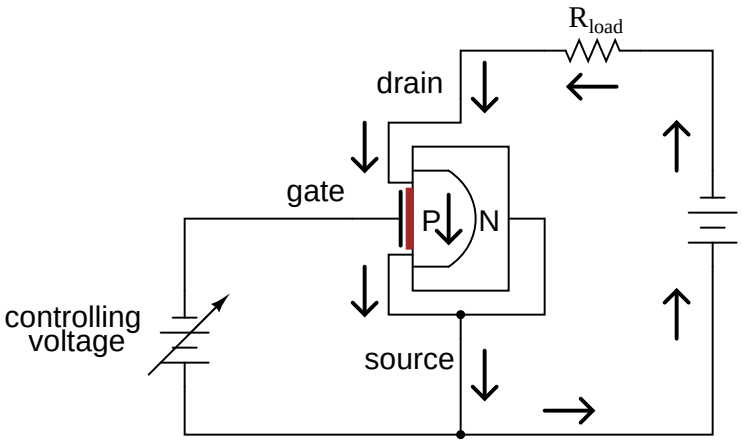
Reversing the controlling voltage's polarity has the opposite effect, attracting the depletion region and narrowing the channel, consequently reducing channel conductivity:



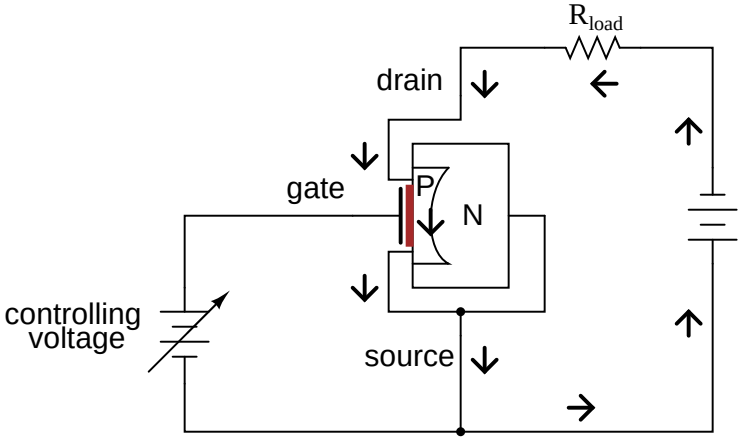
The insulated gate allows for controlling voltages of any polarity without danger of forward-biasing a junction, as was the concern with JFETs. This type of IGFET, although its called a "depletion-type," actually has the capability of having its channel *either* depleted (channel narrowed) *or* enhanced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

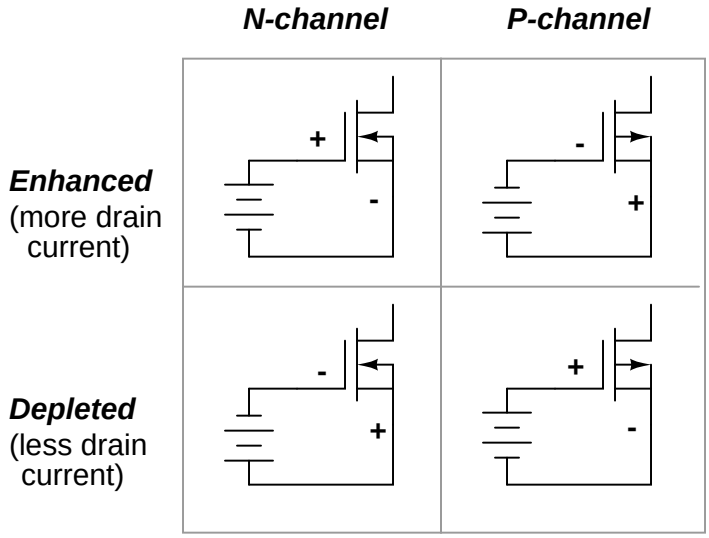


Channel expands for greater conductivity



Channel narrows for less conductivity

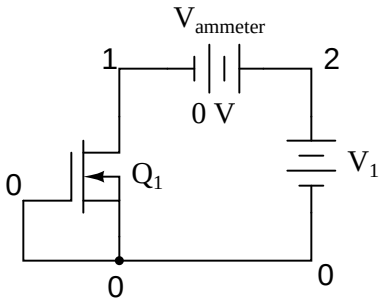
Illustrating the proper biasing polarities with standard IGFET symbols:



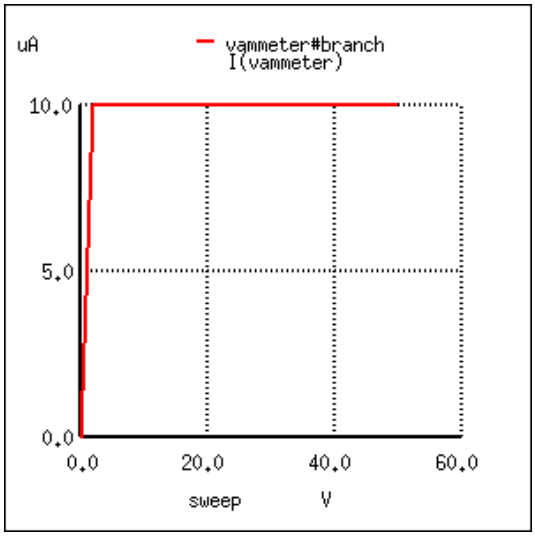
When there is zero voltage applied between gate and source, the IGFET will conduct current between source and drain, but not as much current as it would if it were enhanced by the proper gate voltage. This places the depletion-type, or simply *D-type*, IGFET in a category of its own in the transistor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

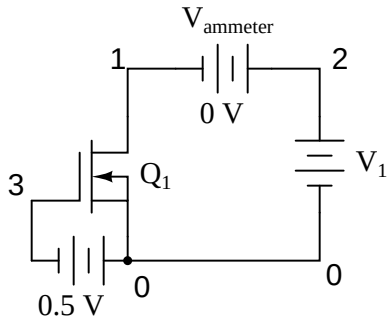
A few SPICE analyses will demonstrate the current-regulating behavior of D-type IGFETs. First, a test with zero input voltage (gate shorted to source) and the power supply swept from 0 to 50 volts. The graph shows drain current:



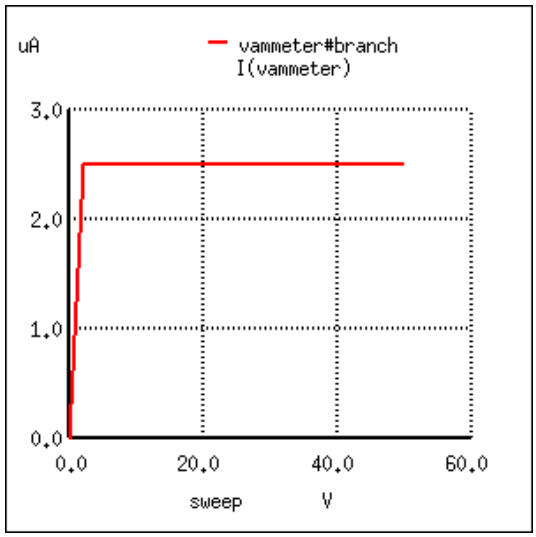

```
n-channel igfet characteristic curve
m1 1 0 0 0 mod1
vammeter 2 1 dc 0
v1 2 0
.model mod1 nmos vto=-1
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



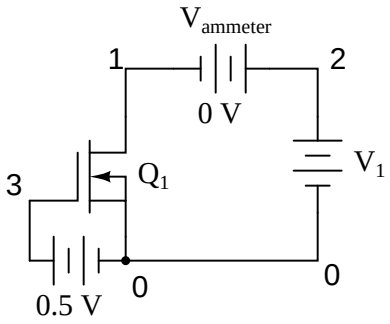
As expected for any transistor, the controlled current holds steady at a regulated value over a wide range of power supply voltages. In this case, that regulated point is 10 μA (1.000E-05). Now let's see what happens when we apply a negative voltage to the gate (with reference to the source) and sweep the power supply over the same range of 0 to 50 volts:



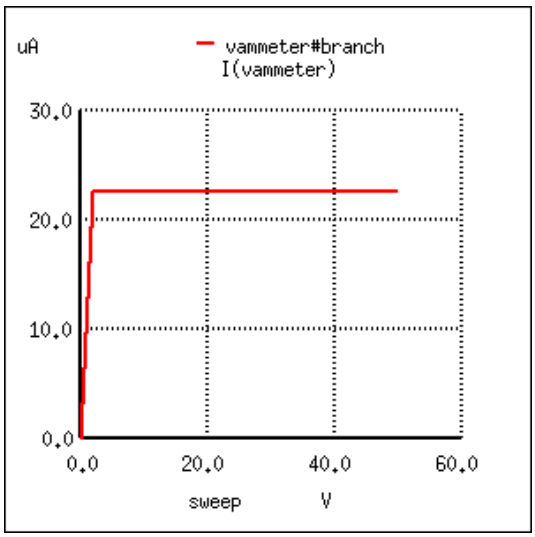

```
n-channel igfet characteristic curve
m1 1 3 0 0 mod1
vin 0 3 dc 0.5
vammeter 2 1 dc 0
v1 2 0
.model mod1 nmos vto=-1
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



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:



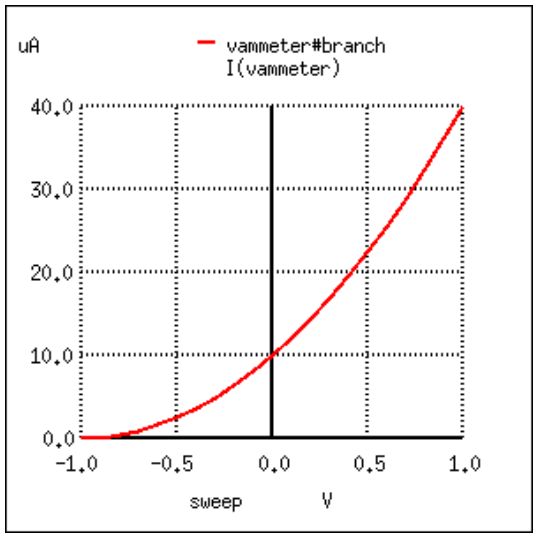

```
n-channel igfet characteristic curve
m1 1 3 0 0 mod1
vin 3 0 dc 0.5
vammeter 2 1 dc 0
v1 2 0
.model mod1 nmos vto=-1
.dc v1 0 50 2
.plot dc i(vammeter)
.end
```



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

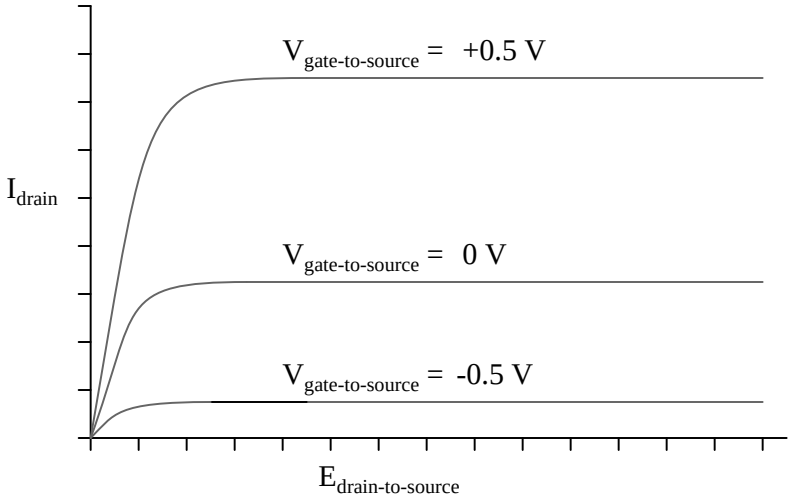
can use SPICE to plot the drain current over a range of input voltage values, sweeping from a negative (depleting) figure to a positive (enhancing) figure, maintaining the power supply voltage of V_1 at a constant value:

```
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:**
-
-
-

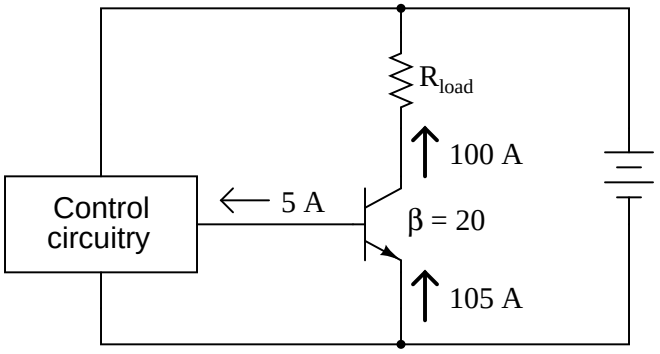
6.11 MESFETs – PENDING

- **REVIEW:**
-
-
-

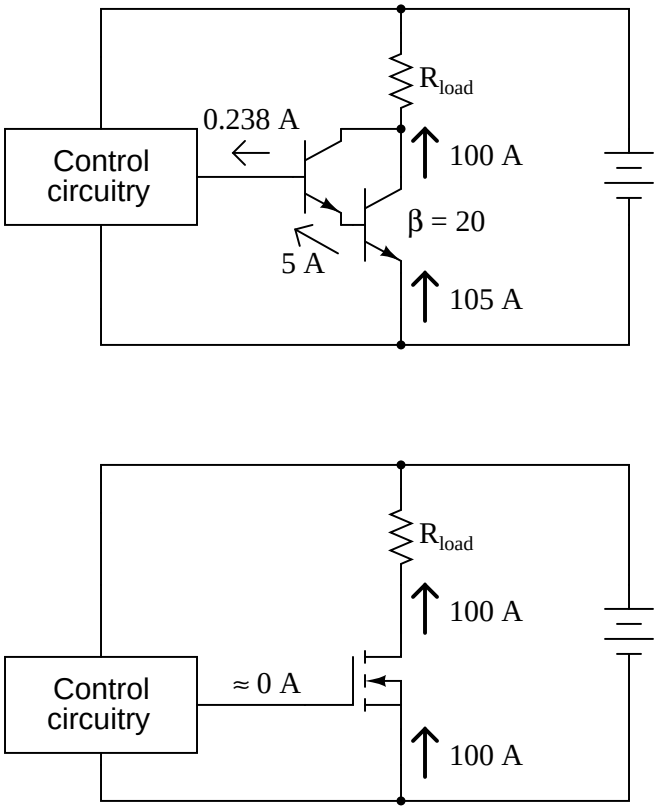
6.12 IGBTs

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:

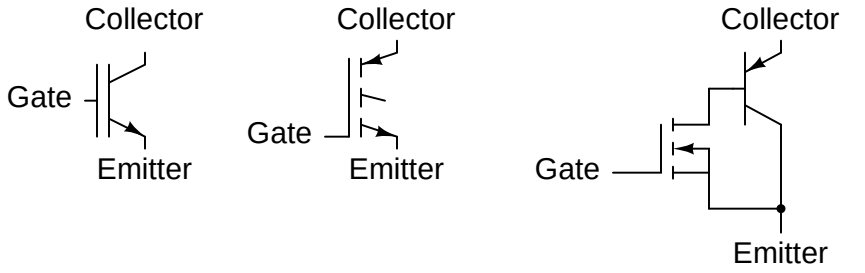


Unfortunately, though, IGFETs have problems of their own controlling high current: they typically exhibit greater drain-to-source voltage drop while saturated than the collector-to-emitter voltage drop of a saturated BJT. This greater voltage drop equates to higher power dissipation for the same amount 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)

Schematic symbols *Equivalent circuit*



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, its 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

7.1 Hysteresis . . . 315
7.2 Gas discharge tubes . . . 316
7.3 The Shockley Diode . . . 320
7.4 The DIAC . . . 326
7.5 The Silicon-Controlled Rectifier (SCR) . . . 327
7.6 The TRIAC . . . 339
7.7 Optothyristors . . . 342
7.8 The Unijunction Transistor (UJT) . . . 342
7.9 The Silicon-Controlled Switch (SCS) . . . 348
7.10 Field-effect-controlled thyristors . . . 350
Bibliography . . . 352

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-hysteric devices. That is, these 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 these will remain that way on their own, even after the cause of the state change is taken away. As such, these 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 hysteric device (a thyristor) be made from non-hysteric 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 hysteric) 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 dust in air lowers its resistance some, but it is still an insulator for most practical purposes. When enough 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 called a *plasma* rather than a *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 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 in Figure 7.1.

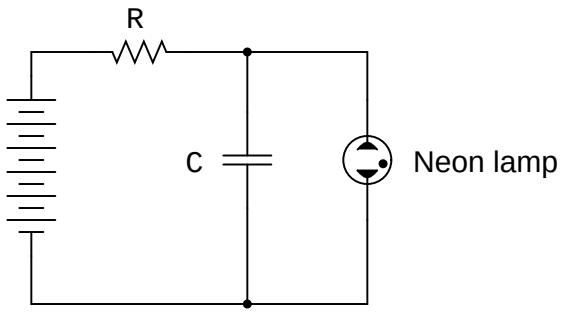


Figure 7.1: *Simple relaxation oscillator*

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 is 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 hysteric 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.

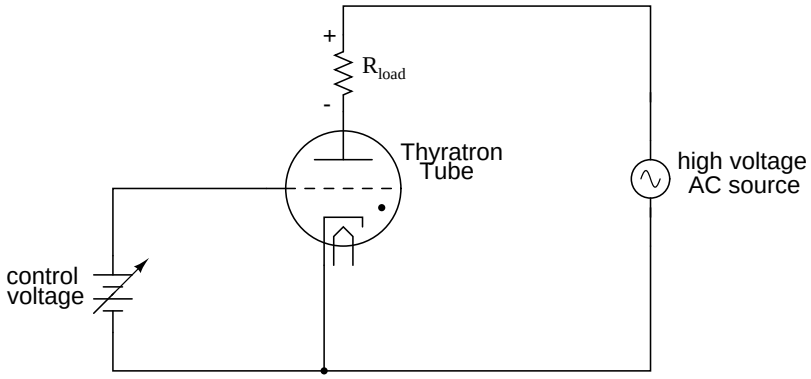


Figure 7.2: *Simple thyratron control circuit*

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 in Figure 7.2. 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 about how the thyratron turns off after its 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 enough voltage is 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 Figure 7.3.

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. Load voltage exists (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, no voltage develops across the load through the negative half-cycle of AC. In

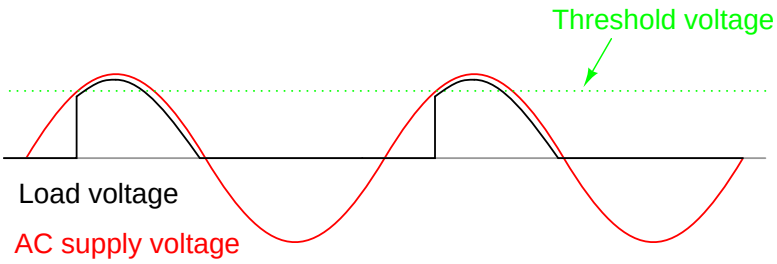


Figure 7.3: *Thyratron waveforms*

practical thyratron circuits, multiple tubes arranged in some form of full-wave rectifier circuit to facilitate full-wave DC power to the load.

The thyratron tube has been applied to a relaxation oscillator circuit. [1] The frequency is controlled by a small DC voltage between grid and cathode. (See Figure 7.4) This voltage-controlled oscillator is known as a *VCO*. Relaxation oscillators produce a very non-sinusoidal output, and they exist mostly as demonstration circuits (as is the case here) or in applications where the harmonic rich waveform is desirable. [2]

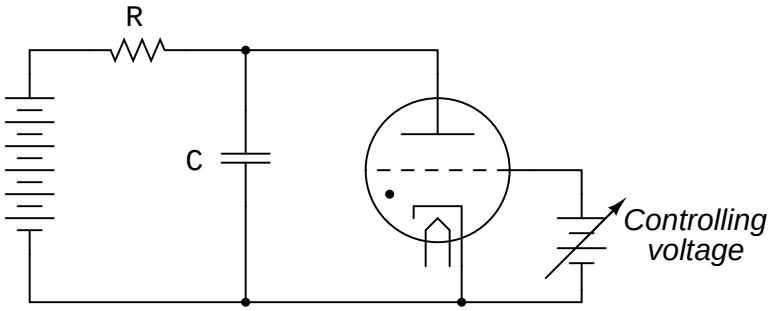


Figure 7.4: *Voltage controlled thyratron relaxation oscillator*

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 *hyster*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, Figure 7.5.

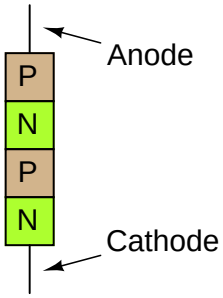


Figure 7.5: *Shockley or 4-layer diode*

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 in Figure 7.6.

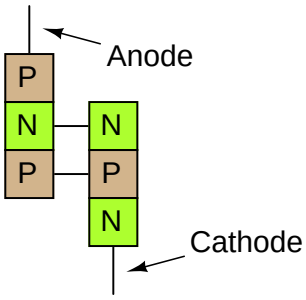


Figure 7.6: *Transistor equivalent of Shockley diode*

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 (Figure 7.7)

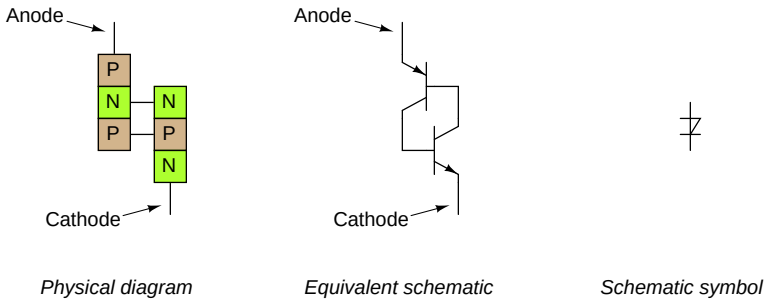


Figure 7.7: *Shockley diode: physical diagram, equivalent schematic diagram, and schematic symbol.*

Let's connect one of these devices to a source of variable voltage and see what happens: (Figure 7.8)

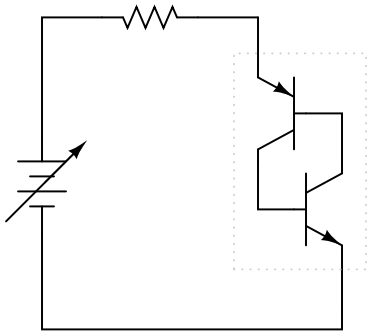


Figure 7.8: *Powered Shockley diode equivalent circuit.*

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 no base current flows, 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 each can withstand before one breaks down and conduct. If two real transistors are connected in this fashion to form a Shockley diode, each one will conduct if sufficient voltage is 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 each 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 too little current flows 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 each one was 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 evident. First, we will observe the circuit as the DC voltage source (battery) is set to zero voltage: (Figure 7.9)

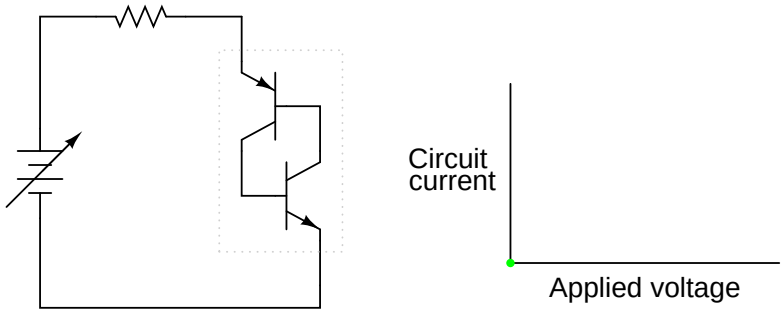


Figure 7.9: *Zero applied voltage; zero current*

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: (Figure 7.10)

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: (Figure 7.11)

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 quickly seen in the circuit: (Figure 7.12)

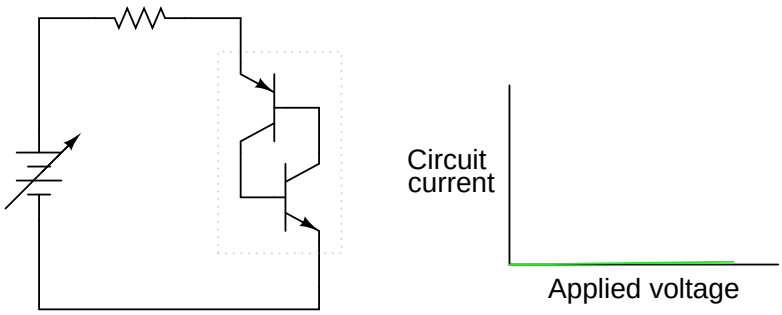


Figure 7.10: *Some applied voltage; still no current*

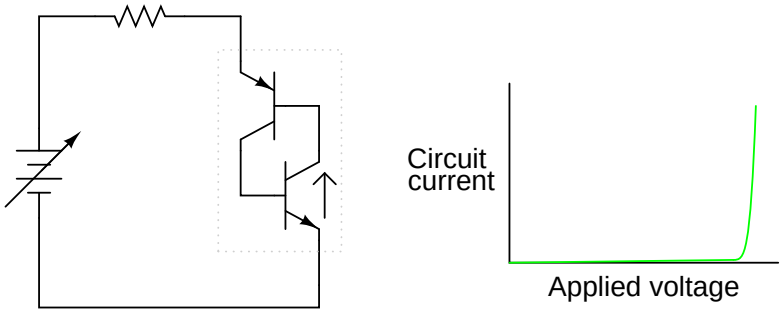


Figure 7.11: *More voltage applied; lower transistor breaks down*

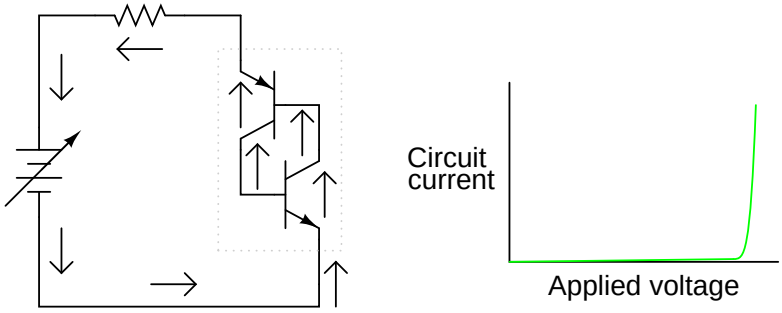


Figure 7.12: *Transistors are now fully conducting.*

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, each 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): (Figure 7.13)

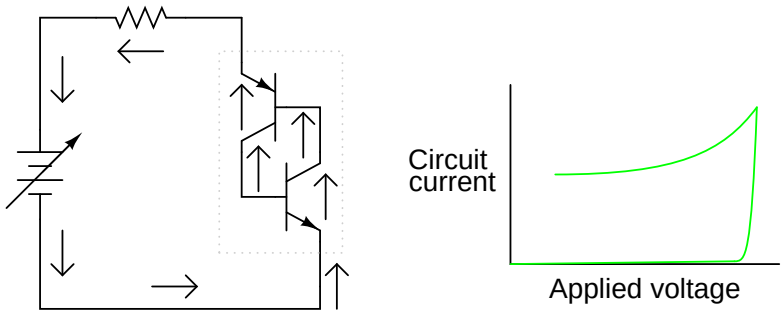


Figure 7.13: *Current maintained even when voltage is reduced*

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: (Figure 7.14)

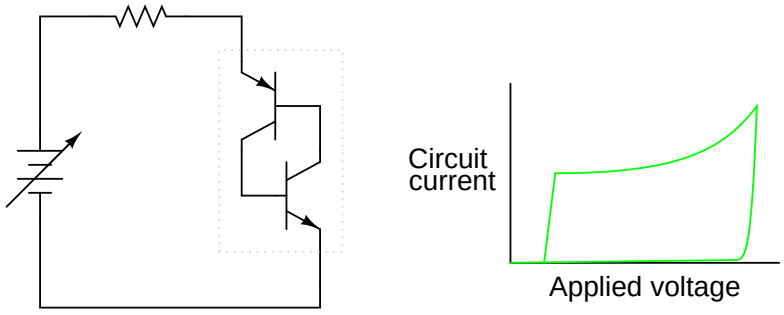


Figure 7.14: *If voltage drops too low, both transistors shut off.*

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 classically 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: (Figure 7.15)

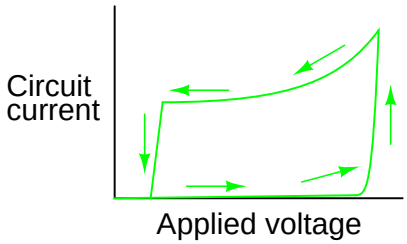


Figure 7.15: *Hysteretic curve*

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

A few special terms apply 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. To get a Shockley diode to latch, the applied voltage must be increased until *breakover* is attained. Though this action is best described as transistor break*down*, the term break*over* is used instead because the result is a pair of transistors in mutual saturation rather than destruction of the transistor. A latched Shockley diode is re-set back into its nonconducting state by reducing current through it until *low-current dropout* occurs.

Note that Shockley diodes may be fired in a way other than breakover: excessive *voltage rise*, or dv/dt. If the applied voltage across the diode increases at a high rate of change, it may trigger. 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 parallel resistor-capacitor networks called *snubbers*: (Figure 7.16)

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. These behave as a pair of interconnected PNP and NPN transistors.

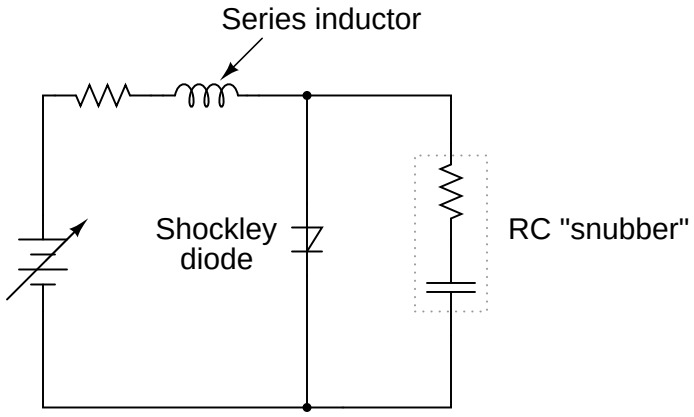


Figure 7.16: *Both the series inductor and parallel resistor-capacitor "snubber" circuit help minimize the Shockley diode's exposure to excessively rising voltage.*

- Like all thyristors, Shockley diodes tend to stay on once turned on (*latched*), and stay off once turned off.
- To latch a Shockley diode exceed the anode-to-cathode *breakover* voltage, or exceed the anode-to-cathode *critical rate of voltage rise*.
- To cause a Shockley diode to stop conducting, 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, these 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*: (Figure 7.17)

DIAC equivalent circuit *DIAC schematic symbol*

Figure 7.17: *The DIAC*

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 voltage is 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 the current waveform in Figure 7.18.

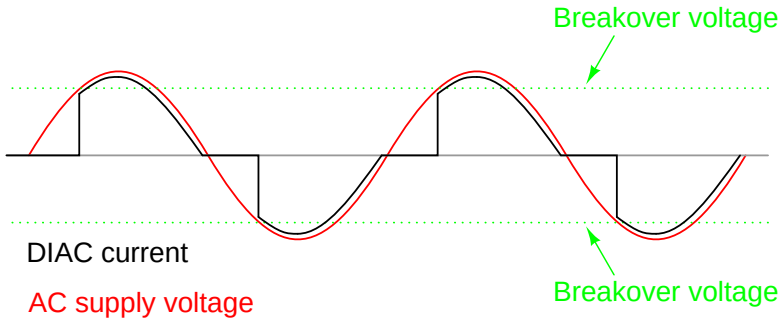


Figure 7.18: *DIAC waveforms*

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, each becomes true amplifying devices (if only in an on/off mode), and we refer to these 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: (Figure 7.19)

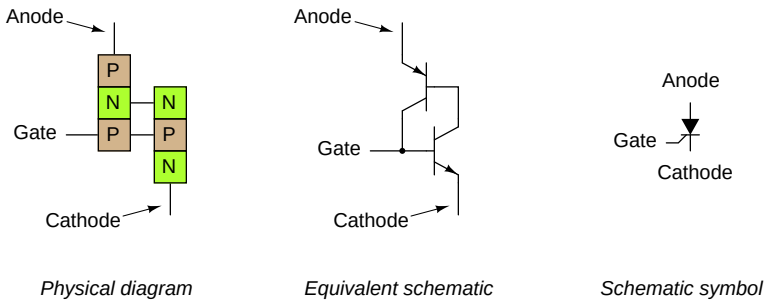


Figure 7.19: *The Silicon-Controlled Rectifier (SCR)*

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: (Figure 7.20)

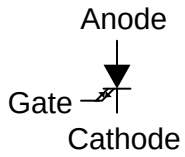


Figure 7.20: *The Gate Turn-Off thyristor (GTO)*

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: (Figure 7.21)

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.

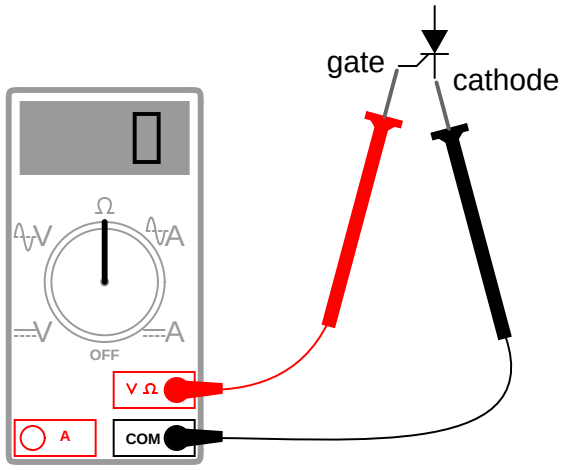


Figure 7.21: *Rudimentary test of SCR*

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: (Figure 7.22)

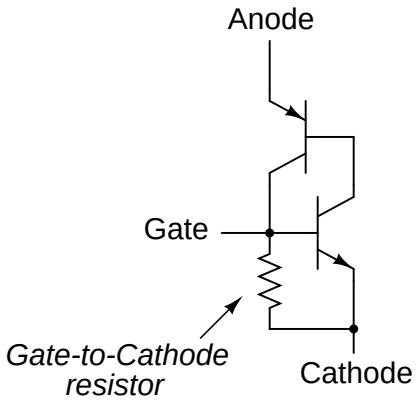


Figure 7.22: *Larger SCRs have gate to cathode resistor.*

”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: (Figure 7.23)

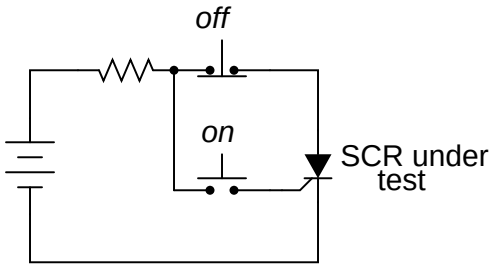


Figure 7.23: *SCR testing circuit*

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. A certain minimum amount of load current is 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: (Figure 7.24)

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 placing a direct short-circuit on the output of that supply to prevent

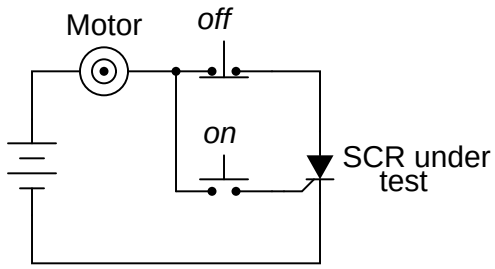


Figure 7.24: *DC motor start/stop control circuit*

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: (Figure 7.25)

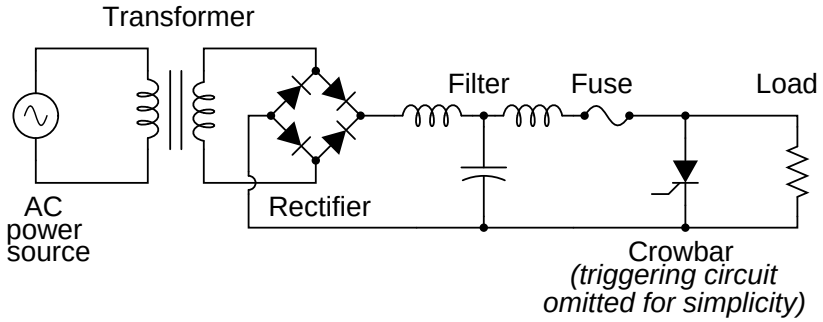


Figure 7.25: *Crowbar circuit used in DC power supply*

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, 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: (Figure 7.26)

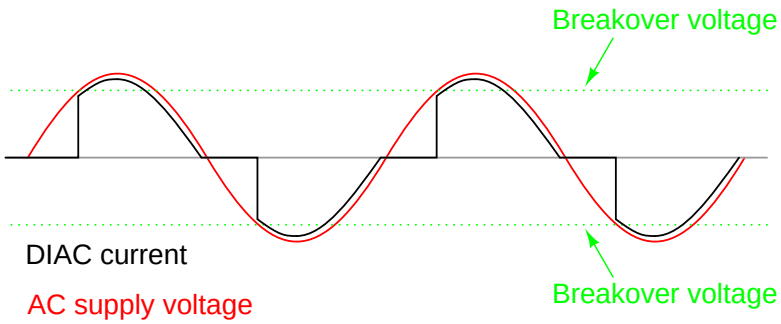


Figure 7.26: *DIAC bidirectional response*

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 circuit in Figure 7.27 as an example. Here, an SCR is positioned in a circuit to control power to a load from an AC source.

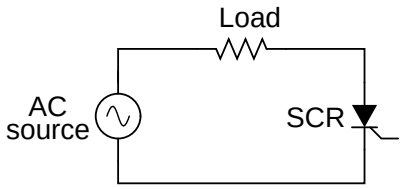


Figure 7.27: *SCR control of AC power*

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 standard 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: (Figure 7.28)

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 enough gate current triggers the SCR. In other words, if we make it harder for electrons to flow through the gate by

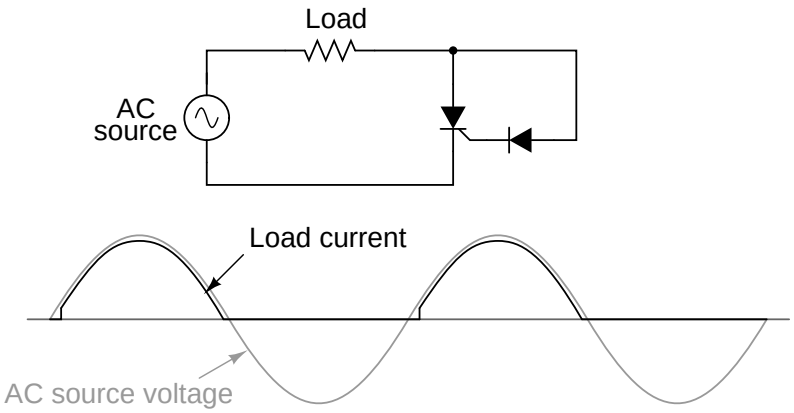


Figure 7.28: *Gate connected directly to anode through a diode; nearly complete half-wave current through load.*

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 is in Figure 7.29.

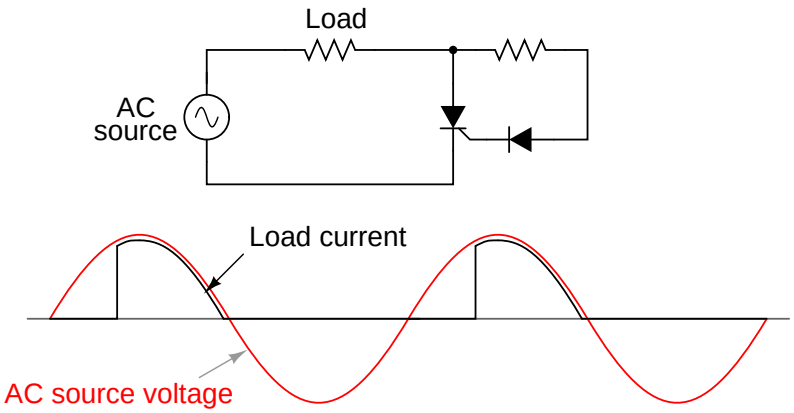


Figure 7.29: *Resistance inserted in gate circuit; less than half-wave current through load.*

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: (Figure 7.30)

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, it is not possible 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: (Figure 7.31)

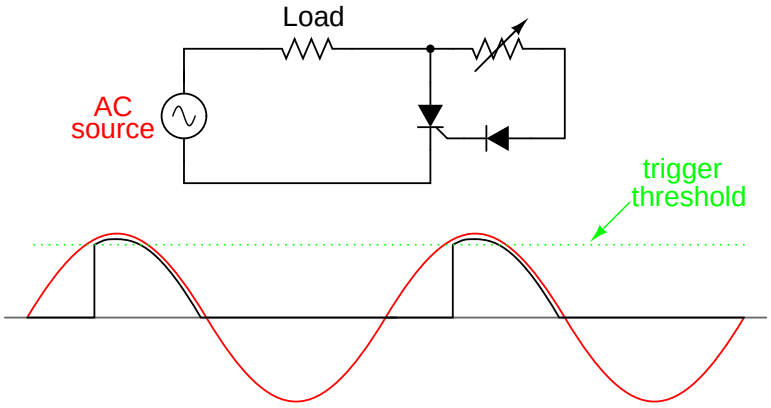


Figure 7.30: *Increasing the resistance raises the threshold level, causing less power to be delivered to the load. Decreasing the resistance lowers the threshold level, causing more power to be delivered to the load.*

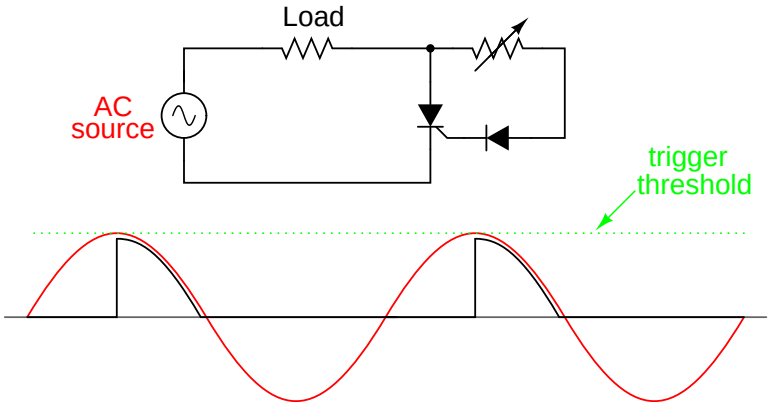


Figure 7.31: *Circuit at minimum power setting*

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: (Figure 7.32)

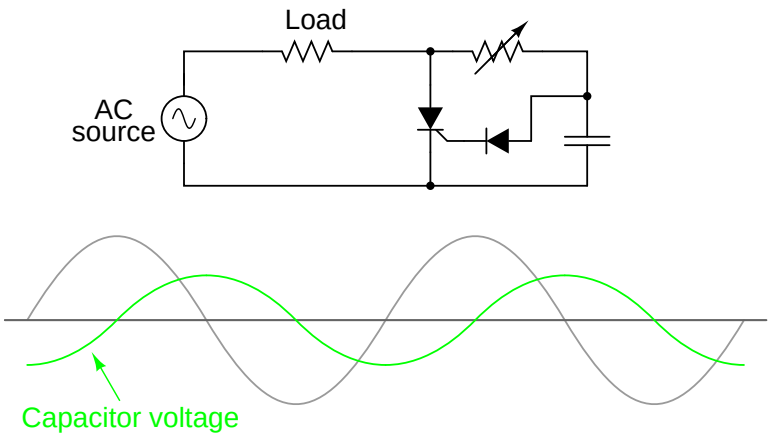


Figure 7.32: *Addition of a phase-shifting capacitor to the circuit*

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 with 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.

With enough voltage across the capacitor to periodically trigger the SCR, the resulting load current waveform will look something like Figure 7.33)

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; thus, 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: (Figure 7.34)

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 in Figure 7.35.

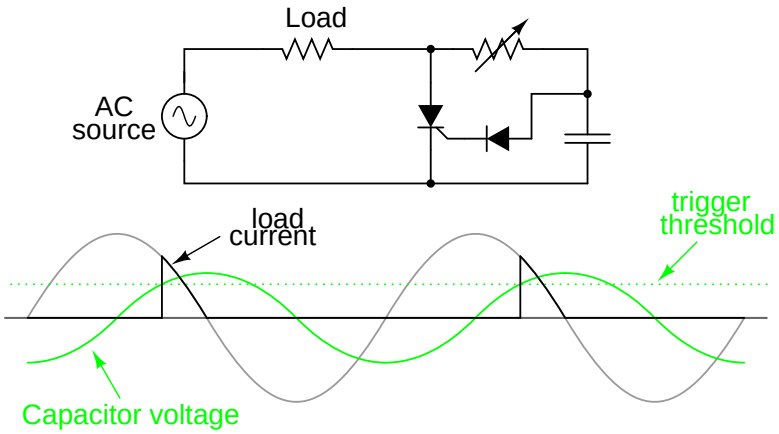


Figure 7.33: *Phase-shifted signal triggers SCR into conduction.*

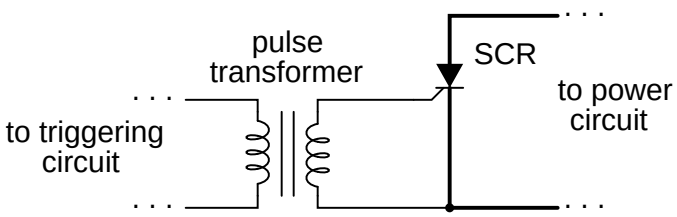


Figure 7.34: *Transformer coupling of trigger signal provides isolation.*

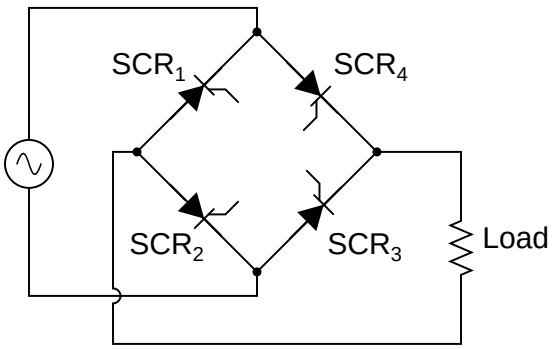


Figure 7.35: *Controlled bridge rectifier*

In any bridge rectifier circuit, the rectifying diodes (in this example, the rectifying SCRs) must conduct in opposite pairs. SCR_1 and SCR_3 must be fired simultaneously, and 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: (Figure 7.36)

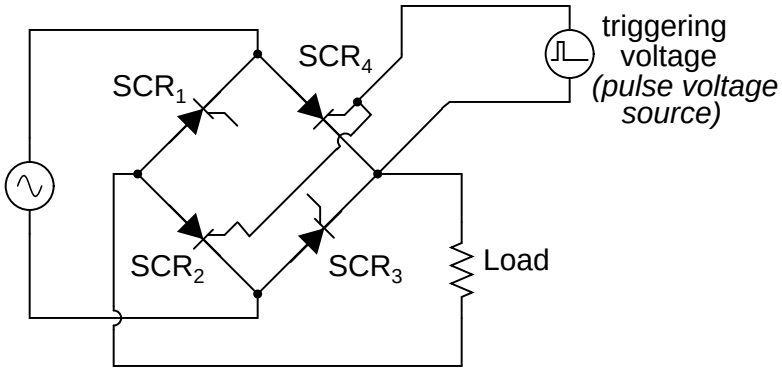


Figure 7.36: *This strategy will not work for triggering SCR_2 and SCR_4 as a pair.*

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: (Figure 7.37)

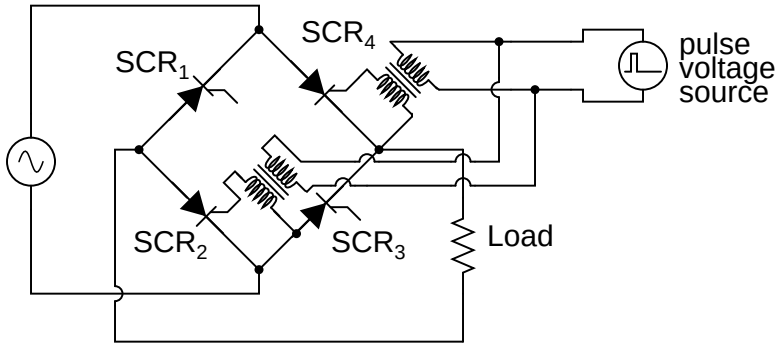


Figure 7.37: *Transformer coupling of the gates allows triggering of SCR_2 and SCR_4 .*

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 Figure 7.38.

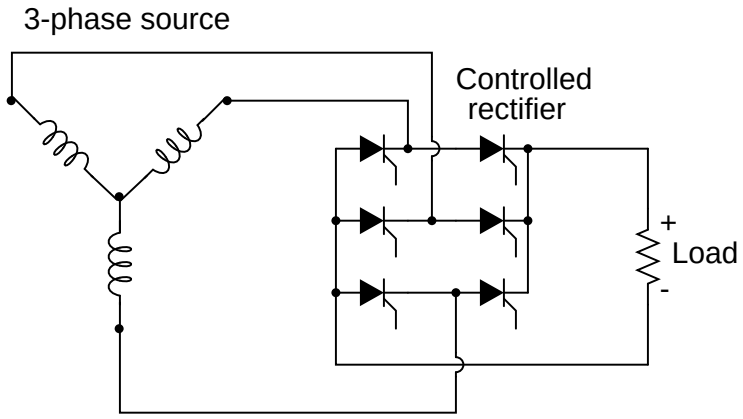


Figure 7.38: *Three-phase bridge SCR control of load*

- **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*: (Figure 7.39)

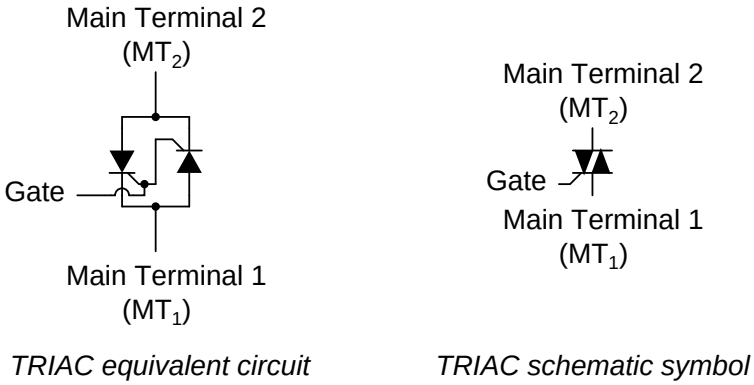


Figure 7.39: *The TRIAC SCR equivalent and, TRIAC schematic symbol*

Because individual SCRs are more flexible to use in advanced control systems, these are more commonly seen in circuits like motor drives; TRIACs are usually seen in simple, low-power applications like household dimmer switches. A simple lamp dimmer circuit is shown in Figure 7.40, complete with the phase-shifting resistor-capacitor network necessary for after-peak firing.

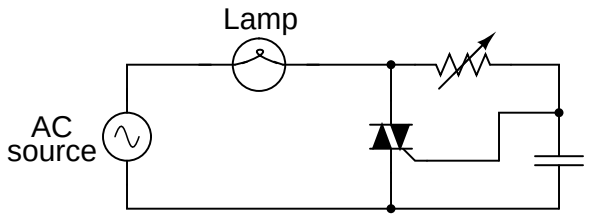


Figure 7.40: *TRIAC phase-control of power*

TRIACs are notorious for not firing *symmetrically*. This means these 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 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: (Figure 7.41)

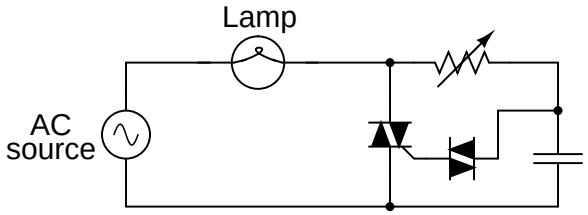


Figure 7.41: *DIAC improves symmetry of control*

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. These 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: (Figure 7.42)

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: (Figure 7.43)

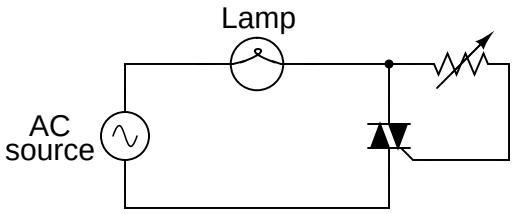


Figure 7.42: *This circuit with the gate to MT_2 does function.*

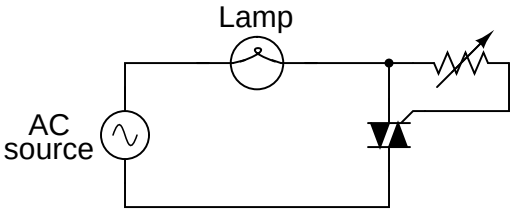


Figure 7.43: *With the gate swapped to MT_1, this circuit does not function.*

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 Figure 7.44.

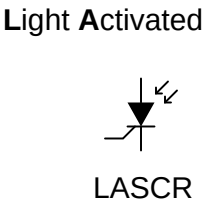


Figure 7.44: *Light activated SCR*

Optically-controlled TRIACs don't receive the honor of having their own acronym, but instead are humbly known as opto-TRIACs. Their schematic symbol is shown in Figure 7.45.

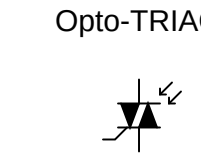


Figure 7.45: *Opto-TRIAC*

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)

Unijunction transistor: Although a unijunction transistor is not a thyristor, this device can trigger larger thyristors with a pulse at base B1. A *unijunction transistor* is composed of a bar of N-type silicon having a P-type connection in the middle. See Figure 7.46(a). The connections at the ends of the bar are known as bases B1 and B2; the P-type mid-point is the emitter. With the emitter disconnected, the total resistance R_{BBO}, a datasheet item, is the sum of R_{B1} and R_{B2} as shown in Figure 7.46(b). R_{BBO} ranges from 4-12kΩ for different device types. The intrinsic standoff ratio η is the ratio of R_{B1} to R_{BBO}. It varies from 0.4 to 0.8 for different devices. The schematic symbol is Figure 7.46(c)

The Unijunction emitter current vs voltage characteristic curve (Figure 7.47(a)) shows that as V_E increases, current I_E increases up I_P at the peak point. Beyond the peak point, current increases as voltage decreases in the negative resistance region. The voltage reaches a minimum at the valley point. The resistance of R_{B1}, the saturation resistance is lowest at the valley point.

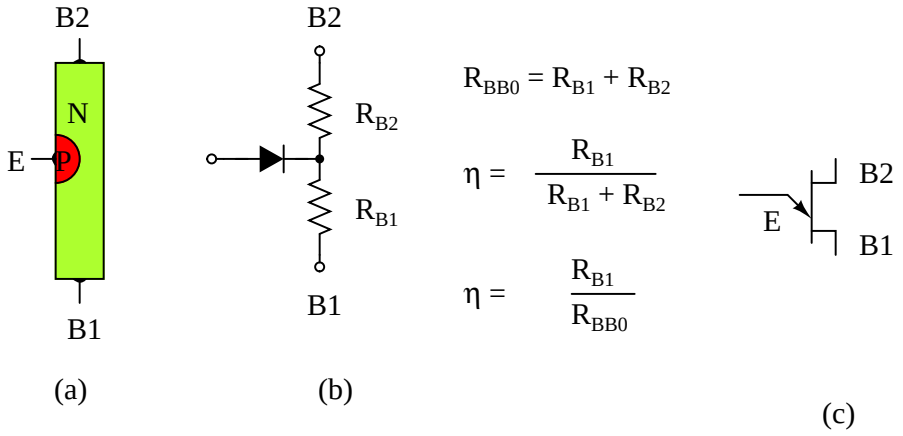


Figure 7.46: *Unijunction transistor: (a) Construction, (b) Model, (c) Symbol*

I_P and I_V, are datasheet parameters; For a 2n2647, I_P and I_V are 2μA and 4mA, respectively. [5] V_P is the voltage drop across R_{B1} plus a 0.7V diode drop; see Figure 7.47(b). V_V is estimated to be approximately 10% of V_{BB}.

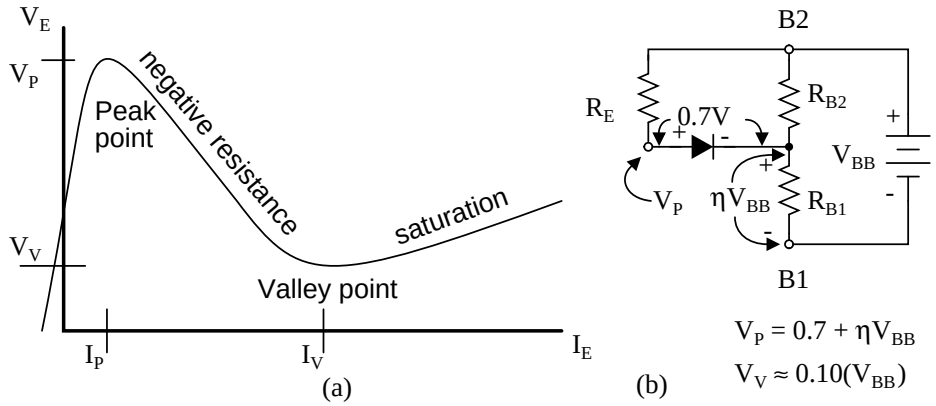
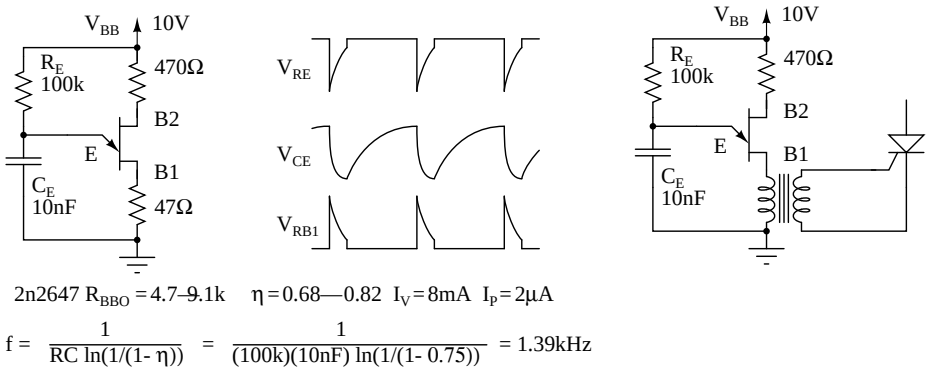


Figure 7.47: *Unijunction transistor: (a) emitter characteristic curve, (b) model for V_P.*

The relaxation oscillator in Figure 7.48 is an application of the unijunction oscillator. R_E charges C_E until the peak point. The unijunction emitter terminal has no effect on the capacitor until this point is reached. Once the capacitor voltage, V_E, reaches the peak voltage point V_P, the lowered emitter-base1 E-B1 resistance quickly discharges the capacitor. Once the capacitor discharges below the valley point V_V, the E-RB1 resistance reverts back to high resistance, and the capacitor is free to charge again.

During capacitor discharge through the E-B1 saturation resistance, a pulse may be seen on the external B1 and B2 load resistors, Figure 7.48. The load resistor at B1 needs to be low to



2n2647 R_{BBO} = 4.7–9.1k η = 0.68—0.82 I_V = 8mA I_P = 2μA

$$f = \frac{1}{RC \ln(1/(1-\eta))} = \frac{1}{(100k)(10nF)\ln(1/(1-0.75))} = 1.39kHz$$

Figure 7.48: *Unijunction transistor relaxation oscillator and waveforms. Oscillator drives SCR.*

not affect the discharge time. The external resistor at B2 is optional. It may be replaced by a short circuit. The approximate frequency is given by 1/f = T = RC. A more accurate expression for frequency is given in Figure 7.48.

The charging resistor R_E must fall within certain limits. It must be small enough to allow I_P to flow based on the V_{BB} supply less V_P. It must be large enough to supply I_V based on the V_{BB} supply less V_V. [6] The equations and an example for a 2n2647:

2n2647 R_{BBO} = 4.7–9.1k η = 0.68—0.82 I_V = 8mA I_P = 2μA

$V_P = 0.7 + \eta V_{BB}$ $\quad$ $V_P = 0.7 + 0.75(10) = 8.2V$

$V_V = 0.10(V_{BB})$ $\quad$ $V_V = 0.10(10) = 1V$

$$\frac{V_{BB} - V_V}{I_V} < R_E < \frac{V_{BB} - V_P}{I_P} \qquad \frac{10 - 1}{8mA} < R_E < \frac{10 - 8.2}{2\mu A}$$

$$1.125k < R_E < 900k$$

Programmable Unijunction Transistor (PUT): Although the unijunction transistor is listed as obsolete (read expensive if obtainable), the programmable unijunction transistor is alive and well. It is inexpensive and in production. Though it serves a function similar to the unijunction transistor, the PUT is a three terminal thyristor. The PUT shares the four-layer structure typical of thyristors shown in Figure 7.49. Note that the gate, an N-type layer near the anode, is known as an "anode gate". Moreover, the gate lead on the schematic symbol is attached to the anode end of the symbol.

The characteristic curve for the programmable unijunction transistor in Figure 7.49 is similar to that of the unijunction transistor. This is a plot of anode current I_A versus anode voltage V_A. The gate lead voltage sets, programs, the peak anode voltage V_P. As anode current inceases, voltage increases up to the peak point. Thereafter, increasing current results in decreasing voltage, down to the valley point.

The PUT equivalent of the unijunction transistor is shown in Figure 7.50. External PUT resistors R1 and R2 replace unijunction transistor internal resistors R_{B1} and R_{B2}, respectively. These resistors allow the calculation of the intrinsic standoff ratio η.

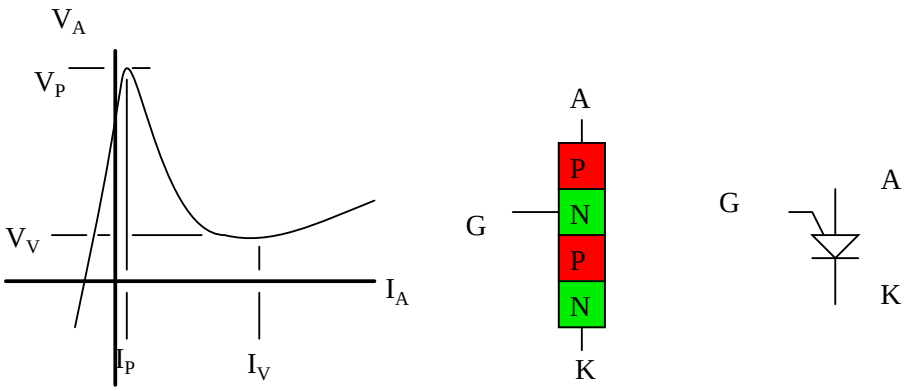


Figure 7.49: *Programmable unijunction transistor: Characteristic curve, internal construction, schematic symbol.*

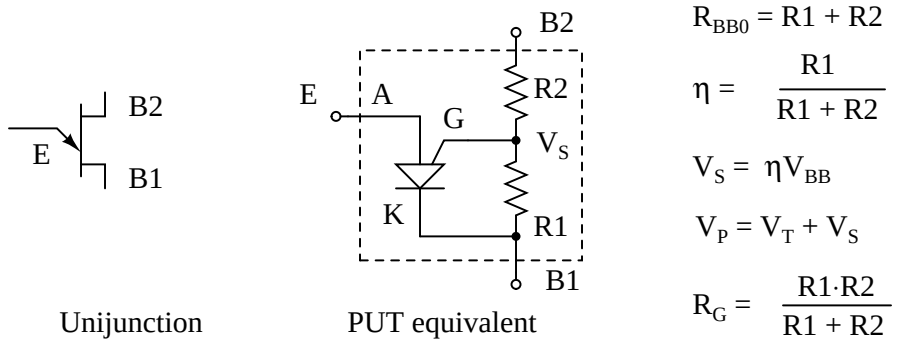


Figure 7.50: *PUT equivalent of unijunction transistor*

Figure 7.51 shows the PUT version of the unijunction relaxation oscillator Figure 7.48. Resistor R charges the capacitor until the peak point, Figure 7.49, then heavy conduction moves the operating point down the negative resistance slope to the valley point. A current spike flows through the cathode during capacitor discharge, developing a voltage spike across the cathode resistors. After capacitor discharge, the operating point resets back to the slope up to the peak point.

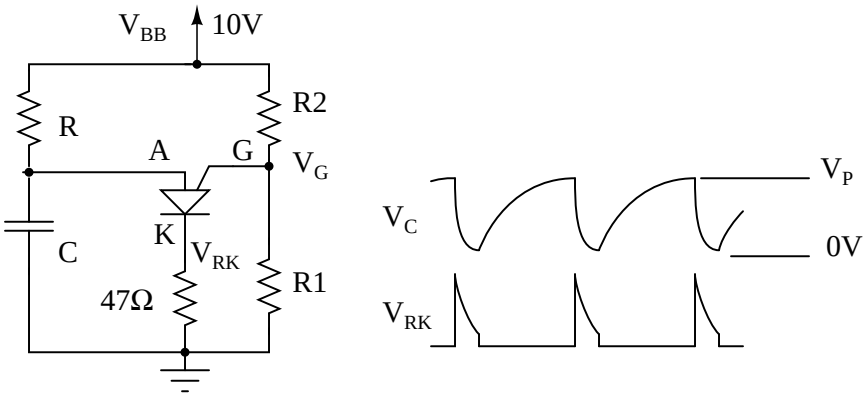


Figure 7.51: *PUT relaxation oscillator*

Problem: What is the range of suitable values for R in Figure 7.51, a relaxation oscillator? The charging resistor must be small enough to supply enough current to raise the anode to V_P the peak point (Figure 7.49) while charging the capacitor. Once V_P is reached, anode voltage decreases as current increases (negative resistance), which moves the operating point to the valley. It is the job of the capacitor to supply the valley current I_V. Once it is discharged, the operating point resets back to the upward slope to the peak point. The resistor must be large enough so that it will never supply the high valley current I_P. If the charging resistor ever could supply that much current, the resistor would supply the valley current after the capacitor was discharged and the operating point would never reset back to the high resistance condition to the left of the peak point.

We select the same V_{BB}=10V used for the unijunction transistor example. We select values of R1 and R2 so that η is about 2/3. We calculate η and V_S. The parallel equivalent of R1, R2 is R_G, which is only used to make selections from Table 7.1. Along with V_S=10, the closest value to our 6.3, we find V_T=0.6V, in Table 7.1 and calculate V_P.

$$R1 = 27k \quad R2 = 16k \qquad V_{BB} = 10V$$

$$\eta = \frac{R1}{R1 + R2} \qquad \eta = \frac{27}{27 + 16} = 0.6279$$

$$V_S = \eta V_{BB} \qquad V_S = 0.6279(10) = 6.279V$$

$$R_G = \frac{R1 \cdot R2}{R1 + R2} \qquad R_G = \frac{27k \cdot 16k}{27k + 16k} = 10k$$

For R_G=10k and V_S=10V, V_T = 0.6V

$$V_P = V_T + V_S \qquad V_P = 0.6 + 6.3 = 6.9V$$

We also find I_P and I_V, the peak and valley currents, respectively in Table 7.1. We still need V_V, the valley voltage. We used 10% of V_{BB}= 1V, in the previous unijunction example. Consulting the datasheet, we find the forward voltage V_F=0.8V at I_F=50mA. The valley current I_V=70μA is much less than I_F=50mA. Therefore, V_V must be less than V_F=0.8V. How much less? To be safe we set V_V=0V. This will raise the lower limit on the resistor range a little.

For R_G=10k and V_S=10V, I_P = 4.0µA

For R_G=10k and V_S=10V, I_V = 70µA

$$V_V = 0.10(V_{BB}) \text{ not used} \qquad V_V = 0V$$

$$\frac{V_{BB} - V_V}{I_V} < R_E < \frac{V_{BB} - V_P}{I_P} \qquad \frac{10 - 0}{70\mu A} < R_E < \frac{10 - 6.9}{4\mu A}$$

$$143k < R_E < 755k$$

Choosing R $>$ 143k guarantees that the operating point can reset from the valley point after capacitor discharge. R $<$ 755k allows charging up to V_P at the peak point.

Table 7.1: *Selected 2n6027 PUT parameters, adapted from 2n6027 datasheet. [4]*

Parameter	**Conditions**	**min**	**typical**	**max**	**units**
V_T				V	
	V_S=10V, R_G=1Meg	0.2	0.7	1.6	
	V_S=10V, R_G=10k	0.2	0.35	0.6	
I_P					μA
	V_S=10V, R_G=1Meg	-	1.25	2.0	
	V_S=10V, R_G=10k	-	4.0	5.0	
I_V					μA
	V_S=10V, R_G=1Meg	-	18	50	
	V_S=10V, R_G=10k	70	150	-	
	V_S=10V, R_G=200Ω	1500	-	-	
V_F	I_F=50mA	-	0.8	1.5	V

Figure 7.52 show the PUT relaxation oscillator with the final resistor values. A practical application of a PUT triggering an SCR is also shown. This circuit needs a V_{BB} unfiltered

supply (not shown) divided down from the bridge rectifier to reset the relaxation oscillator after each power zero crossing. The variable resistor should have a minimum resistor in series with it to prevent a low pot setting from hanging at the valley point.

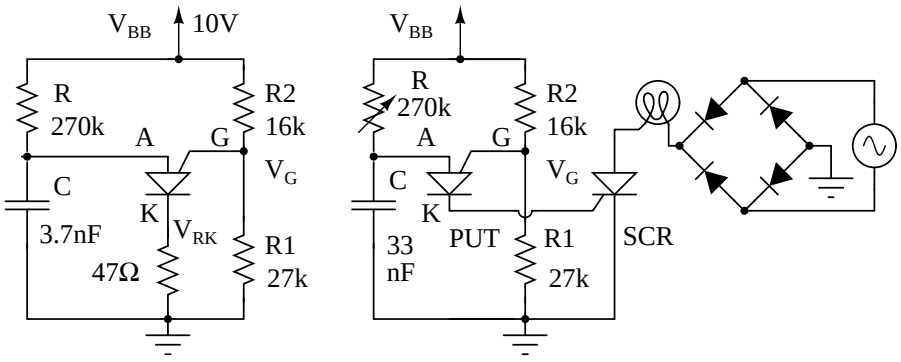


Figure 7.52: *PUT relaxation oscillator with component values. PUT drives SCR lamp dimmer.*

PUT timing circuits are said to be usable to 10kHz. If a linear ramp is required instead of an exponential ramp, replace the charging resistor with a constant current source such as a FET based constant current diode. A substitute PUT may be built from a PNP and NPN silicon transistor as shown for the SCS equivalent circuit in Figure 7.53 by omitting the cathode gate and using the anode gate.

- **REVIEW:**
- A unijunction transistor consists of two bases (B1, B2) attached to a resistive bar of silicon, and an emitter in the center. The E-B1 junction has negative resistance properties; it can switch between high and low resistance.
- A PUT (programmable unijunction transistor) is a 3-terminal 4-layer thyristor acting like a unijunction transistor. An external resistor network "programs" η.
- The intrinsic standoff ratio is η=R1/(R1+R2) for a PUT; substitute R_{B1} and R_{B2}, respectively, for a unijunction transistor. The trigger voltage is determined by η.
- Unijunction transistors and programmable unijunction transistors are applied to oscillators, timing circuits, and thyristor triggering.

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*: (Figure 7.53)

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

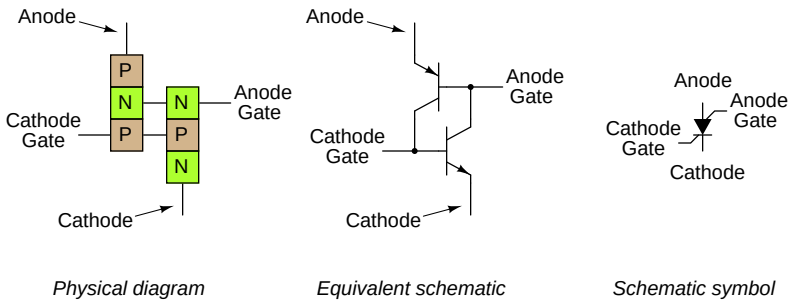


Figure 7.53: *The Silicon-Controlled Switch(SCS)*

current through the device has not yet fallen below the holding current value. Note that the motor is in the anode gate circuit in Figure 7.54. This is correct, although it doesn't look right. The anode lead is required to switch the SCS off. Therefore the motor cannot be in series with the anode.

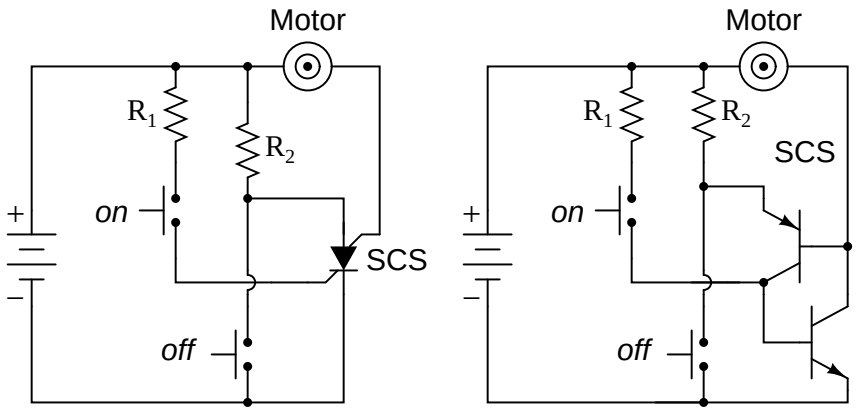


Figure 7.54: *SCS: Motor start/stop circuit, equivalent circuit with two transistors.*

When the "on" pushbutton switch is actuated, the voltage applied between the cathode gate and the cathode, forward-biases 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. [3] 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 standard 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: (Figure 7.55)

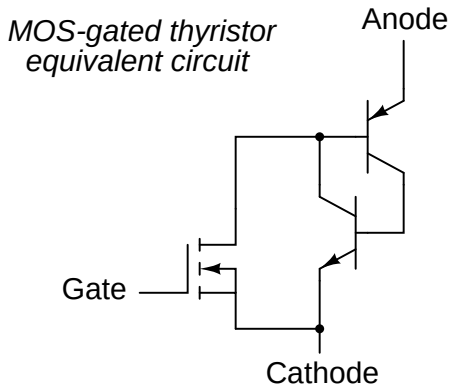


Figure 7.55: *MOS-gated thyristor equivalent circuit*

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: (Figure 7.56)

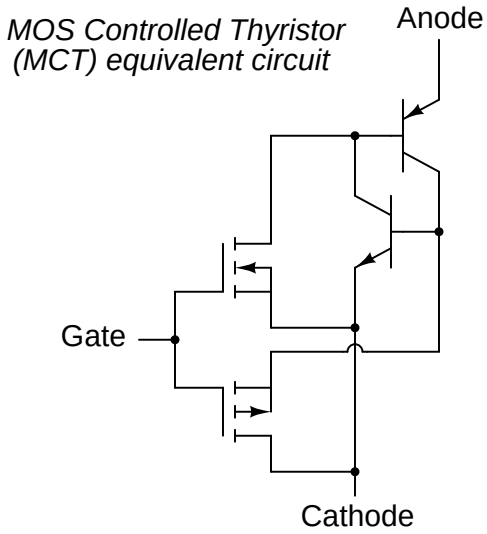


Figure 7.56: *MOS-controlled thyristor (MCT) equivalent circuit*

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 as 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 zero voltage is 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; 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).

Bibliography

[1] "Phattytron PT-1 Vacuum Tube Synthesizer", The Audio Playground Synthesizer Museum at http://www.keyboardmuseum.com/ar/m/meta/pt1.html

[2] "At last, a pitch source with tube power", METASONIX, PMB 109, 881 11th Street, Lakeport CA 95453 USA at http://www.metasonix.com/index.php?option=com_content&task=view&id=14&Itemid=31

[3] "Silicon Contolled Switches", GE Transistor Manual, The General Electric Company, 1964, Figure 16.19(M).

[4] "2N6027, 2N6028 Programmable Unijunction Transistor ", datasheet at http://www.onsemi.com/pub_link/Collateral/2N6027-D.PDF

[5] "Unijunction Transistor ", American Microsemiconductor, at http://www.americanmicrosemi.com/tutorials/unijunction.htm

[6] Matthew H. Williams, "Unijunction Transistor ", at http://baec.tripod.com/DEC90/uni_tran.htm Unijunction Transistor by http://baec.tripod.com/DEC90/uni_tran.htm

Chapter 8

OPERATIONAL AMPLIFIERS

Contents

8.1 Introduction . . . 353
8.2 Single-ended and differential amplifiers . . . 354
8.3 The "operational" amplifier . . . 358
8.4 Negative feedback . . . 364
8.5 Divided feedback . . . 367
8.6 An analogy for divided feedback . . . 370
8.7 Voltage-to-current signal conversion . . . 376
8.8 Averager and summer circuits . . . 378
8.9 Building a differential amplifier . . . 380
8.10 The instrumentation amplifier . . . 382
8.11 Differentiator and integrator circuits . . . 383
8.12 Positive feedback . . . 386
8.13 Practical considerations . . . 390
8.13.1 Common-mode gain . . . 391
8.13.2 Offset voltage . . . 394
8.13.3 Bias current . . . 396
8.13.4 Drift . . . 402
8.13.5 Frequency response . . . 402
8.13.6 Input to output phase shift . . . 403
8.14 Operational amplifier models . . . 404
8.15 Data . . . 411

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

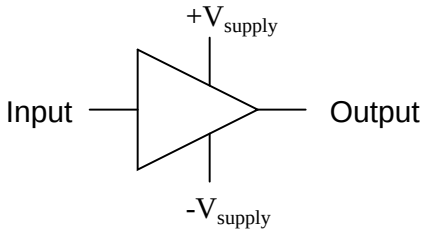
of analog signal processing tasks. It is also quite affordable, most general-purpose amplifiers selling for under a dollar apiece. Modern designs have been engineered with durability in mind as well: several "op-amps" are manufactured that can sustain direct short-circuits on their outputs without damage.

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.

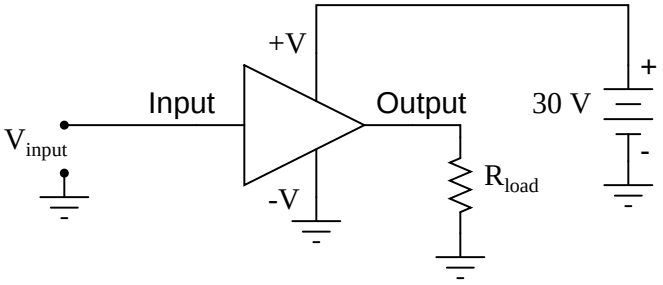
8.2 Single-ended and differential amplifiers

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



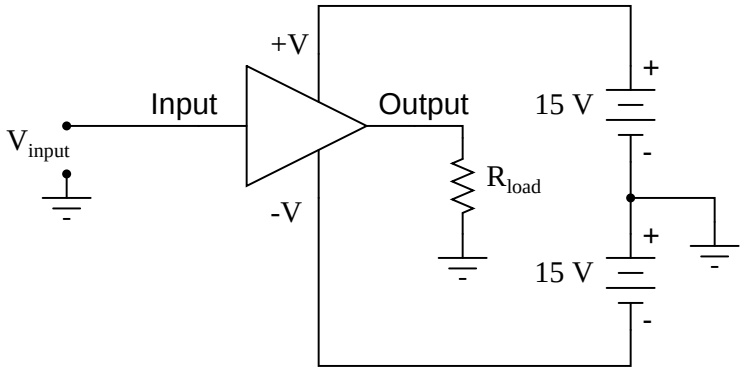
The +V and -V connections denote the positive and negative sides of the DC power supply, respectively. The input and output voltage connections are shown as single conductors, because it is assumed that all signal voltages are referenced to a common connection in the circuit called *ground*. Often (but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

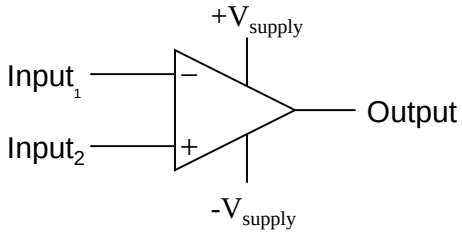
If it is necessary for an amplifier to be able to output true AC voltage (reversing polarity) to the load, a *split* DC power supply may be used, whereby the ground point is electrically "centered" between +V and -V. Sometimes the split power supply configuration is referred to as a *dual* power supply.



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

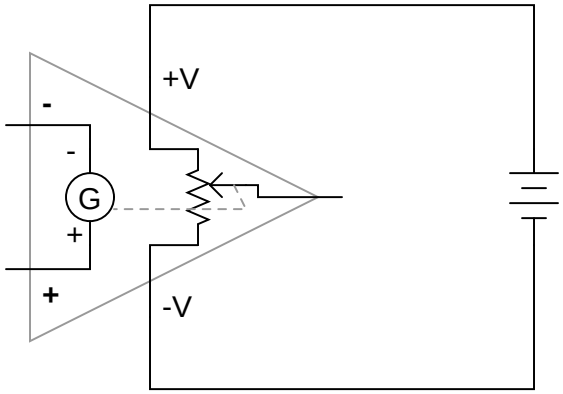
Voltage output equation: $V_{out} = A_V(Input_2 - Input_1)$

or

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

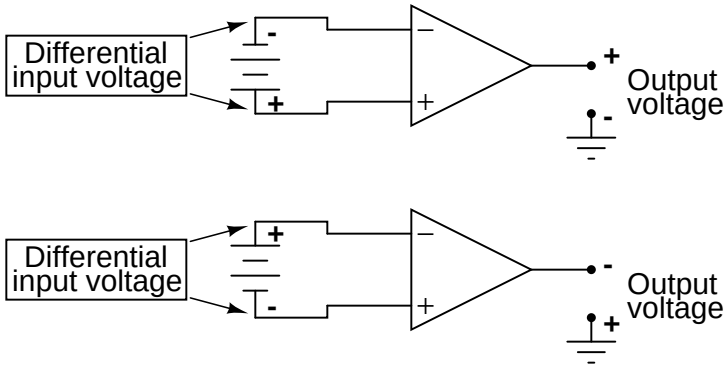
An increasingly positive voltage on the (+) input tends to drive the output voltage more positive, and an increasingly positive voltage on the (-) input tends to drive the output voltage more negative. Likewise, an increasingly negative voltage on the (+) input tends to drive the output negative as well,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

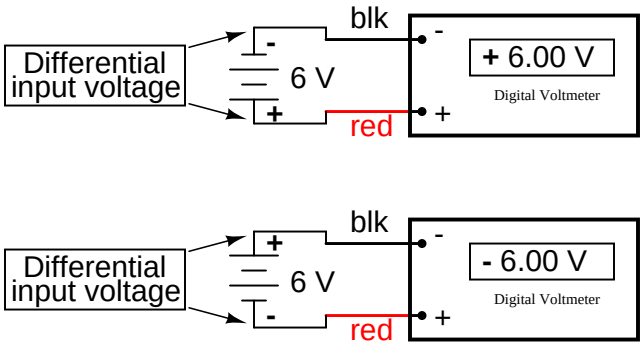


Bear in mind that the above illustration is only a *model* to aid in understanding the behavior of a differential amplifier. It is not a realistic schematic of its actual design. The "G" symbol represents a galvanometer, a sensitive voltmeter movement. The potentiometer connected between +V and -V provides 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, its 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.

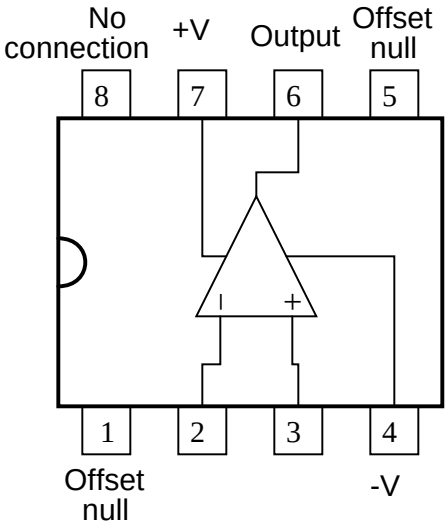
Electronic circuits are very easy and inexpensive to create compared to complex physical systems, so this kind of analog electronic simulation was widely used in the research and development of mechanical systems. For realistic simulation, though, amplifier circuits of high accuracy and easy configurability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

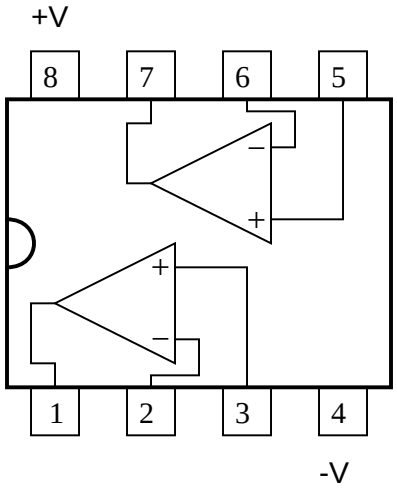
The following diagram shows the pin connections for single op-amps (741 included) when housed in an 8-pin DIP (**D**ual **I**nline **P**ackage) integrated circuit:

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

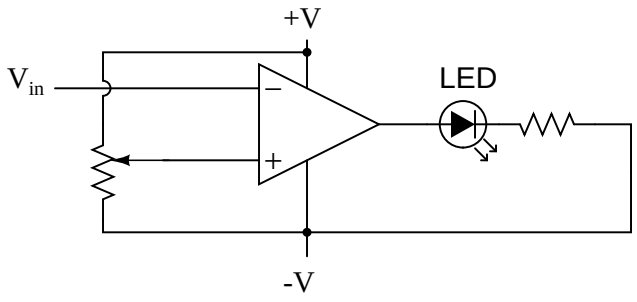


Operational amplifiers are also available four to a package, usually in 14-pin DIP arrangements. Unfortunately, pin assignments aren’t as standard for these ”quad” op-amps as they are for the ”dual” or single units. Consult the manufacturer datasheet(s) for details.

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

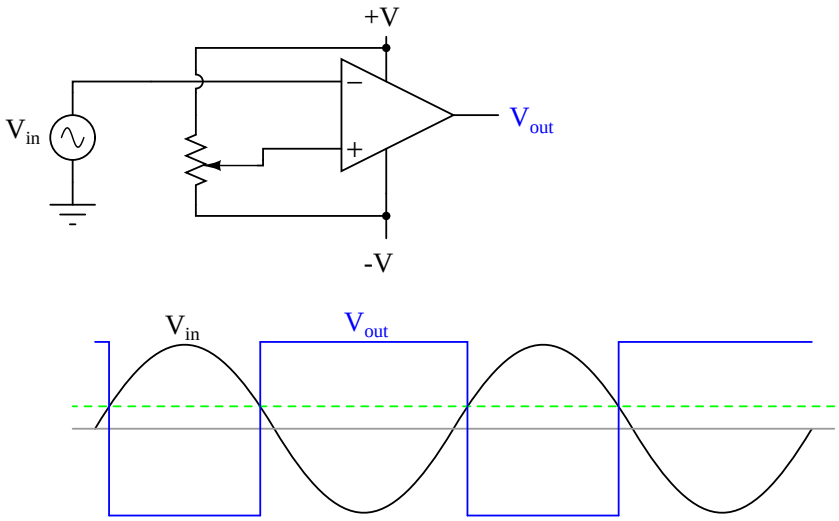
makes them almost useless as an analog differential amplifier by themselves. For an op-amp with a voltage gain (A_V) of 200,000 and a maximum output voltage swing of +15V/-15V, all it would take is a differential input voltage of 75 μV (microvolts) to drive it to saturation or cutoff! Before we take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

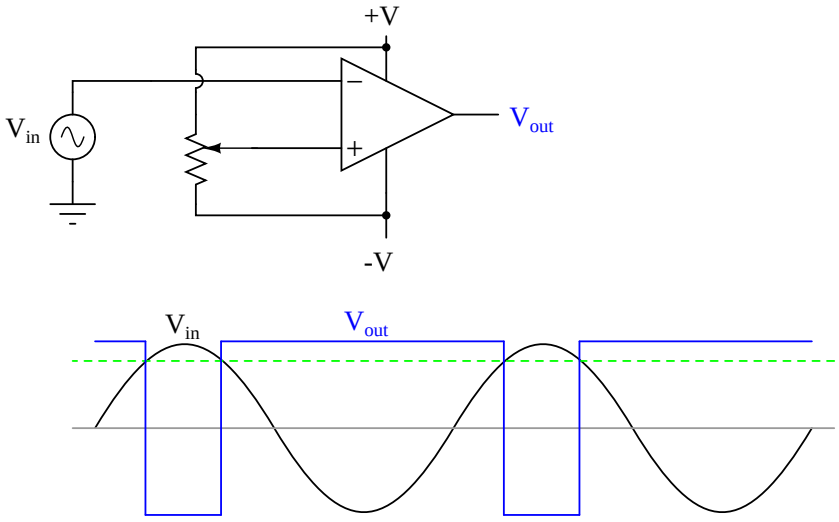


In the above circuit, we have an op-amp connected as a comparator, comparing the input voltage with a reference voltage set by the potentiometer (R_1). If V_{in} drops below the voltage set by R_1, the op-amp's output will saturate to +V, thereby lighting up the LED. Otherwise, if V_{in} is above the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



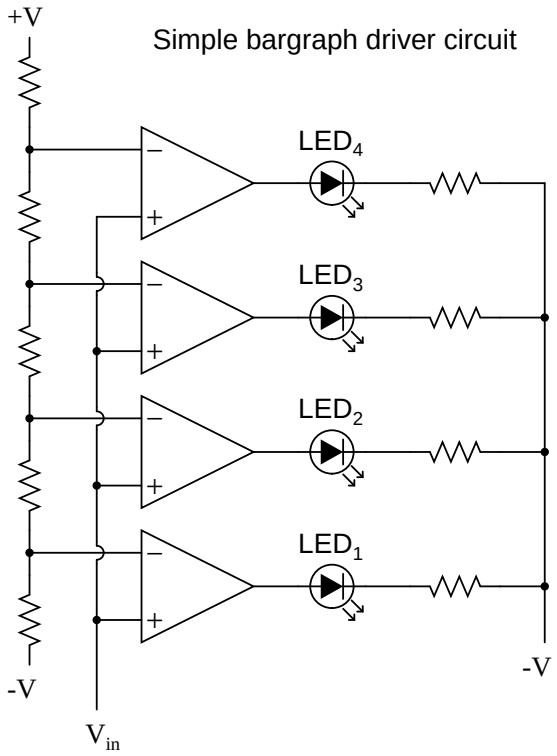
Adjustments to the potentiometer setting would change the reference voltage applied to the noninverting (+) input, which would change the points at which the sine wave would cross, changing the on/off times, or *duty cycle* of the square wave:



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.

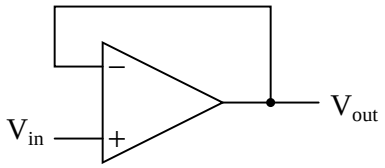
This very same technology is used in some analog-to-digital signal converters, namely the *flash converter*, to translate an analog signal quantity into a series of on/off voltages representing a digital number.

- **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):



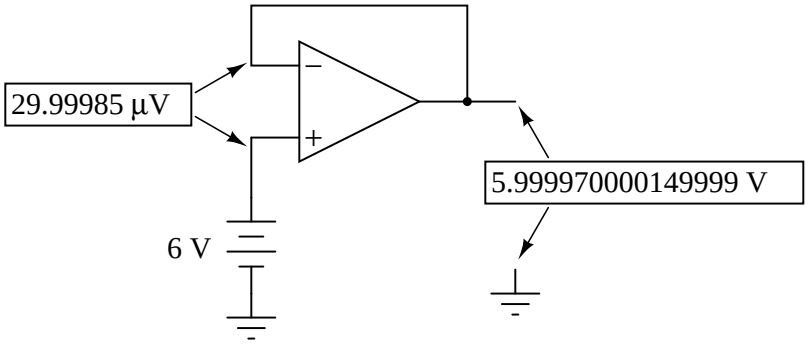
As V_{in} increases, V_{out} will increase in accordance with the differential gain. However, as V_{out} increases, that output voltage is fed back to the inverting input, thereby acting to decrease the voltage differential between inputs, which acts to bring the output down. What will happen for any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

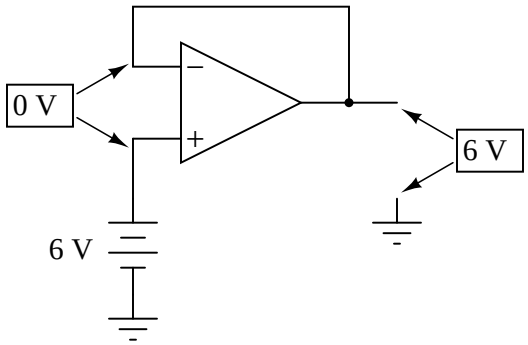
opposed to merely being saturated fully "on" or "off" as it was when used as a comparator, with no feedback at all.

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



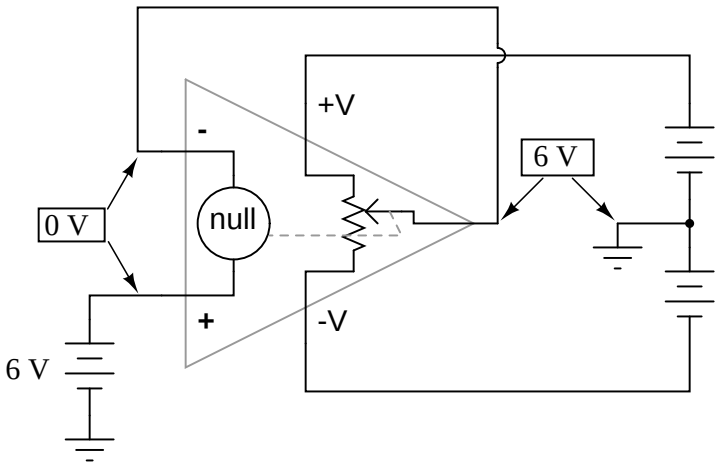
The effects of negative feedback (rounded figures)



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 its 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

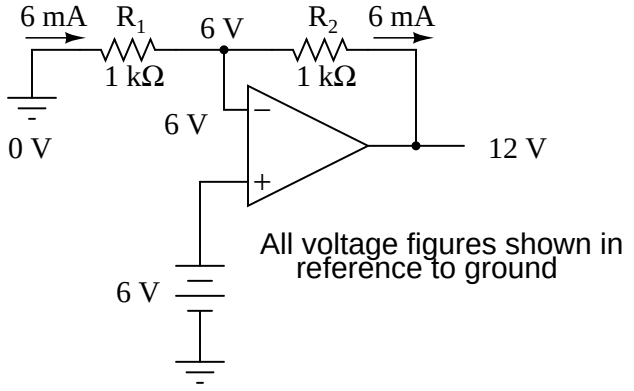
as the *positive saturation voltage* and *negative saturation voltage*, respectively. Other op-amps, such as the model 3130 with field-effect transistors in the final output stage, have the ability to swing their output voltages within millivolts of either power supply *rail* voltage. Consequently, their 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):

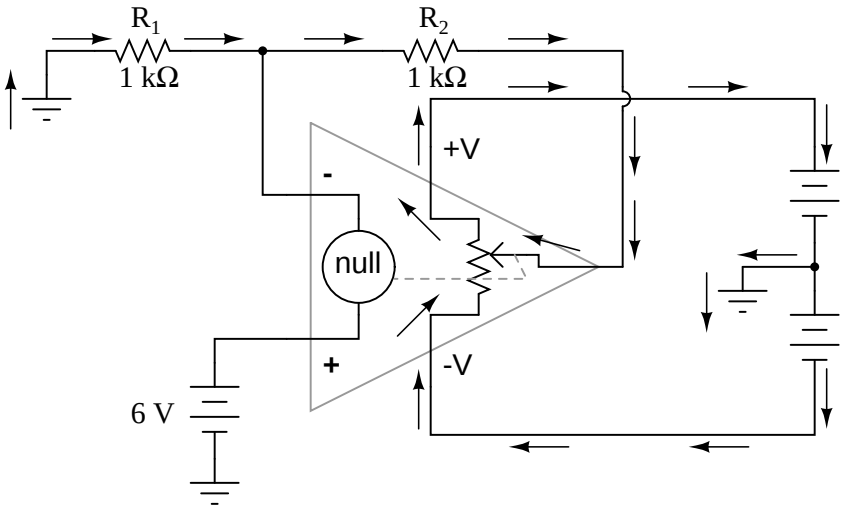
The effects of divided negative feedback



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 6 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:

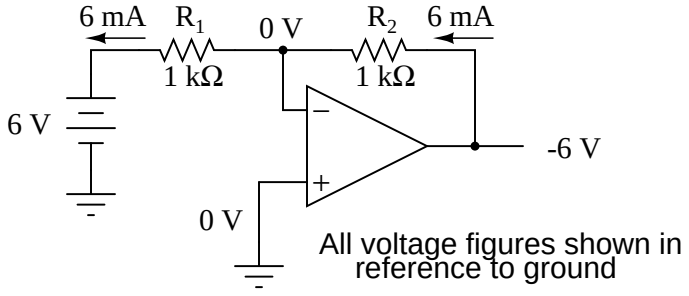
$$A_V = \frac{R_2}{R_1} + 1$$

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 its 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.

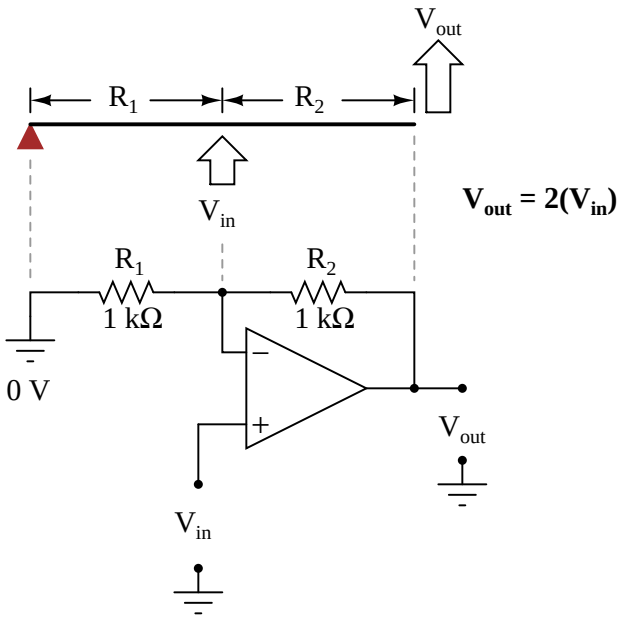
- A negative-feedback op-amp circuit with the input signal going to the noninverting (+) input is called a *noninverting amplifier*. The output voltage will be the same polarity as the input. Voltage gain is given by the following equation: $A_V = (R_2/R_1) + 1$

- 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: $A_V = -R_2/R_1$

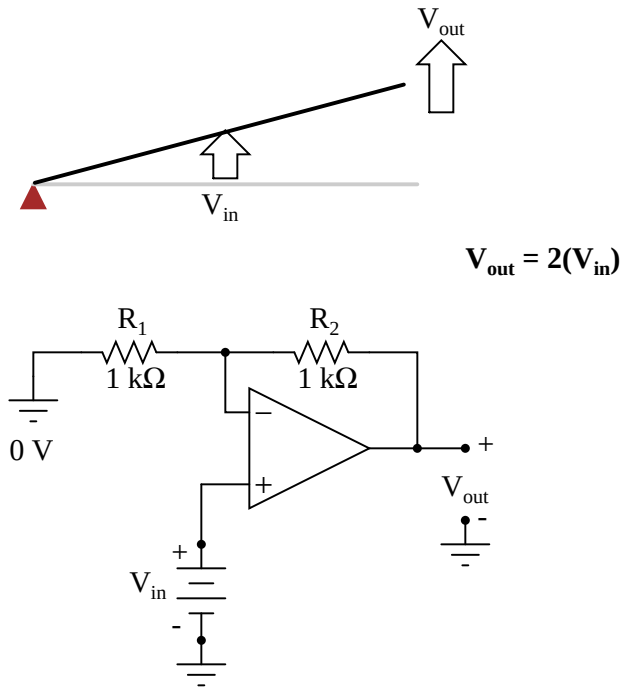
8.6 An analogy for divided feedback

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.

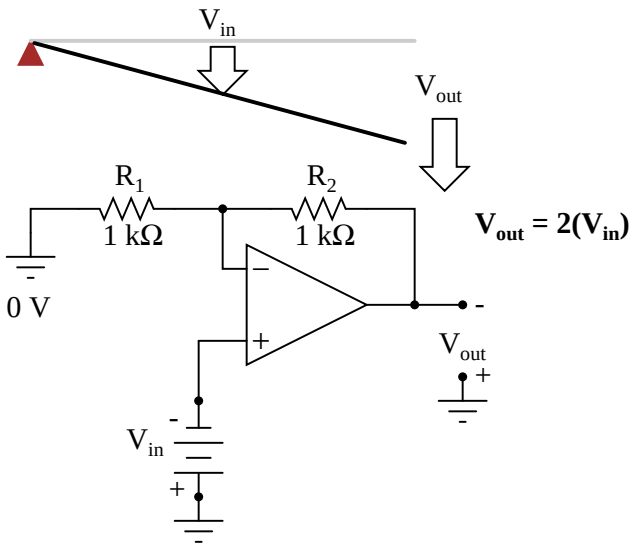
Take for example the following noninverting op-amp circuit. We know from the prior section that the voltage gain of a noninverting amplifier configuration can never be less than unity (1). If we draw a lever diagram next to the amplifier schematic, with the distance between fulcrum and lever ends representative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



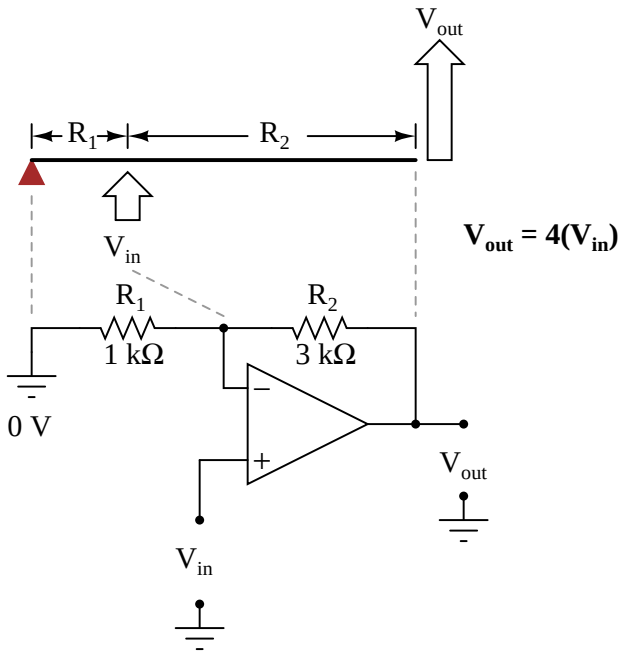
Physicists call this type of lever, with the input force (effort) applied between the fulcrum and output (load), a *third-class* lever. It is characterized by an output displacement (motion) at least as large than the input displacement – a "gain" of at least 1 – and in the same direction. Applying a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



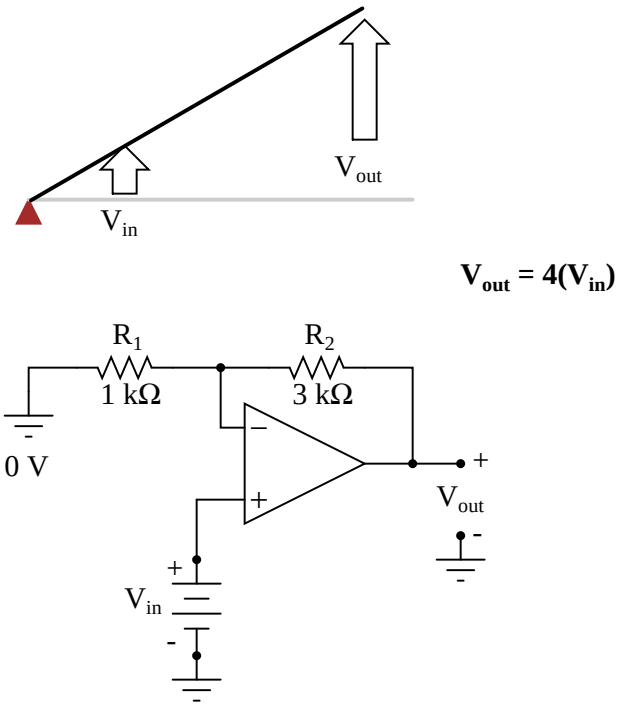
Due to the displacement-amplifying characteristics of the lever, the "output" point will move twice as far as the "input" point, and in the same direction. In the electronic circuit, the output voltage will equal twice the input, with the same polarity. Applying a negative input voltage is analogous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



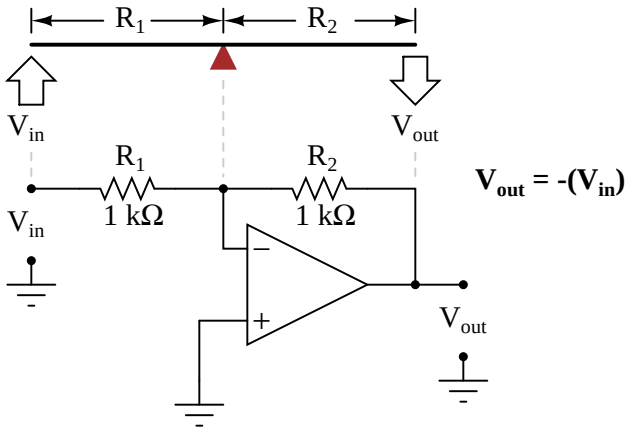
If we alter the resistor ratio R_2/R_1, we change the gain of the op-amp circuit. In lever terms, this means moving the input point in relation to the fulcrum and lever end, which similarly changes the displacement "gain" of the machine:



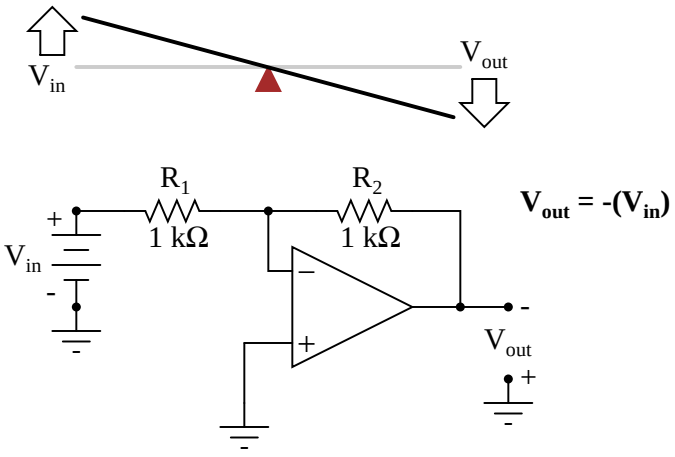
Now, any input signal will become amplified by a factor of four instead of by a factor of two:



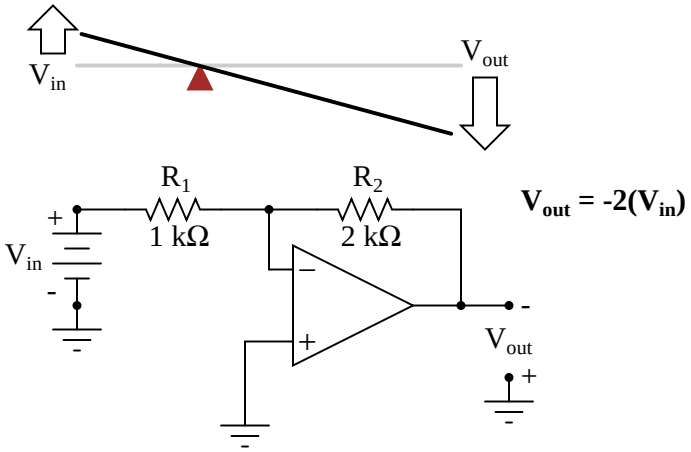
Inverting op-amp circuits may be modeled using the lever analogy as well. With the inverting configuration, the ground point of the feedback voltage divider is the op-amp's inverting input with the input to the left and the output to the right. This is mechanically equivalent to a *first-class* lever,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



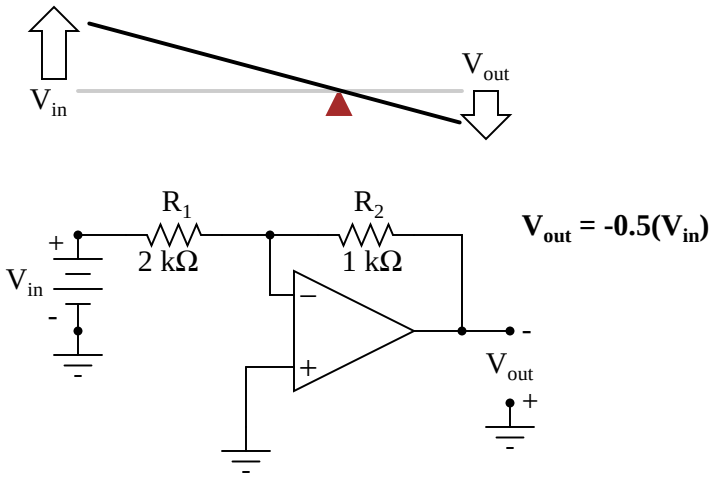
With equal-value resistors (equal-lengths of lever on each side of the fulcrum), the output voltage (displacement) will be equal in magnitude to the input voltage (displacement), but of the opposite polarity (direction). A positive input results in a negative output:



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:



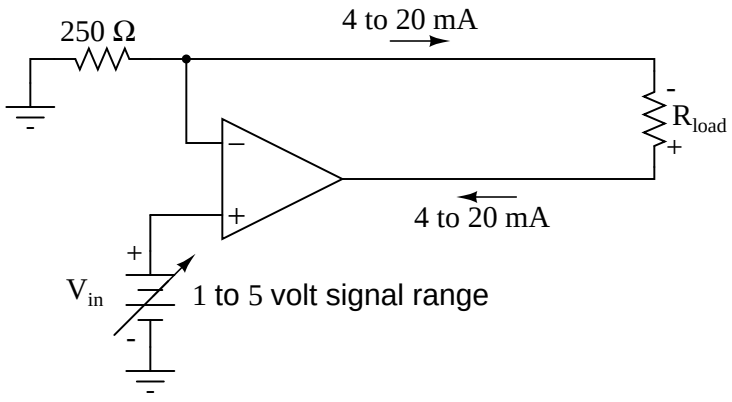
With the inverting amplifier configuration, though, gains of less than 1 are possible, just as with first-class levers. Reversing R_2 and R_1 values is analogous to moving the fulcrum to its complementary position on the lever: one-third of the way from the output end. There, the output displacement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

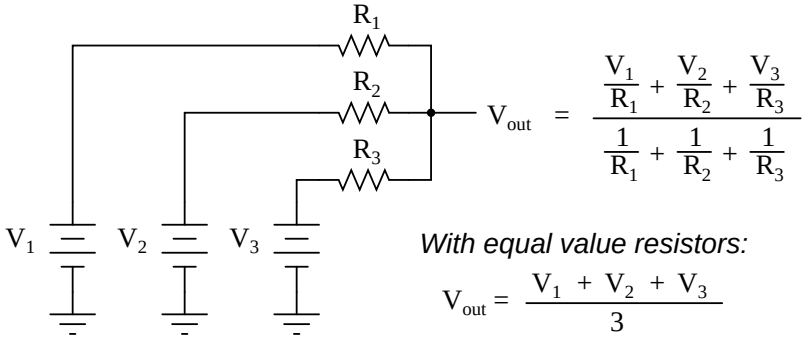
- In industry, DC current signals are often used in preference to DC voltage signals as analog representations of physical quantities. Current in a series circuit is absolutely equal at all points in that circuit regardless of wiring resistance, whereas voltage in a parallel-connected circuit may vary 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.

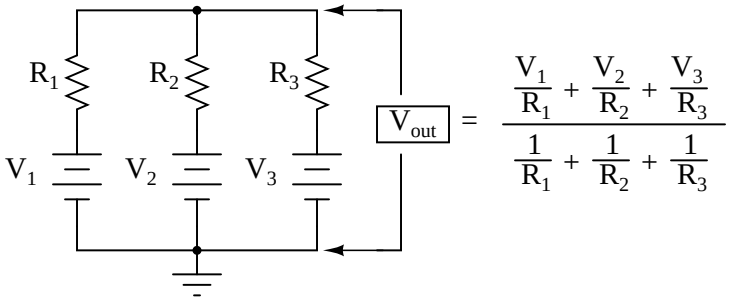
8.8 Averager and summer circuits

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.

"Passive averager" circuit



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



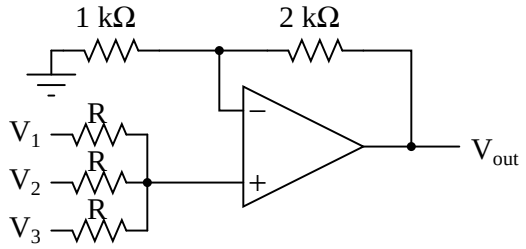
This circuit is commonly known as a *passive averager*, because it generates an average voltage with non-amplifying components. *Passive* simply means that it is an unamplified circuit. The large equation to the right of the averager circuit comes from Millman's Theorem, which describes the voltage produced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

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

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

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

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

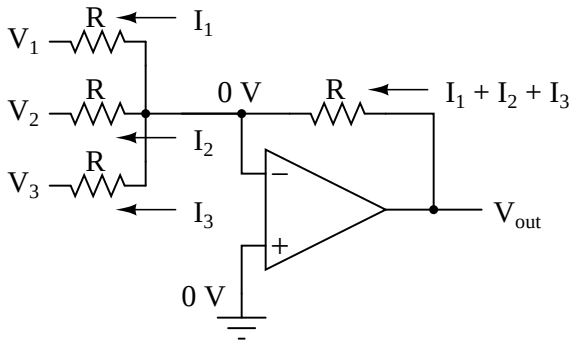


With a voltage divider composed of a 2 kΩ / 1 kΩ combination, the noninverting amplifier circuit will have a voltage gain of 3. By taking the voltage from the passive averager, which is the sum of V_1, V_2, and V_3 divided by 3, and multiplying that average by 3, we arrive at an output voltage equal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

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

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

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



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

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

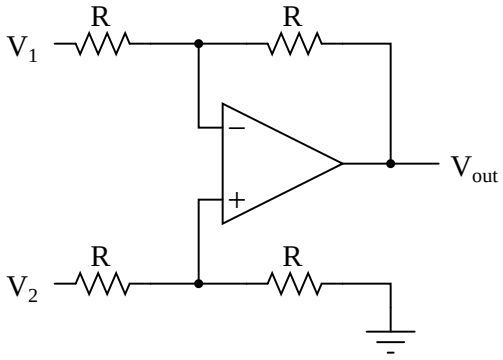
Summer (adder) circuits are quite useful in analog computer design, just as multiplier and divider circuits would be. Again, it is the extremely high differential gain of the op-amp which allows us to build these useful circuits with a bare minimum of components.

- **REVIEW:**

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

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.

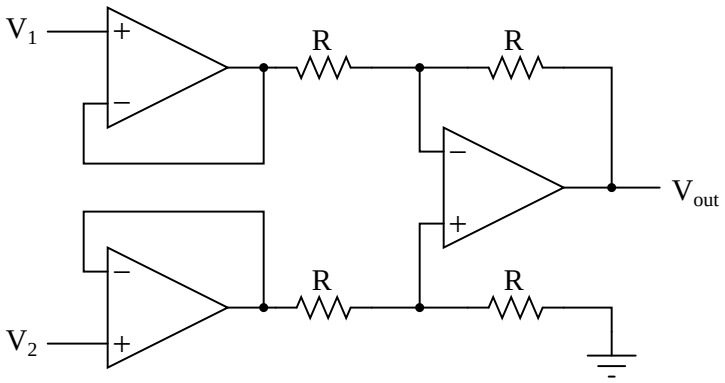


If all the resistor values are equal, this amplifier will have a differential voltage gain of 1. The analysis of this circuit is essentially the same as that of an inverting amplifier, except that the noninverting input (+) of the op-amp is at a voltage equal to a fraction of V_2, rather than being connected 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.

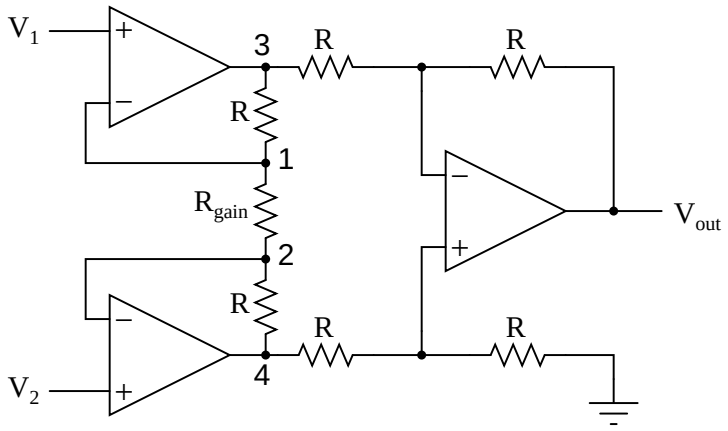
Another limitation of this amplifier design is the fact that its input impedances are rather low compared to that of some other op-amp configurations, most notably the noninverting (single-ended input) amplifier. Each input voltage source has to drive current through a resistance, which constitutes far 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:

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

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

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

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,

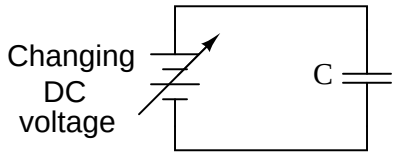
but this would necessitate *balanced* resistor value changes for the circuit to remain symmetrical. Please note that the lowest gain possible with the above circuit is obtained with R_{gain} completely open (infinite resistance), and that gain value is 1.

- **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:



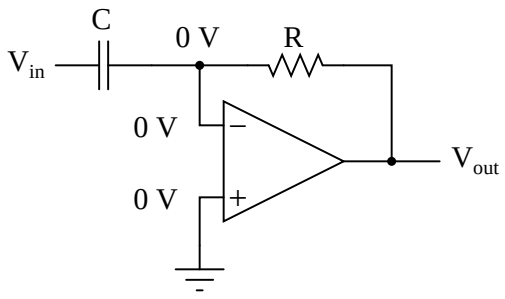
$$i = C \frac{dv}{dt}$$

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:

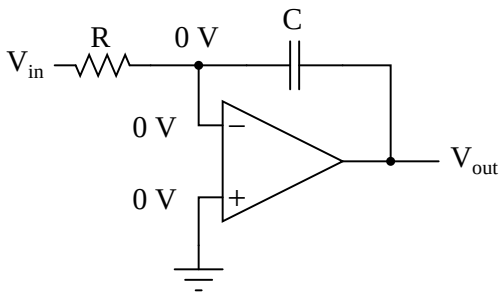
$$V_{out} = -RC \frac{dv_{in}}{dt}$$

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.

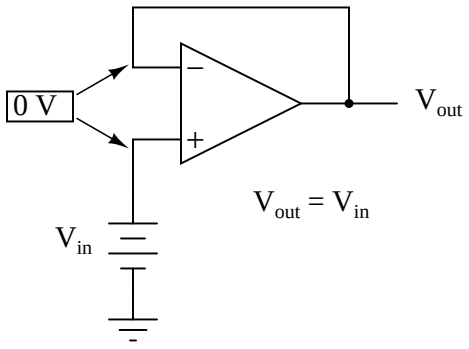
8.12 Positive feedback

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

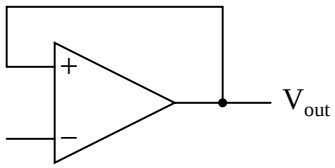


Whether the output is directly fed back to the inverting (-) input or coupled through a set of components, the effect is the same: the extremely high differential voltage gain of the op-amp will be "tamed" and the circuit will respond according to the dictates of the feedback "loop" connecting output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

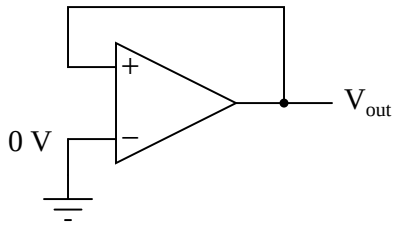
Another type of feedback, namely *positive feedback*, also finds application in op-amp circuits. Unlike negative feedback, where the output voltage is "fed back" to the inverting (-) input, with positive feedback the output voltage is somehow routed back to the noninverting

(+) 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:

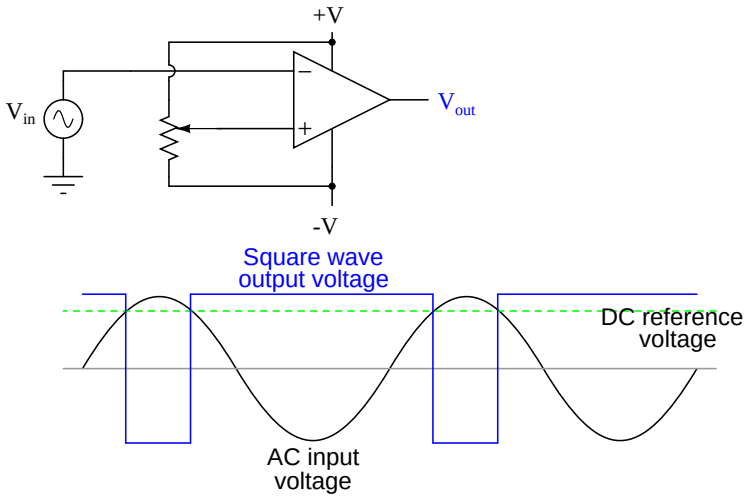


With the inverting input grounded (maintained at zero volts), the output voltage will be dictated by the magnitude and polarity of the voltage at the noninverting input. If that voltage happens to be positive, the op-amp will drive its output positive as well, feeding that positive voltage back to the 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.

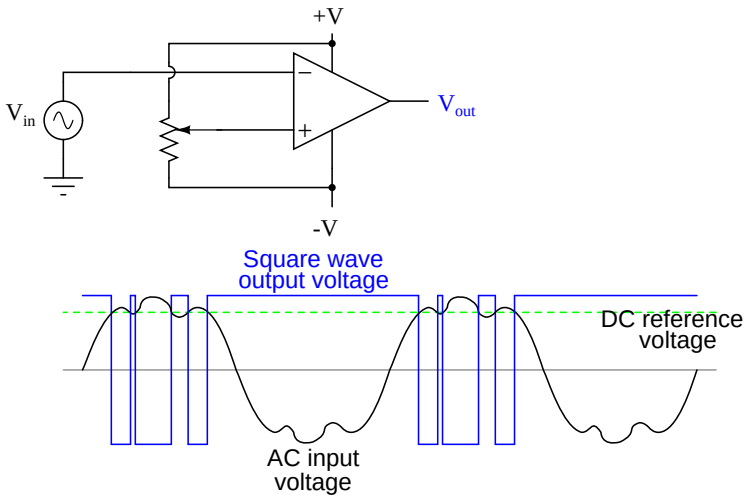
So, an op-amp with positive feedback tends to stay in whatever output state its already in. It "latches" between one of two states, saturated positive or saturated negative. Technically, this is known as *hysteresis*.

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.



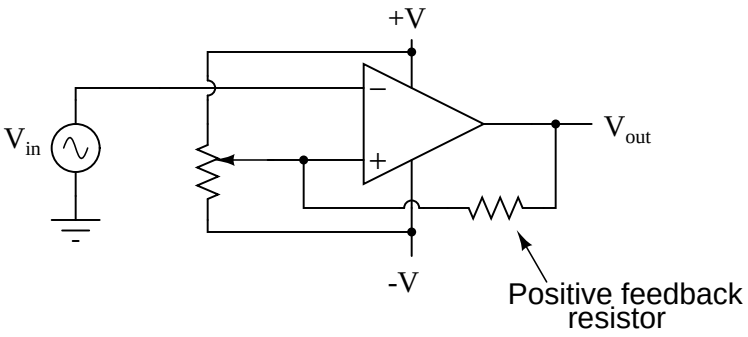
A "clean" AC input waveform produces predictable transition points on the output voltage square wave

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:

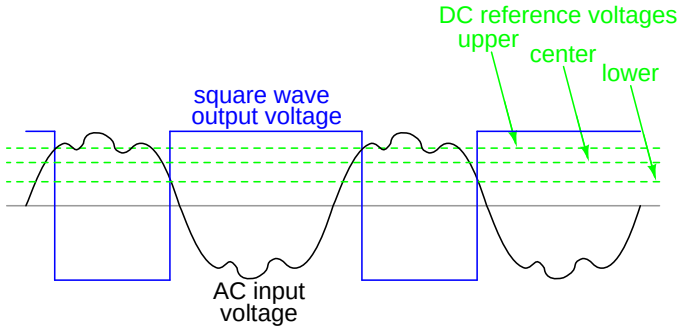


Any time there is a transition through the reference voltage level, no matter how tiny that transition may be, the output of the comparator will switch states, producing a square wave with ”glitches.”

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.



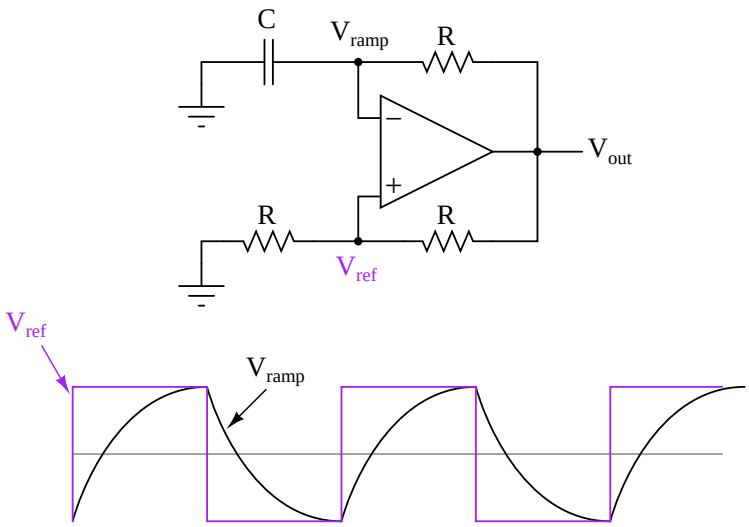
What this feedback resistor creates is a dual-reference for the comparator circuit. The voltage applied to the noninverting (+) input as a reference which to compare with the incoming AC voltage changes depending on the value of the op-amp's output voltage. When the op-amp output is saturated positive, 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.

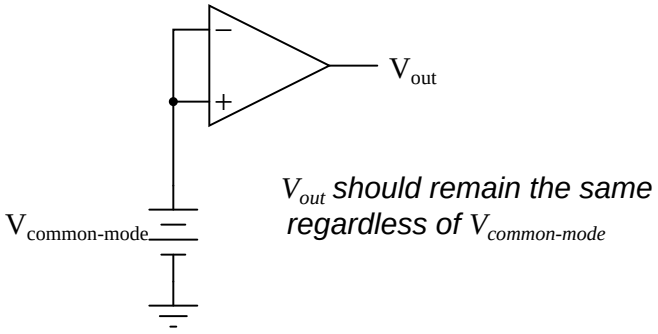
- **REVIEW:**
- Negative feedback creates a condition of *equilibrium* (balance). Positive feedback creates a condition of *hysteresis* (the tendency to "latch" in one of two extreme states).
- An *oscillator* is a device producing an alternating or pulsing output voltage.

8.13 Practical considerations

Real operational have some imperfections compared to an "ideal" model. A real device deviates from a perfect difference amplifier. One minus one may not be zero. It may have have an offset like an analog meter which is not zeroed. The inputs may draw current. The characteristics may drift with age and temperature. Gain may be reduced at high frequencies, and phase may shift from input to output. These imperfection may cause no noticable errors in some applications, unacceptable errors in others. In some cases these errors may be compensated for. Sometimes a higher quality, higher cost device is required.

8.13.1 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:

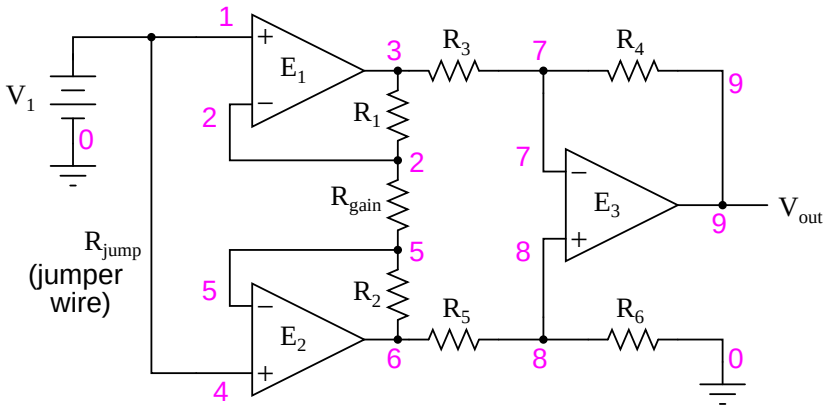
$$\text{CMRR} = \frac{\text{Differential } A_V}{\text{Common-mode } A_V}$$

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, its 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.

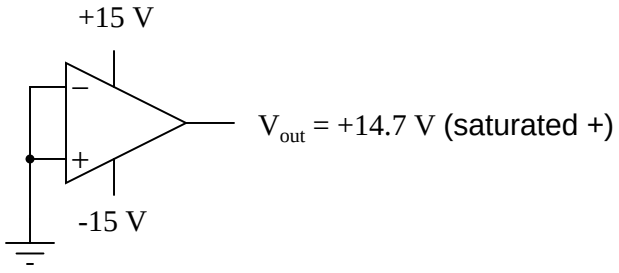
One quirk of some op-amp models is that of output *latch-up*, usually caused by the common-mode input voltage exceeding allowable limits. If the common-mode voltage falls outside of the manufacturer's specified limits, the output may suddenly "latch" in the high mode (saturate at full output voltage). 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!

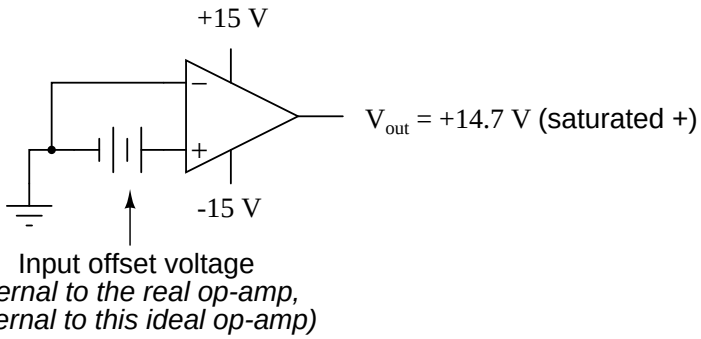
8.13.2 Offset voltage

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*.

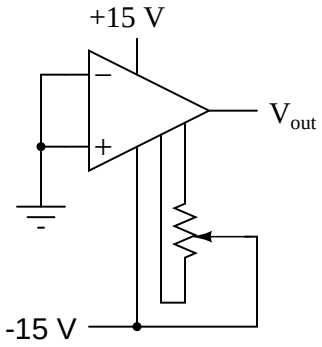


A perfect op-amp would output exactly zero volts with both its inputs shorted together and grounded. However, most op-amps off the shelf will drive their outputs to a saturated level, either negative or positive. In the example shown above, the output voltage is saturated at a value of positive 14.7 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:



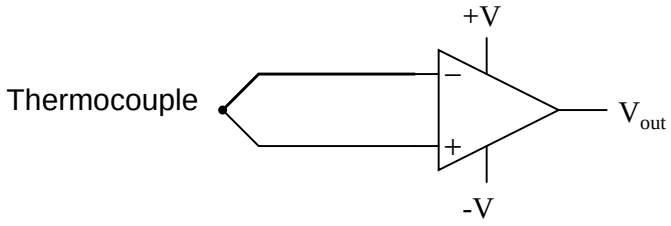
Potentiometer adjusted so that
V_{out} = 0 volts with inputs shorted together

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.

8.13.3 Bias current

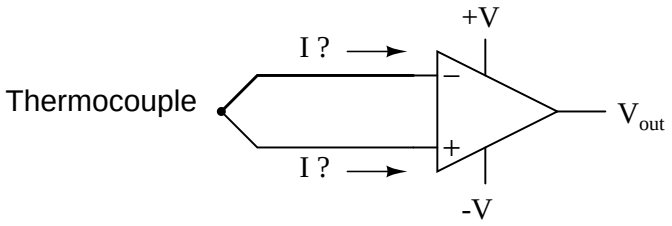
Inputs on an op-amp have extremely high input impedances. That is, the input currents entering or exiting an op-amp's two input signal connections are extremely small. For most purposes of op-amp circuit analysis, we treat them as though they don't exist at all. We analyze the circuit as though there 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:



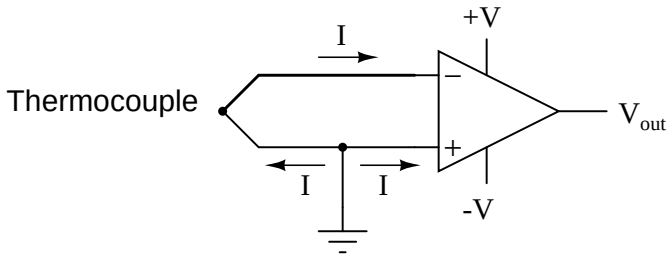
At first glance, we see no apparent problems with this circuit. A thermocouple, generating a small voltage proportional to temperature (actually, a voltage proportional to the *difference* in temperature between the measurement junction and the "reference" junction formed when the alloy thermocouple 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*

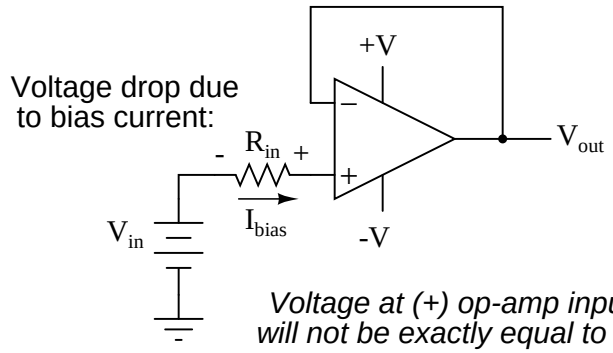
In order for this circuit to work properly, we must ground one of the input wires, thus providing a path to (or from) ground for both currents:



*This comparator circuit **will** work*

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:



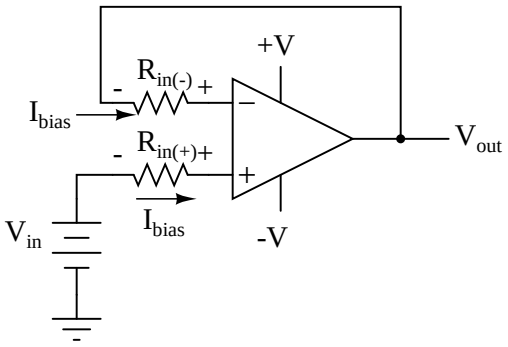
Voltage at (+) op-amp input will not be exactly equal to V_{in}

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.

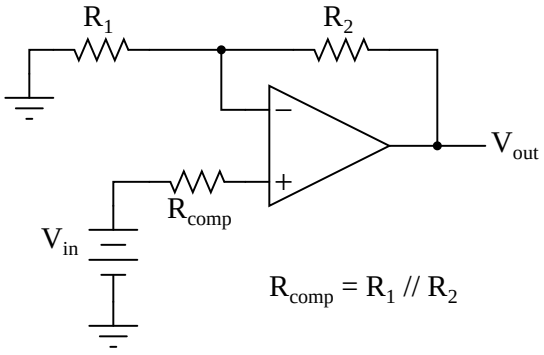
One way to do so is based on the assumption that the two input bias currents will be the same. In reality, they are often close to being the same, the difference between them referred to as the *input offset current*. If they are the same, then we should be able to cancel out the effects of input resistance voltage drop by inserting an equal amount of resistance in series with the other input, like this:



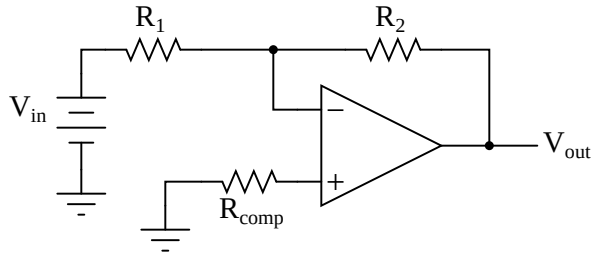
With the additional resistance added to the circuit, the output voltage will be closer to V_{in} than before, even if there is some offset between the two input currents.

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



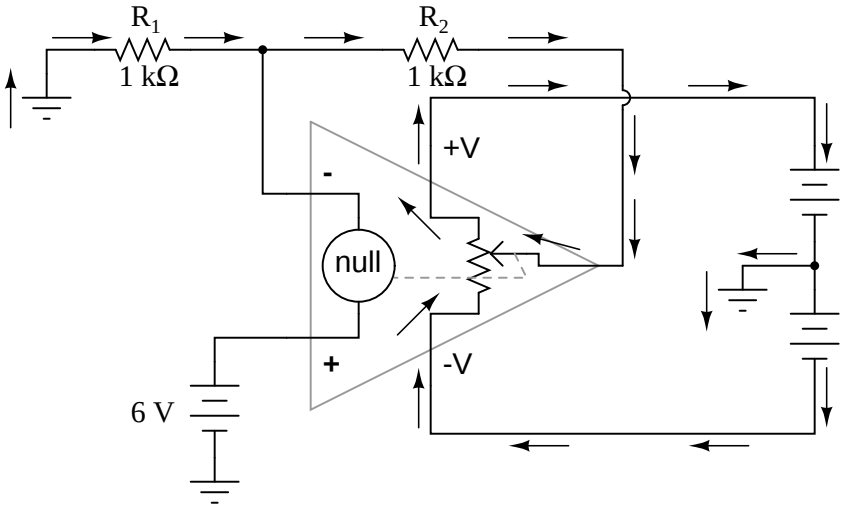
Inverting amplifier with compensating resistor



$$R_{comp} = R_1 \,//\, R_2$$

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.

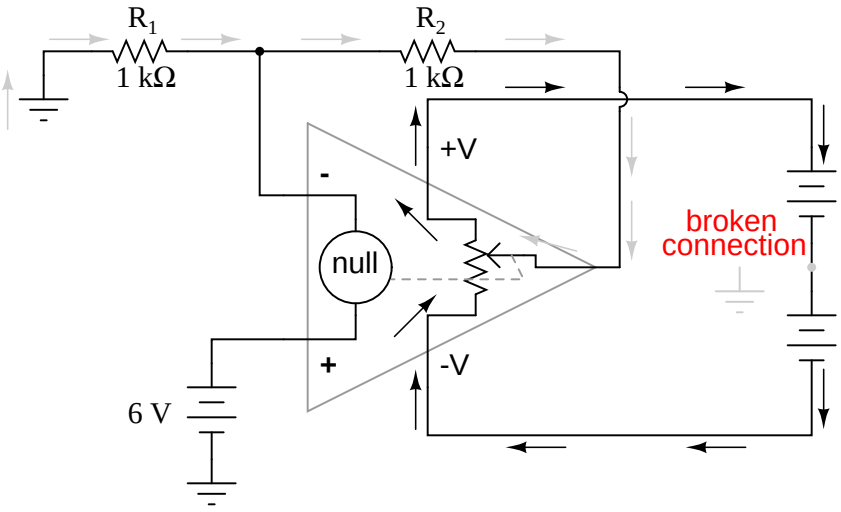
A related problem, occasionally experienced by students just learning to build operational amplifier circuits, is caused by a lack of a common ground connection to the power supply. It is *imperative* to proper op-amp function that some terminal of the DC power supply be common to the "ground" connection 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:

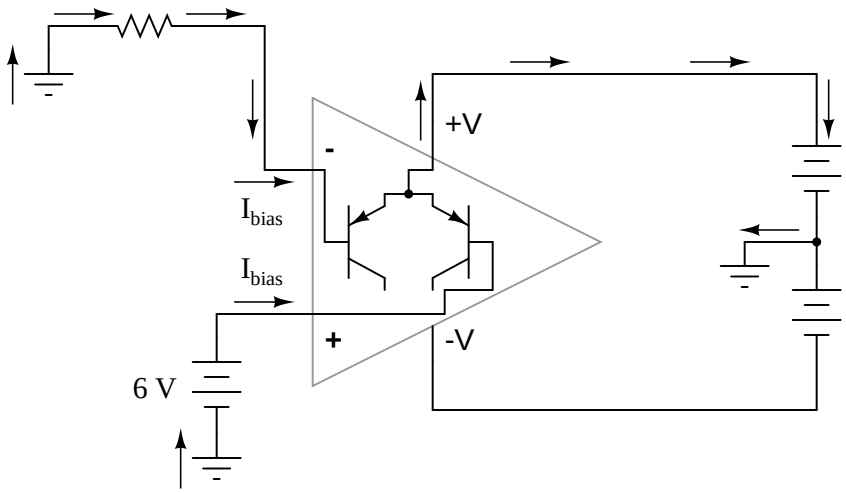
A power supply ground is essential to circuit operation!



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.

8.13.4 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.

8.13.5 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.

8.13.6 Input to output phase shift

In order to illustrate the phase shift from input to output of an operational amplifier (op-amp), the OPA227 was tested in our lab. The OPA227 was constructed in a typical non-inverting configuration (Figure 8.1).

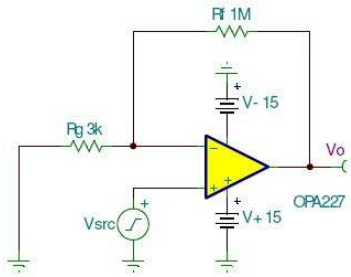


Figure 8.1: *OPA227 Non-inverting stage*

The circuit configuration calls for a signal gain of ≅34 V/V or ≅50 dB. The input excitation at Vsrc was set to 10 mVp, and three frequencies of interest: 2.2 kHz, 22 kHz, and 220 MHz. The OPA227's open loop gain and phase curve vs. frequency is shown in Figure 8.2.

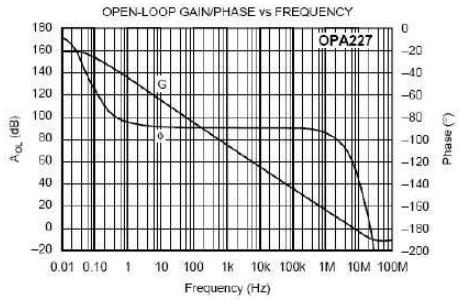


Figure 8.2: A_V *and* Φ *vs. Frequency plot*

To help predict the closed loop phase shift from input to output, we can use the open loop gain and phase curve. Since the circuit configuration calls for a closed loop gain, or $1/\beta$, of

≅50 dB, the closed loop gain curve intersects the open loop gain curve at approximately 22 kHz. After this intersection, the closed loop gain curve rolls off at the typical 20 dB/decade for voltage feedback amplifiers, and follows the open loop gain curve.

What is actually at work here is the negative feedback from the closed loop modifies the open loop response. Closing the loop with negative feedback establishes a closed loop pole at 22 kHz. Much like the dominant pole in the open loop phase curve, we will expect phase shift in the closed loop response. How much phase shift will we see?

Since the new pole is now at 22 kHz, this is also the -3 dB point as the pole starts to roll off the closed loop again at 20 dB per decade as stated earlier. As with any pole in basic control theory, phase shift starts to occur one decade in frequency before the pole, and ends at 90^o of phase shift one decade in frequency after the pole. So what does this predict for the closed loop response in our circuit?

This will predict phase shift starting at 2.2 kHz, with 45^o of phase shift at the -3 dB point of 22 kHz, and finally ending with 90^o of phase shift at 220 kHz. The three Figures shown below are oscilloscope captures at the frequencies of interest for our OPA227 circuit. Figure 8.3 is set for 2.2 kHz, and no noticeable phase shift is present. Figure 8.4 is set for 220 kHz, and ≅45^o of phase shift is recorded. Finally, Figure 8.5 is set for 220 MHz, and the expected ≅90^o of phase shift is recorded. The scope plots were captured using a LeCroy 44x Wavesurfer. The final scope plot used a x1 probe with the trigger set to HF reject.

Figure 8.3: *OPA227 Av=50dB @ 2.2 kHz*

8.14 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

Figure 8.4: *OPA227 Av=50dB @ 22 kHz*

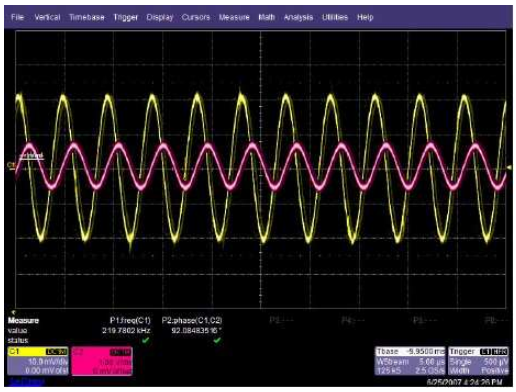
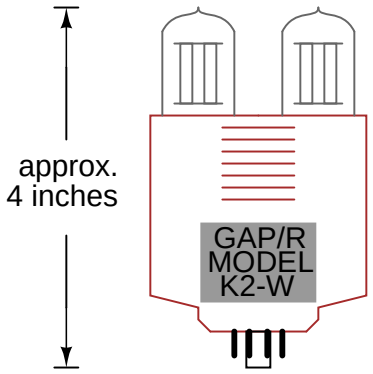


Figure 8.5: *OPA227 Av=50dB @ 220 kHz*

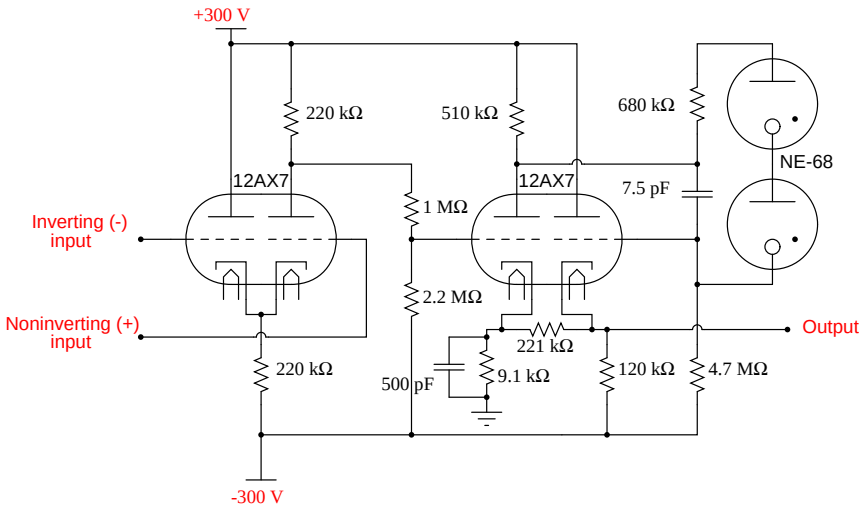
built around two twin-triode tubes mounted in an assembly with an octal (8-pin) socket for easy installation and servicing in electronic equipment chassis of that era. The assembly looked something like this:

The Philbrick Researches op-amp, model K2-W



The schematic diagram shows the two tubes, along with ten resistors and two capacitors, a fairly simple circuit design even by 1952 standards:

The Philbrick Researches op-amp, model K2-W



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 in Figure 8.6.

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 shown in Figure 8.7.

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

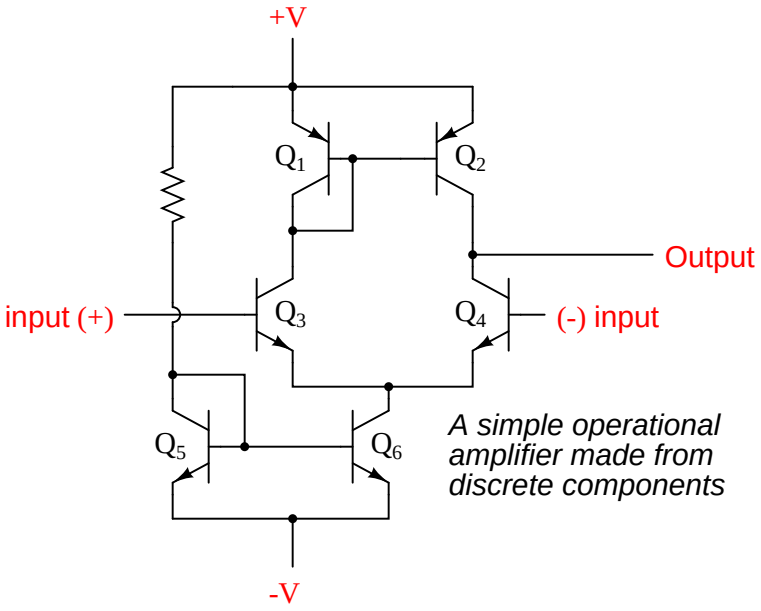


Figure 8.6: *A simple operational amplifier made from discrete components.*

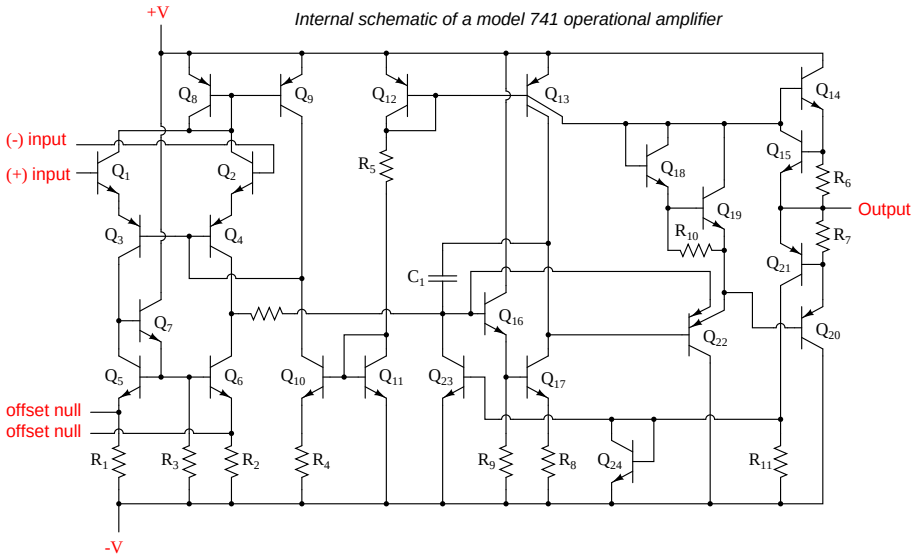


Figure 8.7: *Schematic diagram of a model 741 op-amp.*

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.

Table 8.1: *Widely used operational amplifiers*

Model number	**Devices/ package (count)**	**Power supply (V)**	**Bandwidth (MHz)**	**Bias current (nA)**	**Slew rate (V/μS)**	**Output 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
741C	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

Listed in Table 8.1 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 the op-amps, both designed for high bandwidth in Table 8.2.

Table 8.2: *High bandwidth operational amplifiers*

Model number	**Devices/ package (count)**	**Power supply (V)**	**Bandwidth (MHz)**	**Bias current (nA)**	**Slew rate (V/μS)**	**Output current (mA)**
CLC404	1	10 / 14	232	44,000	2600	70
CLC425	1	5 / 14	1900	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. Some op-amps, designed for high power output are listed in Table 8.3.

Table 8.3: *High current operational amplifiers*

Model number	**Devices/ package (count)**	**Power supply (V)**	**Bandwidth (MHz)**	**Bias current (nA)**	**Slew rate (V/μS)**	**Output current (mA)**
LM12CL	1	15 / 80	0.7	1000	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.

8.15 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).

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.

Wayne Little (June 2007): Author, "Input to output phase shift" subsection, in "Practical considerations" section.

Chapter 9

PRACTICAL ANALOG SEMICONDUCTOR CIRCUITS

Contents

9.1 ElectroStatic Discharge 413
9.1.1 ESD Damage Prevention 414
9.1.2 Storage and Transportation of ESD sensitive component and boards . . . 418
9.1.3 Conclusion 418
9.2 Power Supply circuits 419
9.2.1 Power Supply types 419
9.2.2 Power Supply Introduction 420
9.2.3 Linear power supplies 423
9.3 Amplifier circuits – PENDING 432
9.4 Oscillator circuits – INCOMPLETE 433
9.4.1 Varactor multiplier 434
9.5 Phase-locked loops – PENDING 434
9.6 Radio circuits – INCOMPLETE 434
9.7 Computational circuits 444
9.8 Measurement circuits – INCOMPLETE 465
9.9 Control circuits – PENDING 466
Bibliography 466

***** INCOMPLETE *****

9.1 ElectroStatic Discharge

Volume I chapter 1.1 discusses static electricity, and how it is created. This has a lot more significance than might be first assumed, as control of static electricity plays a large part in

modern electronics and other professions. An ElectroStatic Discharge event is when a static charge is bled off in an uncontrolled fashion, and will be referred to as ESD hereafter.

ESD comes in many forms, it can be as small as 50 volts of electricity being equalized up to tens of thousands of volts. The actual power is extremely small, so small that no danger is generally offered to someone who is in the discharge path of ESD. It usually takes several thousand volts for a person to even notice ESD in the form of a spark and the familiar zap that accompanies it. The problem with ESD is even a small discharge that can go completely unnoticed can ruin semiconductors. A static charge of thousands of volts is common, however the reason it is not a threat is there is no current of any substantial duration behind it. These extreme voltages do allow ionization of the air and allow other materials to break down, which is the root of where the damage comes from.

ESD is not a new problem. Black powder manufacturing and other pyrotechnic industries have always been dangerous if an ESD event occurs in the wrong circumstance. During the era of tubes (AKA valves) ESD was a nonexistent issue for electronics, but with the advent of semiconductors, and the increase in miniaturization, it has become much more serious.

Damage to components can, and usually do, occur when the part is in the ESD path. Many parts, such as power diodes, are very robust and can handle the discharge, but if a part has a small or thin geometry as part of their physical structure then the voltage can break down that part of the semiconductor. Currents during these events become quite high, but are in the nanosecond to microsecond time frame. Part of the component is left permanently damaged by this, which can cause two types of failure modes. Catastrophic is the easy one, leaving the part completely nonfunctional. The other can be much more serious. Latent damage may allow the problem component to work for hours, days or even months after the initial damage before catastrophic failure. Many times these parts are referred to as ”walking wounded”, since they are working but bad. Figure 9.1 is shown an example of latent (”walking wounded”) ESD damage. If these components end up in a life support role, such as medical or military use, then the consequences can be grim. For most hobbyists it is an inconvenience, but it can be an expensive one.

Even components that are considered fairly rugged can be damaged by ESD. Bipolar transistors, the earliest of the solid state amplifiers, are not immune, though less susceptible. Some of the newer high speed components can be ruined with as little as 3 volts. There are components that might not be considered at risk, such as some specialized resistors and capacitors manufactured using MOS (Metal Oxide Semiconductor) technology, that can be damaged via ESD.

9.1.1 ESD Damage Prevention

Before ESD can be prevented it is important to understand what causes it. Generally materials around the workbench can be broken up into 3 categories. These are ESD Generative, ESD Neutral, and ESD Dissipative (or ESD Conductive). ESD Generative materials are active static generators, such as most plastics, cat hair, and polyester clothing. ESD Neutral materials are generally insulative, but don’t tend to generate or hold static charges very well. Examples of this include wood, paper, and cotton. This is not to say they can not be static generators or an ESD hazard, but the risk is somewhat minimized by other factors. Wood and wood products, for example, tend to hold moisture, which can make them slightly conductive. This is true of a lot of organic materials. A highly polished table would not fall under this category, because

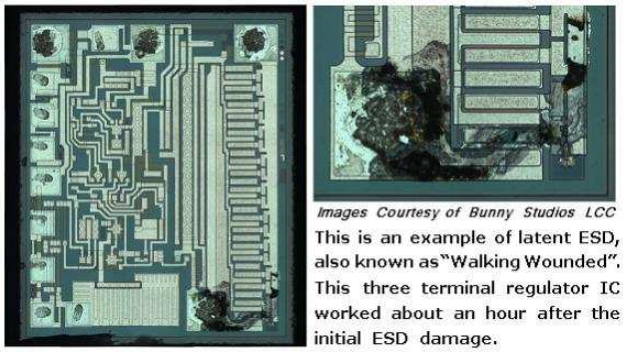


Figure 9.1:

the gloss is usually plastic, or varnish, which are highly efficient insulators. ESD Conductive materials are pretty obvious, they are the metal tools laying around. Plastic handles can be a problem, but the metal will bleed a static charge away as fast as it is generated if it is on a grounded surface. There are a lot of other materials, such as some plastics, that are designed to be conductive. They would fall under the heading of ESD Dissipative. Dirt and concrete are also conductive, and fall under the ESD Dissipative heading.

There are a lot of activities that generate static, which you need to be aware of as part of an ESD control regimen. The simple act of pulling tape off a dispenser can generate extreme voltage. Rolling around in a chair is another static generator, as is scratching. In fact, any activity that allows 2 or more surfaces to rub against each other is pretty certain to generate some static charge. This was mentioned in the beginning of this book, but real world examples can be subtle. This is why a method for continuously bleeding off this voltage is needed. Things that generate huge amounts of static should be avoided while working on components.

Plastic is usually associated with the generation of static. This has been gotten around in the form of conductive plastics. The usual way to make conductive plastic is an additive that changes the electrical characteristics of the plastic from an insulator to a conductor, although it will likely still have a resistance of millions of ohms per square inch. Plastics have been developed that can be used as conductors is in low weight applications, such as those in the airline industries. These are specialist applications, and are not generally associated with ESD control.

It is not all bad news for ESD protection. The human body is a pretty decent conductor. High humidity in the air will also allow a static charge to dissipate harmlessly away, as well as making ESD Neutral materials more conductive. This is why cold winter days, where the humidity inside a house can be quite low, can increase the number of sparks on a doorknob. Summer, or rainy days, you would have to work quite hard to generate a substantial amount of static. Industry clean rooms and factory floors go the effort to regulate both temperature and humidity for this reason. Concrete floors are also conductive, so there may be some existing components in the home that can aid in setting up protections.

To establish ESD protection there has to be a standard voltage level that everything is referenced to. Such a level exists in the form of ground. There are very good safety reasons that ground is used around the house in outlets. In some ways this relates to static, but not

directly. It does give us a place to dump our excess electrons, or acquire some if we are short, to neutralize any charges our bodies and tools might acquire. If everything on a workbench is connected directly or indirectly to ground via a conductor then static will dissipate long before an ESD event has a chance to occur.

A good grounding point can be made several different ways. In houses with modern wiring that is up to code the ground pin on the AC plug in can be used, or the screw that holds the outlets cover plate on. This is because house wiring actually has a wire or spike going into the earth somewhere where the power is tapped from the main power lines. For people whose house wiring isn't quite right a spike driven into the earth at least 3 feet or a simple electrical connection to metal plumbing (worst option) can be used. The main thing is to establish an electrical path to the earth outside the house.

Ten megohms is considered a conductor in the world of ESD control. Static electricity is voltage with no real current, and if a charge is bled off seconds after being generated it is nullified. Generally a 1 to 10 megohm resistor is used to connect any ESD protection for this reason. It has the benefit of slowing the discharge rate during an ESD event, which increases the likelihood of a component surviving undamaged. The faster the discharge, the higher the current spike going though the component. Another reason such a resistance is considered desirable is if the user is accidentally shorted to high voltage, such as household current, it won't be the ESD protections that kill them.

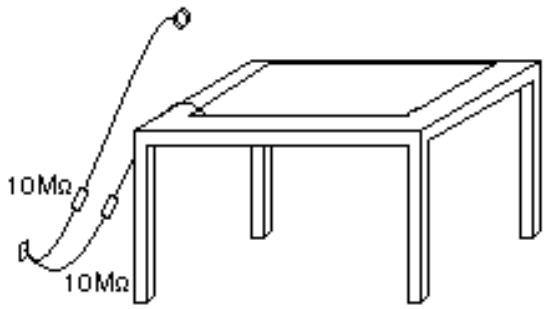
A large industry has grown up around controlling ESD in the electronics industry. The staple of any electronics construction is the workbench with a static conductive or dissipative surface. This surface can be bought commercially, or home made in the form of a sheet of metal or foil. In the case of a metal surface it might be a good idea to lay thin paper on top, although it is not necessary if you are not doing any powered tests on the surface. The commercial version is usually some form of conductive plastic whose resistance is high enough not to be a problem, which is a better solution. If you are making your own surface for the workbench be sure to add the 10 megohm resistor to ground, otherwise you have no protection at all.

The other big item that needs ESD grounded is you. People are walking static generators. Your body being conductive it is relatively easy to ground it though, this is usually done with a wrist strap. Commercial versions already have the resistor built in, and have a wide strap to offer a good contact surface with your skin. Disposable versions can be bought for a few dollars. A metal watchband is also a good ESD protection connection point. Just add a wire (with the resistor) to your grounding point. Most industries take the issue seriously enough to use real time monitors that will sound an alarm if the operator is not properly grounded.

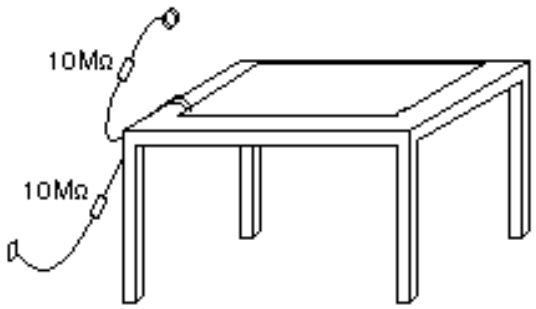
Another way of grounding yourself is a heel strap. A conductive plastic part is wrapped around the heel of your shoe, with a conductive plastic strap going up and under your sock for good contact with the skin. It only works on floors with conductive wax or concrete. The method will keep a person from generating large charges that can overwhelm other ESD protections, and is not considered adequate in and of itself. You can get the same effect by walking barefoot on a concrete floor.

Yet another ESD protection is to wear ESD conductive smocks. Like the heel strap, this is a secondary protection, not meant to replace the wrist strap. They are meant to short circuit any charges that your clothes may generate.

Moving air can also generate substantial static charges. When you blow dust off your electronics their will be static generated. An industrial solution to the problem to this issue is two fold: Firstly, air guns have a small, well shielded radioactive material implanted within the



Correctly grounded station, the table surface and wrist strap each have their own path to ground.



Incorrectly grounded station, the wrist strap is connected to the table surface, in the event of ESD in the wrist strap it will also raise the table's potential. This is better than no protection.

Figure 9.2:

air gun to ionize the air. Ionized air is a conductor, and will bleed off static charges quite well. Secondly, use high voltage electricity to ionize the air coming out of a fan, which has the same effect as the air gun. This will effectively help a workstation reduce the potential for ESD generation by a large amount.

Another ESD protection is the simplest of all, distance. Many industries have rules stating all Neutral and Generative materials will be at least 12 inches or more from any work in progress.

The user can also reduce the possibility of ESD damage by simply not removing the part out of its protective packaging until it is time to insert it into the circuit. This will reduce the likelihood of ESD exposure, and while the circuit will still be vulnerable, the component will have some minor protection from the rest of the components, as the other components will offer different discharge paths for ESD.

9.1.2 Storage and Transportation of ESD sensitive component and boards

It does no good to follow ESD protections on the workbench if the parts are being damaged while storing or carrying them. The most common method is to use a variation of a Faraday cage, an ESD bag. An ESD bag surrounds the component with a conductive shield, and usually has a non static generating insulative layer inside. In permanent Faraday cages this shield is grounded, as in the case of RFI rooms, but with portable containers this isn't practical. By putting a ESD bag on a grounded surface the same thing is accomplished. Faraday cages work by routing the electric charge around the contents and grounding them immediately. A car struck by lightning is an extreme example of a Faraday cage.

Static bags are by far the most common method of storing components and boards. They are made using extremely thin layers of metal, so thin as to be almost transparent. A bag with a hole, even small ones, or one that is not folded on top to seal the content from outside charges is ineffective.

Another method of protecting parts in storage is totes or tubes. In these cases the parts are put into conductive boxes, with a lid of the same material. This effectively forms a Faraday cage. A tube is meant for ICs and other devices with a lot of pins, and stores the parts in a molded conductive plastic tube that keeps the parts safe both mechanically and electrically.

9.1.3 Conclusion

ESD can be a minor unfelt event measuring a few volts, or a massive event presenting real dangers to operators. All ESD protections can be overwhelmed by circumstance, but this can be circumvented by awareness of what it is and how to prevent it. Many projects have been built with no ESD protections at all and worked well. Given that protecting these projects is a minor inconvenience it is better to make the effort.

Industry takes the problem very seriously, as both a potential life threatening issue and a quality issue. Someone who buys an expensive piece of electronics or high tech hardware is not going to be happy if they have to return it in 6 months. When a reputation is on the line it is easier to do the right thing.

These are some of the more common logos associated with anti-static labels. They are used to inform the user that the contents are static sensitive.

Figure 9.3:

9.2 Power Supply circuits

9.2.1 Power Supply types

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.

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 its cheap, simple, and efficient.

See Rectifier circuits in the Diodes chapter for the various configurations of the rectifiers used in unregulated power supplies. Note that those circuits are unfiltered, A low pass filter is normally added to the output of the rectifier circuit to remove some of the ripple.

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.

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.

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.

9.2.2 Power Supply Introduction

Power supply circuits are a class of circuits that are designed to convert electrical energy for some load. Every power supply consists of at least three parts:

- An input power source, which delivers power at some voltage or range of voltages V1

- A load, which requires power delivered at some voltage or range of voltages V2
- Conversion circuitry, which receives voltage V1 as an input and generates voltage V2 as an output

Some devices are simple enough that they can operate properly without any modifications to the voltage and current provided by the input source. For example, the lightbulb inside a low-cost flashlight is designed to emit light when connected in series with a few batteries, meaning the entire conversion circuit is just wires. In a similar way, household incandescent lightbulbs are designed to operate properly when connected to an AC source, operated at a well-regulated voltage and line frequency. But for the majority of electronic devices, it is impractical to operate an entire circuit at voltages commonly available. Computers, cell phones, car stereos, aircraft sensors, traffic lights, and pacemakers all have elements which require drastically different voltages than those delivered by any common power source. Well-designed power supply circuits convert almost all of the energy supplied by batteries, solar cells, AC lines and other power sources to voltage levels suitable for the operation of intricate electronic devices.

These are some of the typical considerations when designing a power supply circuit:

Efficiency

Efficiency is defined as the output power divided by the total input power. The maximum theoretical efficiency of a circuit is 100%, and this makes sense: the only place output power can come from in a power supply is the input power source. Energy that is consumed in the conversion process, and is not delivered as output power, is called power loss. All power supply circuits have some losses, even if those losses are very small. Maximizing efficiency and minimizing losses is of key importance in power supply design. Highly efficient devices can last longer on a single battery charge, cost less money to operate from a utility AC line, and generate less heat.

Heat

Power loss is dissipated away from a power supply circuit as heat. Very small semiconductor components may only be able to dissipate a few hundred milliwatts before they become too hot and fail. On the other hand, very large power supplies can convert multiple kilowatts of power, and routinely see tens of watts dissipated across only a few components. Further complicating issues, many power supplies are designed to operate in hot or cold environments, where temperatures can vary by over 100^o C. At hotter temperatures, devices must be thermally derated to avoid overheating, which significantly reduces the maximum output power available. At colder temperatures, considerable deviations in component values can be expected, and rapid changes in loading can lead to thermal shock effects, where repeated heating and cooling stresses components to failure. In most cases, cold temperature performance can be guaranteed with proper component selection; removing waste heat and preventing damage from overheating receive much greater consideration.

In order to prevent component failures, high dissipation components are usually connected to heat sinks. Sometimes the only heat sink needed is a solid connection to a copper plane in a printed circuit board. But for anything beyond a few watts, components need to be connected to a separate, thermally conductive metal block. By putting long metal fins on these blocks, the surface area can be boosted to increase convective heat transfer. A fan can also be used to increase airflow. Some designs even use water or oil traveling through the block to more effectively remove waste heat.

As a general rule, most semiconductors begin experiencing damage when the circuit's internal temperature reaches 150°C, though some devices are designed to withstand even higher temperatures. Other components such as inductors and capacitors are available in a wide range of operating temperatures and tolerances, with a premium charged for more extreme temperatures and tighter tolerances.

Size

In some devices such as cell phones or smart watches, there can be dozens or hundreds of components made to fit within only a few square centimeters. Power supply circuits in these types of devices must be small leave room for other, feature-rich components. In other devices such as aircraft electronics, the power requirements are large enough that many components must be attached to a heat sink. This can add significant weight to the overall design, which reduces fuel economy of the aircraft. Size is directly related to the amount of power being converted, and the efficiency of the conversion. The more power being converted, the larger the components must be to spread out self-heating and to withstand the high voltages used for larger power conversions. Improvements in efficiency can help to reduce supply size, since less heat sinking is required.

Cost

Unsurprisingly, cost is a critical factor. Generally, as both power and efficiency are increased, the cost of the power supply increases as well. This cost increase comes from a combination of expensive but well-optimized components, increased complexity leading to longer design and test cycles, and costs associated with regulatory compliance. As in any engineering challenge, power supply design is a tradeoff of acceptable performance and cost. Since all electronic devices require one or more power supply circuits, aggressive cost optimization is common. In high-volume manufacturing, saving even a few cents per product can reduce build costs by thousands of dollars.

Line Regulation

Line regulation is a measure of how well a power supply circuit can respond to changes in input source voltage. Many input power sources present a wide voltage range to a power supply input: battery voltages can vary by 30% or more across one charge cycle, solar cell voltages vary proportionally to incident sunlight, and AC line voltages can (on rare occasion) deviate by as much as 20% in either direction. Line regulation is defined as the output voltage at the maximum/minimum input voltage, minus the output voltage at the nominal input voltage. It can also be given as a percentage of the nominal output voltage value. An ideal power supply has perfect line regulation, $\pm$0V or $\pm$0% change. It is not uncommon for modern power supplies to see values $< \pm$5mV or $< \pm$0.1%.

Load Regulation

Load regulation is a measure of how well a power supply circuit can respond to changes in output loading. As output power increases, heating from power loss causes changes in reference parameters used by the circuit to control the output. Power supply designers use carefully designed reference circuits to minimize the effects of temperature variations, but observable effects still exist. Load regulation is defined as the output voltage at full load, minus the output voltage at no load. It can also be given as a percentage of the nominal output voltage value. An ideal power supply has perfect load regulation, 0V or 0% change. Modern power supply circuits can achieve values similar to line regulation.

Ripple Rejection

For many power supply circuits with an AC line as input, the line frequency is coupled

through the supply to the output. Some power supply circuits specify a ripple rejection, usually in dB, which is defined as the magnitude of a specific frequency on the output (commonly 100Hz or 120Hz) relative to the magnitude of that same specific frequency on the input.

Quiescent Current

Even at no load, some power is required to keep a power supply in regulation. The housekeeping current used to power the control circuitry of the supply is called the quiescent current. This value has a wide range, spanning from hundreds of milliamps all the way down to hundreds of nanoamps.

Output Impedance

An ideal voltage source has zero output impedance. Practical converters see some small output impedance, which tends to grow at higher frequencies. For a power supply to effectively regulate against loads that change in milliseconds or less, low output impedance is mandatory. Otherwise, sudden changes in load current will produce severe changes in output voltage. Nearly all converters can easily achieve output impedances of less than an ohm; $< 10\text{m}\Omega$ at DC is not uncommon.

Output Voltage Noise

Electrons flowing in resistors and transistors are susceptible to thermodynamic events, statistical fluctuations in current density, and other complex particle-scale phenomena. These tendencies manifest in all circuits, including power supply circuits, as noise on the output voltage. Although the average value of a power supply output is constant, noise can cause the output to experience millivolt excursions on a microsecond or submicrosecond scale. For lower power analog circuits that depend on tightly regulated power supply voltages such as high-resolution analog-to-digital converters or high-frequency oscillators, power supply noise can cripple performance. Because noise sources tend to be a function of frequency, noise is commonly listed as a value integrated over a frequency range (in RMS Volts), or is specified as a plot of noise spectral density comparing noise (in Volts/¡radic¿Hz) vs. frequency. Wideband (10Hz to 100kHz) integrated noise can be controlled to $< 10\mu$Vrms, and noise at high frequencies can approach $<$ 10nV/¡radic¿Hz.

Higher power designs tend to introduce noise on the order of tens or hundreds of millivolts, concentrated at specific frequencies, as a function of their construction. Though some designs exist which can tightly control even high-power supply noise, they are costly and are therefore reserved for specialized test and measurement equipment. Virtually all practical power supplies above a few watts will generate millivolts of noise on the output, and for many types of load this does not affect device performance in the slightest. It is common to use an effective but noisy power supply for insensitive loads, and as the input to a second, quieter power supply.

9.2.3 Linear power supplies

There are two major types of power supplies whose output behavior can be determined according to linear equations: shunt regulators, and series regulators. Shunt regulators (pictured in Figure 9.4 are so named because they shunt away unnecessary load current to keep the output in regulation. In a shunt regulator, high quiescent current is necessary, since the shunt must be able to redirect the full load current at no load conditions. This can lead to high power dissipation, especially for appreciably large full load currents. On the bright side, they are relatively simple, often made of entirely passive elements, and can be reduced to two-terminal devices.

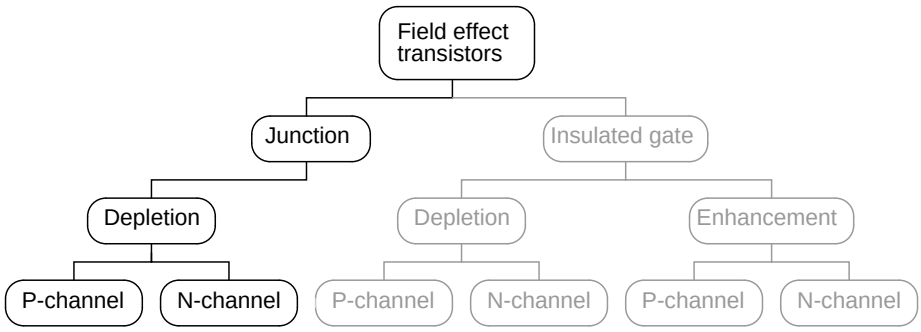


Figure 9.4: *Shunt regulators*

$$I_D = I_{DSS}\left[1 - \frac{V_{GS}}{V_P}\right]^2$$

$V_{GS} = -0.5 \qquad I_D = 200\mu A \qquad I_{DSS(min)} = 2mA$

$$V_P = 1 - \frac{V_{GS}}{\sqrt{\frac{I_D}{I_{DSS}}}} = 1 - \frac{-0.5}{\sqrt{\frac{200\mu A}{2mA}}} = -0.73V$$

Series regulators, in contrast, regulate the input current with a pass element to control the output current delivered to a load pictured in Figure 9.5. This can be utilized to reduce quiescent current to almost nothing at light or no load, though some current must always pass through the series element to ensure proper output voltage regulation. However, the series regulator must control both the voltage drop across and the current through the series element to regulate the output voltage, and no passive element can be used to guarantee this behavior. Because of the need for some form of output sensing circuitry, a three terminal solution is almost always required.

$m = \Delta Y/\Delta X$

$1/R_S = m = (I_{Qmax} - I_{Qmin})/(V_{GSmax} - V_{GSmin})$

$\mathbf{R_S = 1/m = (V_{GSmax} - V_{GSmin})/(I_{Qmax} - I_{Qmin})}$ RS

$I_D = mV_{GS} + b \qquad (Y = mX + b)$

$0 = mV_{GS} + b \qquad (I_D = 0 \text{ at } V_B ; b = V_B)$

$V_B = -b/m \qquad$ need b

$I_{Qmax} = mV_{GSmax} + b \qquad I_{Qmin} = mV_{GSmin} + b$

For the sake of illustrating the common terms seen in power supply design, consider the following specification: Suppose there is a need to take a static 15V output from a converter,

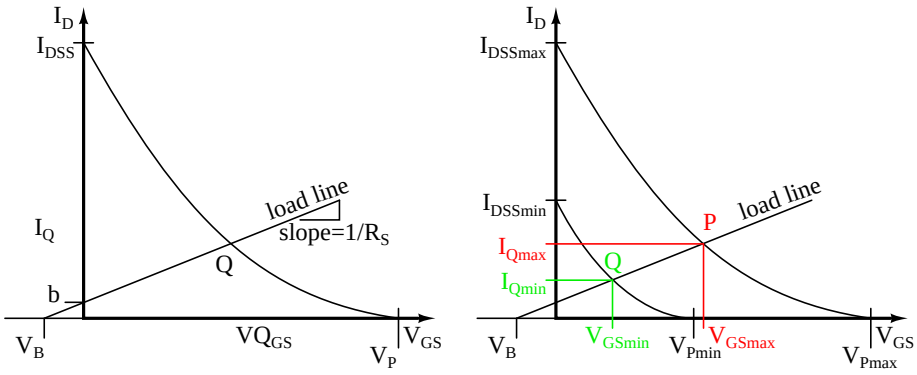
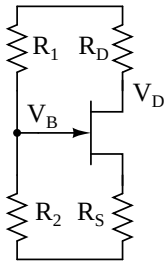


Figure 9.5: *Series regulators*

and step it down to a 5V level. The input voltage may vary by as much as ±3V, and the output current must be 500mA maximum. In the following examples, several basic topologies will be explored, and the relative strengths and weaknesses of the different approaches compared.

Example Supply: Resistor Divider



$2\text{mA} < I_{DSS} < 20\text{mA}$

$-8\text{V} < V_P < -0.73\text{V}$

$I_{Qnom} = 1\text{mA}$ $\qquad V_{DD} = 25\text{V}$

$V_D = 18\text{V} \pm 2\text{V}$

$\Delta V_D = 2\text{V}$

$R_D = (V_{DD} - V_D / I_{Qnom}) = (25\text{V} - 18\text{V})/1\text{mA} = 7\text{k} \quad (6.8\text{k})$

$\Delta I_Q = \Delta V_D / R_D = 2/6.8\text{k} = 0.29412\text{mA}$

$I_{Qmin} = I_{Qnom} - \Delta I_Q = 1 - 0.29412 = 0.70588\text{mA}$

$I_{Qmax} = I_{Qnom} + \Delta I_Q = 1 + 0.29412 = 1.29412\text{mA}$

$$V_{GSmin} = V_{Pmin}\left[1 - \sqrt{\frac{I_{Qmin}}{I_{DSSmin}}}\right] = 0.73\text{V}\left[1 - \sqrt{\frac{0.7059\text{mA}}{2\text{mA}}}\right] = 0.2963\text{V}$$

$$V_{GSmax} = V_{Pmax}\left[1 - \sqrt{\frac{I_{Qmax}}{I_{DSSmax}}}\right] = 8.0\text{V}\left[1 - \sqrt{\frac{1.2941\text{mA}}{20\text{mA}}}\right] = 5.9650\text{V}$$

$$R_S = (V_{GSmax} - V_{GSmin})/(I_{Qmax} - I_{Qmin})$$

$$R_S = (5.965 - 0.2963)/(1.29412\text{mA} - 0.70588\text{mA}) = 9.64\text{k} \quad (9.1\text{k})$$

$$V_B = \frac{V_{GSmin} I_{Qmax} - V_{GSmax} I_{Qmin}}{I_{Qmax} - I_{Qmin}} = \frac{(0.2963 \cdot 1.2941\text{mA}) - (5.9650 \cdot 0.7059\text{mA})}{1.2941\text{mA} - 0.7059\text{mA}} = -6.51$$

Figure 9.6: *Resistor divider power supply*

The humble resistor divider circuit of Figure 9.6 is perhaps the simplest power supply circuit. While its behavior is entirely linear, it is hard to say whether such a supply should be considered a shunt or series regulator, since the output voltage is a function of both the shunt and series elements. For the purposes of this first example, the distinction is unimportant. The nonidealities of this circuit, especially when constrained by the specification above, make it useful to conceptualize the common terms used in power supply design.

The resistor values must be small to simultaneously allow 500mA through the pass element without causing too much of a voltage drop, and maintain a nominal output voltage of 5V from a 15V supply. The values 15Ω and 7.5Ω are selected for R1 and R2, respectively. The behavior of the circuit can be described by a system of equations, first using ohm's law at the output, second by using the standard divider equation considering R2 and the load in parallel:

$$\frac{I_{Qmax} - b}{V_{GSmax}} = m \qquad \frac{I_{Qmin} - b}{V_{GSmin}} = m$$

$$\frac{I_{Qmax} - b}{V_{GSmax}} = \frac{I_{Qmin} - b}{V_{GSmin}}$$

$$V_{GSmin} I_{Qmax} - b V_{GSmin} = V_{GSmax} I_{Qmin} - b V_{GSmax}$$

$$V_{GSmin} I_{Qmax} - V_{GSmax} I_{Qmin} = - b(V_{GSmax} - V_{GSmin})$$

By selecting values for Vin and one other parameter, and solving for the remaining unknowns, the performance of this circuit may be interrogated. The key points are summarized below.

Efficiency: To achieve the maximum output current of 500mA with nominal input conditions (Vin = 15V), solving the system of equations gives:

$$\frac{V_{GSmin} I_{Qmax} - V_{GSmax} I_{Qmin}}{(V_{GSmax} - V_{GSmin})} = -b$$

$$V_B = \frac{-b}{m} = \left[\frac{V_{GSmax} - V_{GSmin}}{I_{Qmax} - I_{Qmin}}\right] \left[\frac{V_{GSmin} I_{Qmax} - V_{GSmax} I_{Qmin}}{(V_{GSmax} - V_{GSmin})}\right]$$

Meanwhile, the input power is the input voltage multiplied by the input current, and the input current is found as:

$$\mathbf{V_{GSmin} = V_{Pmin}\left[1 - \sqrt{\frac{I_{Qmin}}{I_{DSSmin}}}\right]} \qquad \text{VGSMIM}$$

$$\mathbf{V_{GSmax} = V_{Pmax}\left[1 - \sqrt{\frac{I_{Qmax}}{I_{DSSmax}}}\right]} \qquad \text{VGSMAX}$$

Therefore,

$$\mathbf{V_B = \frac{V_{GSmin} I_{Qmax} - V_{GSmax} I_{Qmin}}{I_{Qmax} - I_{Qmin}}} \qquad \text{VB}$$

Our efficiency at full load is then:

$$\mathbf{V_{GSmin} = V_{Pmin}\left[1 - \sqrt{\frac{I_{Qmin}}{I_{DSSmin}}}\right]} \qquad \mathbf{V_{GSmax} = V_{Pmax}\left[1 - \sqrt{\frac{I_{Qmax}}{I_{DSSmax}}}\right]}$$

It can also be shown using calculus that the maximum efficiency is achieved with a load resistance of 5Ω, yielding only 10.1% efficiency. These values are unimpressive. Interestingly, a quick calculation will reveal that this maximum efficiency is the same, regardless of input voltage. This makes sense, since output power is ratiometric with input power for the same circuit.

Quiescent current: At no load, the circuit draws:

$$R_{OUT} = \frac{1}{\frac{1}{R_L} + G_m}$$

$$R_{OUT} \approx R_L$$

This is more than the maximum output by some margin, and is very wasteful compared to what might be achievable with other topologies. Worse still, the total current only increases with increasing load.

Heat: At no load, the circuit dissipates 10W power, and at full load this increases to 12.5W. Under short circuit conditions, this increases to 15W, all dissipated in R1. Both R1 and R2 would need to be large wirewound resistors, or would require active cooling, for this supply to function at ambient temperature. Performance above ambient temperature is more difficult.

Load regulation: At full load, the output voltage drops from 5V to 2.5V. From the load regulation equation, we find that this supply has the following load regulation:

$$A_V = -g_m R_L \qquad \text{ACS}$$

This is atrocious. Any improvement in load regulation is also practically infeasible; to make the parallel combination of R2 and the load negligibly different from the load, even at full load, the value of R1 and R2 would need to be further decreased by more than an order of magnitude, which would necessarily increase the quiescent current and decrease the efficiency by the same degree. It is unreasonable to require over 100W of power dissipation to maintain a reasonable load regulation from a resistor divider. Ideally, it shouldn't even take milliwatts.

Line regulation: At 18V with no load, the output voltage is:

$$Cin = C_{gs} + C_{gd}(1-A_V)$$

or in datasheet parameters:

$$Cin = (C_{iss} - C_{rss}) + C_{rss}(1-A_V)$$

$$Cin = C_{iss} + C_{rss} A_V$$

Meanwhile, at 12V, the output voltage is:

$$R_i \approx \frac{1}{g_m}$$

This corresponds to a line regulation of $\pm$1V, or $\pm$20%. This is quite terrible.

While on the subject of input voltage variations, consider that the values of quiescent current and load regulation will change for different input voltages. As the input voltage increases, the quiescent current increases, and the heat generation increases. The output will deliver 500mA to a 7Ω load at 3.5V. At ¡15V the output voltage decreases substantially, delivering 500mA to a 3Ω load at 1.5V. Since the output voltage is directly proportional to the input voltage, the output is dependent on a stable input voltage, which is not always possible.

Output Impedance: By small signal analysis, the voltage source at Vin is shorted, and the output impedance is plainly the parallel combination of R1||R2, or 5Ω. Since this output impedance is static across all changes in input voltage and output current, it is understandable why the output voltage varies so much with every change in input and load conditions.

Output Noise: Although this resistor will be affected by thermal noise, standard 1/f noise, and excess noise due to resistor construction, ultimately noise is unlikely to be the biggest concern in this design, and true noise analysis will be saved for more deserving circuits.

The performance of this circuit as a power supply is nothing short of abysmal. In fairness to the resistor divider, the most common use for such a circuit is voltage division into high-impedance loads, such as amplifier input pins and transistor gates. For these high-impedance load conditions, the divider may be treated as very close to ideal, and as such it is not often thought of as a power supply circuit. Nevertheless, when operating conditions begin to change (such as with supply voltage variations or even small load current increases), high impedance amplifier inputs and transistors can still be made to misbehave.

In this impractical example it should be clear that a resistor divider is unsuited for any serious power delivery, with completely unusable line and load regulation and horrible overall efficiency. However, with only a minor modification, this circuit can be augmented with vastly improved line and load regulation. This is explored in the following example.

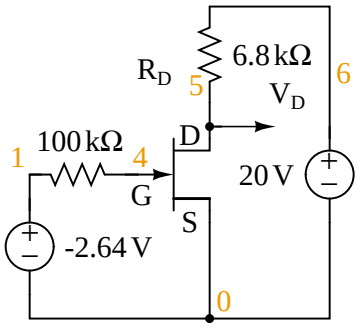
Example Supply: Zener Divider

The circuit of Figure 9.7 is a Zener divider (Zener diodes are discussed in chapter 3). By substituting a reverse-biased Zener diode in place of R2 in the previous circuit, the shifting Zener impedance above a certain reverse current knee point can be exploited to guarantee a stable output voltage over different line and load conditions. Keeping in mind that Zener diodes can only be constructed with certain reverse voltages, the closest stable output to 5V is chosen, giving a Zener voltage of 5.1V. At no load, all available current will be passed through the Zener diode. By choosing this load current to be slightly over 500mA at maximum load (say by 10mA), regulation can be ensured even when the the full load current is delivered to the load. R1 is selected for all voltages within the tolerance of the input voltage range: the worst case, at 12V, requires that:

$$\frac{1}{R_{out}} = g_m + 1/R_S$$

$$R_{out} \approx \frac{1}{g_m}$$

As long as the Zener diode has current through it, a load of 10.2Ω can now be attached, with any input supply voltage in the specified range, and 500mA will be delivered to it. To prove



$I_D = 1mA$

$V_D = 13V \pm 2V$

$R_D = (20\text{-}13)V/1mA = 7k\ (6.8k)$

$-8V < V_P < -0.73V$

$V_P = -(0.73+8)/2 \quad = -4.37V$

$2mA < I_{DSS} < 20mA$

$I_{DSS} = \sqrt{2 \cdot 20} = 6.36mA$

$$V_{GS} = V_P \left[1 - \sqrt{\frac{I_D}{I_{DSS}}}\right] = -4.37 \left[1 - \sqrt{\frac{1mA}{6.36mA}}\right] = -2.64V$$

Figure 9.7: *Zener Divder Power Supply*

this assertion, test the behavior at Vin = 12V and Vin = 18V:

$$C_{in} = C_{gd} + C_{gs}(1 - A_v) = C_{rss}A_v + C_{iss}(1 - A_v)$$

In theory, this design should therefore be capable of meeting all the requirements. A closer examination of the affected parameters offers some caveats.

Efficiency: The Zener regulator efficiency differs depending on input voltage and output loading. The best case efficiency for any input voltage is at full load, and the best case efficiency for any load is at the minimum input voltage. In this condition, we find that:

$$r_E = \frac{KT}{I^m_E}$$

where:

K = 1.38×10^{-23} watt-sec/°C, Boltzman's constant
T = temperature in Kelvins 300
I_E = emitter current
m = varies from 1 to 2 for Silicon

r_E $0.026V/I_E = 26mV/I_E$

This is better than the resistor divider, but not by much, and only at one extreme corner of operation. At the other corner the results are less impressive:

$$P_I = 18V \times 965mA = 17.2W$$

$$\frac{2.55W}{17.2W} = 14.8\%$$

Quiescent current: At no load, the full operating current of the Zener regulator must travel through the Zener diode. Best case, this is always more than the maximum output current; worst case, it can be much greater. At 17V, this regulator consumes almost double the maximum output current!

Heat: Since the Zener regulator quiescent current is always greater than the maximum operating current, the worst case power dissipation leads to a great deal of heat dissipated in both R1 and in the diode. However, as the load current increases, the Zener diode dissipates less and less power, since the current and therefore the power must be diverted from the diode to the output load. Meanwhile, R1 power dissipation remains almost constant across loading, but benefits from a lower input voltage. If for any reason the output current exceeds the quiescent current (such as during a short circuit), the power dissipation in R1 increases above the typical worst case operating point, requiring a larger component or better cooling to endure this stress. Even under normal operating conditions, R1 still dissipates enough to require a large wirewound resistor and probably some form of active cooling:

$$\frac{(18V - 5.1V)^2}{13.5\Omega} = 12.3W$$

It is worth noting, in passing, that at worst case the Zener diode must dissipate close to 5W; while there exist Zener diodes capable of this, 5W is an uncommonly large value for a Zener diode. With smaller maximum load current requirements, low power Zeners may be used at substantially decreased costs.

Line and load regulation: From an ideal standpoint, the Zener voltage is always 5.1V, across all line and load conditions. In reality, however, the Zener diode has some temperature related effects which cause the Zener voltage to change. Worse still, the temperature effects do not all act in the same direction. Low voltage Zener diodes behave predominantly according to the Zener effect, an electron tunneling process, which has a negative temperature coefficient (Zener voltage decreases with increasing heat). Higher voltage Zener diodes behave predominantly according to the avalanche effect, a form of current multiplication that has a positive temperature coefficient (Zener voltage increases with increasing heat). At around 4V to 6V, and dependent on the Zener current, the temperature coefficients of these two mechanisms will combine and can occasionally cancel out almost entirely. Unfortunately, there is still some effect at 5.1V; A 5W rated 1N5338B, for example, can see a difference of almost 0.4V across temperature, typically increasing in voltage.

A basic approximation of this effect can explain the difficulty. With a 15V supply voltage, at no load the Zener current and power are found to be:

$$I_Z = \frac{(15V - 5.1V)}{13.5\Omega} = 733mA$$

$$P_Z = 5.1V \times 733mA = 3.74W$$

Assuming the change in Zener voltage is up to 0.4V at 5W for the given Zener diode, and the change is both linear and positive, the Zener voltage may increase in response to increasing junction temperature by as much as:

$$0.4W \times \frac{3.74W}{5W} = 0.3V$$

This changes the Zener current and power to:

$$I_Z = \frac{(15V - 5.4V)}{13.5\Omega} = 711mA$$

$$P_Z = 5.4V \times 711mA = 3.84W$$

Iteration shows this change in Zener voltage with temperature eventually stabilizes; still, the output is far from its ideal value. As load current increases, Zener current decreases, returning the output voltage to a lower value. Line and load regulation are difficult to estimate precisely, since the exact location of the Zener knee, the effect of process variation on the temperature coefficient, and the variation of the temperature coefficient with Zener current cannot always be predicted. Both are frequently verified experimentally or with a spice simulation. Broadly speaking, with a maximum specified regulation swing of about 0.4V, and assuming this can be either positive or negative for a 5.1V Zener diode, the combined line and load regulation can be stated as:

$$\frac{0.4V}{5.1V} = \pm 7.8\%$$

Output Impedance: For the small signal analysis, voltage sources are shorted. The impedance looking into the output is just R1||RZ. But RZ is a dynamic value, based on the Zener voltage (Vout) and the Zener current (Iz).

$$I_Z = I_{in} - I_{out}$$

$$I_{in} = \frac{(V_{in} - V_{out})}{R_1}$$

$$I_{out} = \frac{V_{out}}{R_L}$$

$$Zout = R_1 \,||\left(\frac{V_{out}}{\frac{(V_{in} - V_{out})}{R_1} - \frac{V_{out}}{R_L}}\right) = \frac{V_{out} \times R_1 \times R_L}{(V_{in} \times R_L - V_{out} \times R_L)}$$

In the limit as RL approaches infinity, Zout becomes Vout x R1 / Vin, approximately 4.59Ω at 15V input. Interestingly, from this equation we can discover that the output impedance increases as a function of increasing load, to a maximum of R1 at a dead short across the output. The output is regulated because the output impedance continuously changes to match the level of loading.

Output Noise: The topic of output noise for Zener diodes is complicated. Due to the differ-

ent mechanisms of Zener diode behavior, there are different sources of noise for different Zener voltages and currents. Some attempt to simplify these topics will be made here.

Low voltage Zener diodes operate on the Zener effect, where discrete electrons tunnel across a barrier. Since this is a discrete, random process centered around a mean value, it follows a Poisson distribution and generates corresponding shot noise. The noise level is proportional to the square root of the number of discrete events. Thus, as current increases, shot noise increases as well. For a given Zener current I_Z and Zener voltage V_Z and recalling the electron elementary charge q = 1.6 x 10-19 coulombs, the shot noise is:

$$E_n = \sqrt{2 \times q \times I_Z} \times \frac{V_z}{I_z} = 5.66 \times 10^{-10} \times \frac{V_z}{\sqrt{I_z}} \left[nV_{rms}/\sqrt{Hz} \right]$$

At no load, this effect is almost negligible, since it is inversely proportional to I_Z. But at full load, I_Z shrinks considerably. The noise at full load for V_{in} = 12V is 7x worse than the noise at no load. At 18V, since the difference in I_Z at no load and at full load is smaller, the effect is much less pronounced.

High voltage Zener diodes operate on the avalanche effect, where one carrier collides with many others and causes an avalanche multiplication of carrier movements, resulting in wide-bandwidth noise that can exceed simple shot noise by orders of magnitude. In fact, the equation is almost identical, but depends to some extent on the recombination lifetime of each new electron in the avalanche. Without getting too deeply into the physics, it is usually sufficient to introduce some large multiplier to the original shot noise equation. Whereas a low voltage Zener diode might measure its wideband noise in the hundreds of nV, a high voltage Zener diode might measure its wideband noise in hundreds of μvolts or even low millivolts.

To keep a Zener diode at the lowest possible noise, there are only two requirements: first, use a low voltage Zener diode, to minimize avalanche noise; second, use a large Zener current, even at full load. Though increasing the current increases the power dissipation, potentially leading to greater thermal noise, remember that thermal noise is proportional to Zener impedance, and that Zener impedance shrinks faster than absolute temperature grows. In power supply design Zener noise is once again rarely an issue, since other regulators can be created with less noise, more efficiency, and better line and load regulation.

In general, shunt regulators are used in cases where the power dissipation is negligible, and the load current is small (tens of milliamps or fewer). More complex shunt regulators can incorporate compensation schemes which minimize the effects of line, load, and temperature variations. The exact mechanisms of these compensation schemes are beyond the scope of this discussion, but line and load regulation values of ¡1% are achievable with shunt regulation schemes, over a very wide range of temperatures and input voltages.

9.3 Amplifier circuits – PENDING

Note, Q_3 and Q_4 in Figure 9.8 are complementary, NPN and PNP respectively. This circuit works well for moderate power audio amplifiers. For an explanation of this circuit see “Direct coupled complementary-pair,”

(page 255).

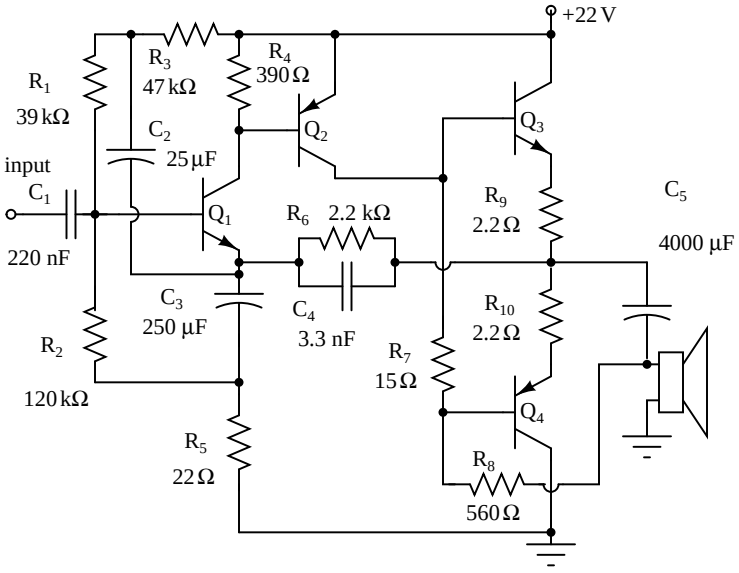


Figure 9.8: *Direct coupled complementary symmetry 3 w audio amplifier. After Mullard. [2]*

9.4 Oscillator circuits – INCOMPLETE

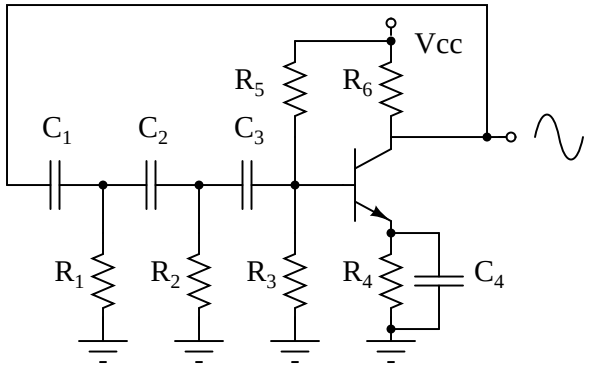


Figure 9.9: *Phase shift oscillator. R_1C_1, R_2C_2, and R_3C_3 each provide 60^o of phase shift.*

The phase shift oscillator of Figure 9.9 produces a sinewave output in the audio frequency range. Resistive feedback from the collector would be negative feedback due to 180^o phasing (base to collector phase inversion). However, the three 60^o RC phase shifters (R_1C_1, R_2C_2, and R_3C_3) provide an additional 180^o for a total of 360^o. This in-phase feedback constitutes positive feedback. Oscillations result if transistor gain exceeds feedback network losses.

9.4.1 Varactor multiplier

A Varactor or variable capacitance diode with a nonlinear capacitance vs frequency characteristic distorts the applied sinewave f1 in Figure 9.10, generating harmonics, f3.

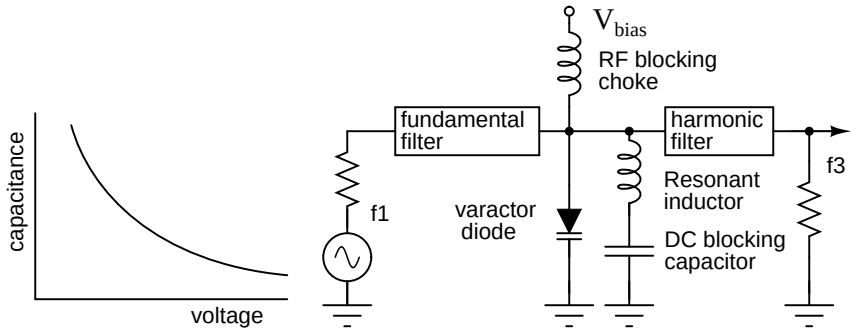


Figure 9.10: *Varactor diode, having a nonlinear capacitance vs voltage characteristic, serves in frequency multiplier.*

The fundamental filter passes f1, blocking the harmonics from returning to the generator. The choke passes DC, and blocks radio frequencies (RF) from entering the V_{bias} supply. The harmonic filter passes the desired harmonic, say the 3rd, to the output, f3. The capacitor at the bottom of the inductor is a large value, low reactance, to block DC but ground the inductor for RF. The varicap diode in parallel with the indctor constitutes a parallel resonant network. It is tuned to the desired harmonic. Note that the reverse bias, V_{bias}, is fixed.

The varicap multiplier is primarily used to generate microwave signals which cannot be directly produced by oscillators. The lumped circuit representation in Figure 9.10 is actually stripline or waveguide sections. Frequenies up to hundreds of gHz may be produced by varactor multipliers.

9.5 Phase-locked loops – PENDING

9.6 Radio circuits – INCOMPLETE

An antenna ground system, tank circuit, peak detector, and headphones are the the main components of a *crystal radio*. See Figure 9.11 (a). The antenna absorbs transimtted radio signals (b) which flow to ground via the other components. The combination of C1 and L1 comprise a resonant circuit, refered to as a *tank circuit*. Its purpose is to select one out of many available radios signals. The variable capacitor C1 allows for *tuning* to the various signals. The diode passes the positive half cycles of the RF, removing the negative half cycles (c). C2 is sized to filter the radio frequencies from the RF envelope (c), passing audio frequencies (d) to the headset. Note that no power supply is required for a crystal radio. A germanium diode, which has a lower forward voltage drop provides greater sensitvity than a silicon diode.

While 2000Ω magnetic headphones are shown above, a ceramic earphone, sometimes called

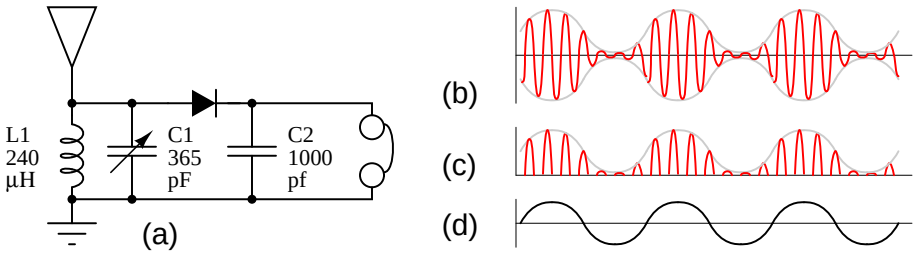


Figure 9.11: *(a) Crystal radio. (b) Modulated RF at antenna. (c) Rectified RF at diode cathode, without C2 filter capacitor. (d) Demodualted audio to headphones.*

a crystal earphone, is more sensitive. The ceramic earphone is desirable for all but the strongest radio signals

The circuit in Figure 9.12 produces a stronger output than the crystal detector. Since the transistor is not biased in the linear region (no base bias resistor), it only conducts for positive half cycles of RF input, detecting the audio modulation. An advantage of a transistor detector is amplification in addition to detection. This more powerful circuit can readily drive 2000Ω magnetic headphones. Note the transistor is a germanuim PNP device. This is probably more sensitive, due to the lower 0.2V V_{BE}, compared with silicon. However, a silicon device should still work. Reverse battery polarity for NPN silicon devices.

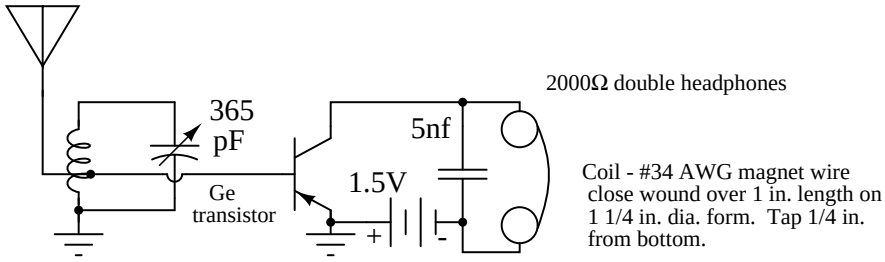


Figure 9.12: *TR One, one transistor radio. No-bias-resistor causes operation as a detector. After Stoner, Figure 4.4A. [8]*

The 2000Ω headphones are no longer a widely available item. However, the low impedance earbuds commonly used with portable audio equipment may be substituted when paired with a suitable audio transformer. See Volume 6 Experiments, AC Circuits, Sensitive audio detector for details.

The circuit in Figure 9.13 adds an audio amplifier to the crystal detector for greater headphone volume. The original circuit used a germanium diode and transistor. [8] A schottky diode may be substituted for the germanium diode. A silicon transistor may be used if the base-bias resistor is changed according to the table.

For more crystal radio circuits, simple one-transistor radios, and more advanced low transistor count radios, see Wenzel [9]

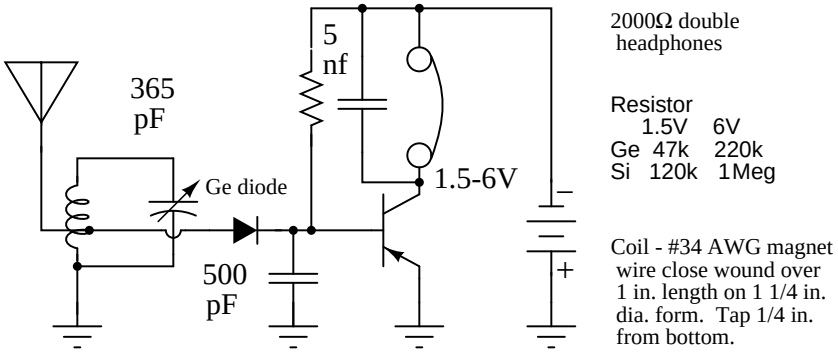


Figure 9.13: *Crystal radio with one transistor audio amplifer, base-bias. After Stoner, Figure 4.3A. [8]*

The circuit in Figure 9.15 is an integrated circuit AM radio containing all the active radio frequency circuitry within a single IC. All capacitors and inductors, along with a few resistors, are external to the IC. The 370 Pf variable capacitor tunes the desired RF signal. The 320 pF variable capacitor tunes the local oscillator 455 KHz above the RF input signal. The RF signal and local oscillator frequencies mix producing the sun and difference of the two at pin 15. The external 455 KHz ceramic filter between pins 15 and 12, selects the 455 KHz difference frequency. Most of the amplification is in the intermediate frequency (IF) amplifier between pins 12 and 7. A diode at pin 7 recovers audio from the IF. Some automatic gain control (AGC) is recovered and filtered to DC and fed back into pin 9.

Figure 9.16 shows conventional mecahnical tuning (a) of the RF input tuner and the local oscillator with varactor diode tuning (b). The meshed plates of a dual variable capacitor make for a bulky component. It is ecconomic to replace it with varicap tuning diodes. Increasing the reverse bias V_{tune} decreases capacitance which increases frequency. V_{tune} could be produced by a potentiometer.

Figure 9.17 shows an even lower parts count AM radio. Sony engineers have included the intermediate frequency (IF) bandpass filter within the 8-pin IC. This eliminates external IF transformers and an IF ceramic filter. L-C tuning components are still required for the radio frequency (RF) input and the local oscillator. Though, the variable capacitors could be replaced by varicap tuning diodes.

Figure 9.18 shows a low-parts-count FM radio based on a TDA7021T integrated circuit by NXP Wireless. The bulky external IF filter transformers have been replaced by R-C filters. The resistors are integrated, the capacitors external. This circuit has been simplified from Figure 5 in the NXP Datasheet. See Figure 5 or 8 of the datasheet for the omitted signal strength circuit. The simple tuning circuit is from the Figure 5 Test Circuit. Figure 8 has a more elaborate tuner. Datasheet Figure 8 shows a stereo FM radio with an audio amplifier for driving a speaker. [7]

For a construction project, the simplified FM Radio in Figure 9.18 is recommended. For the 56nH inductor, wind 8 turns of #22 AWG bare wire or magnet wire on a 0.125 inch drill bit or other mandrel. Remove the mandrel and strech to 0.6 inch length. The tuning capacitor may be a miniature trimmer capacitor.

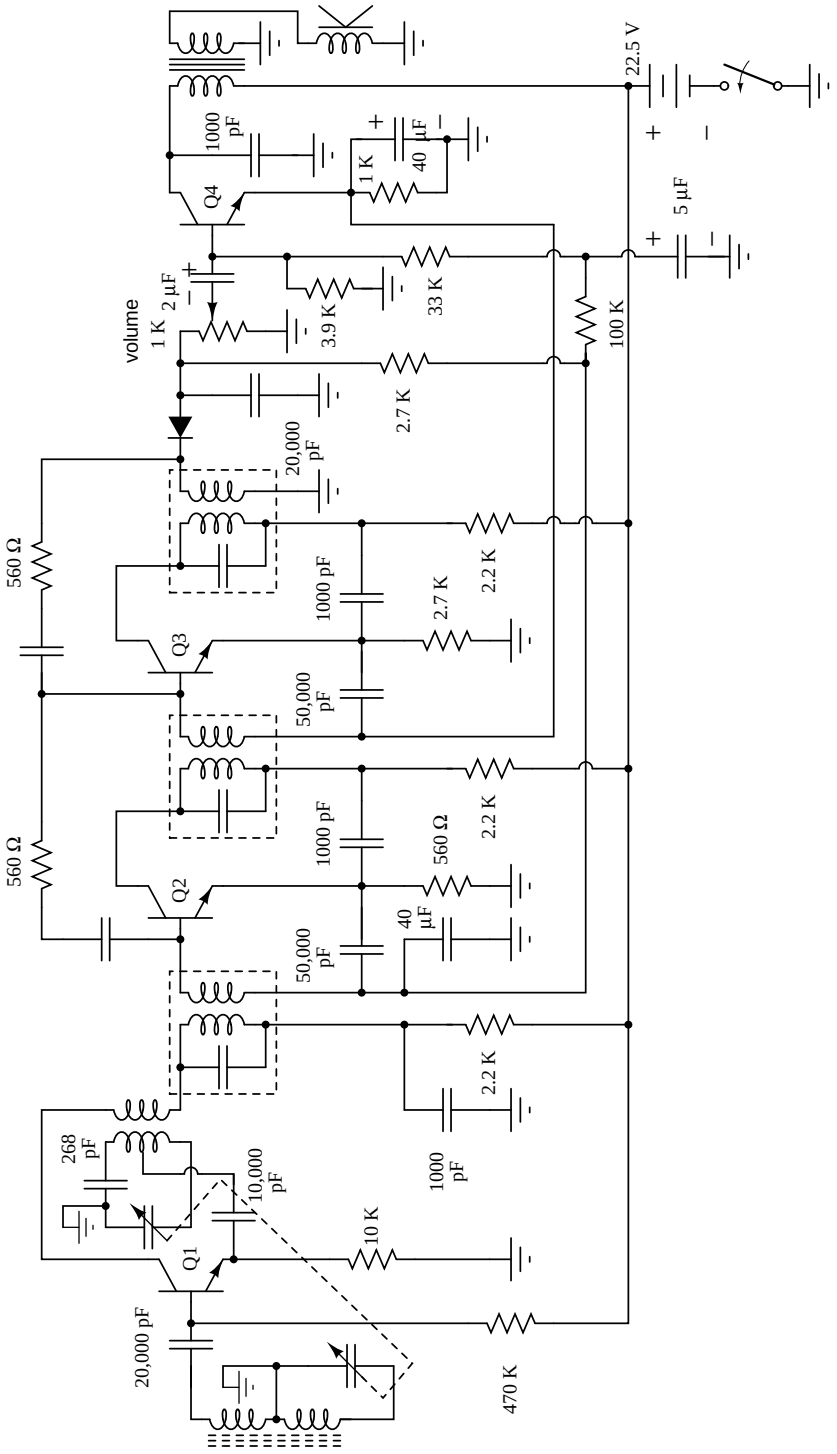


Figure 9.14: *Regency TR1: First mass produced transistor radio, 1954.*

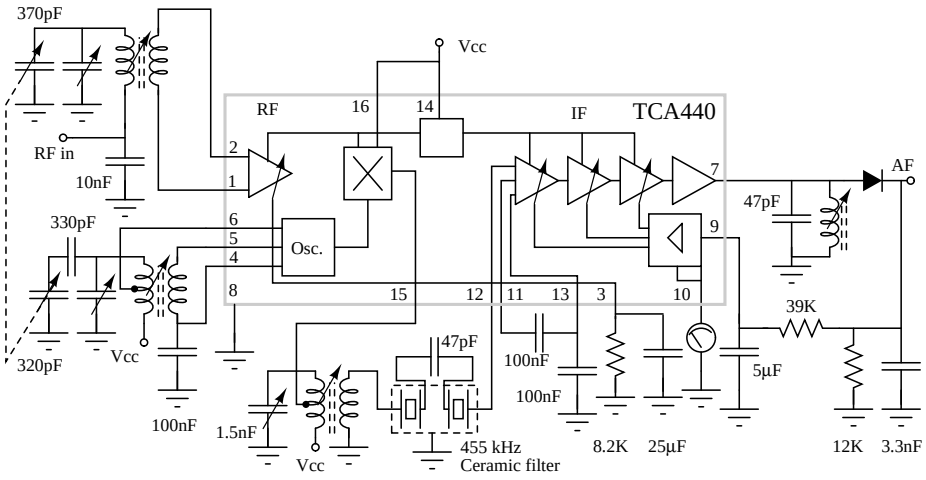


Figure 9.15: *IC radio, After Signetics [3]*

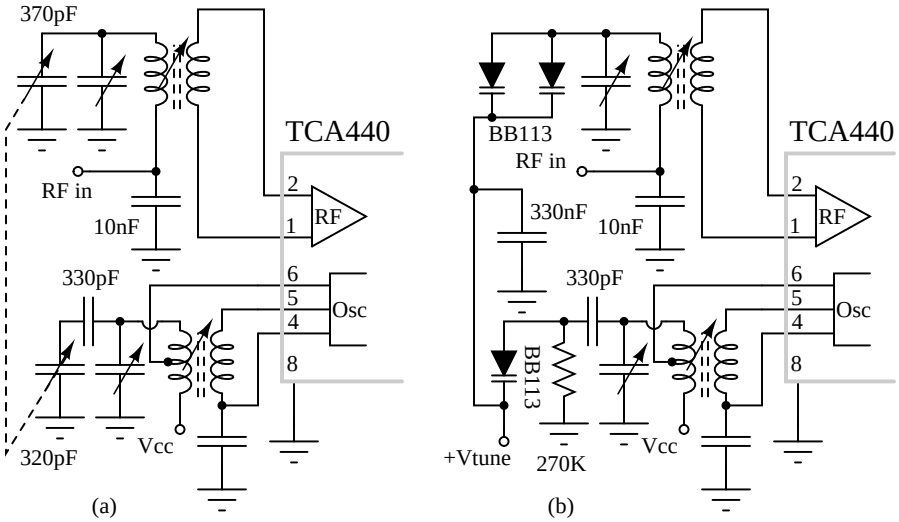


Figure 9.16: *IC radio comparison of (a) mechanical tuning to (b) electronic varicap diode tuning.[3]*

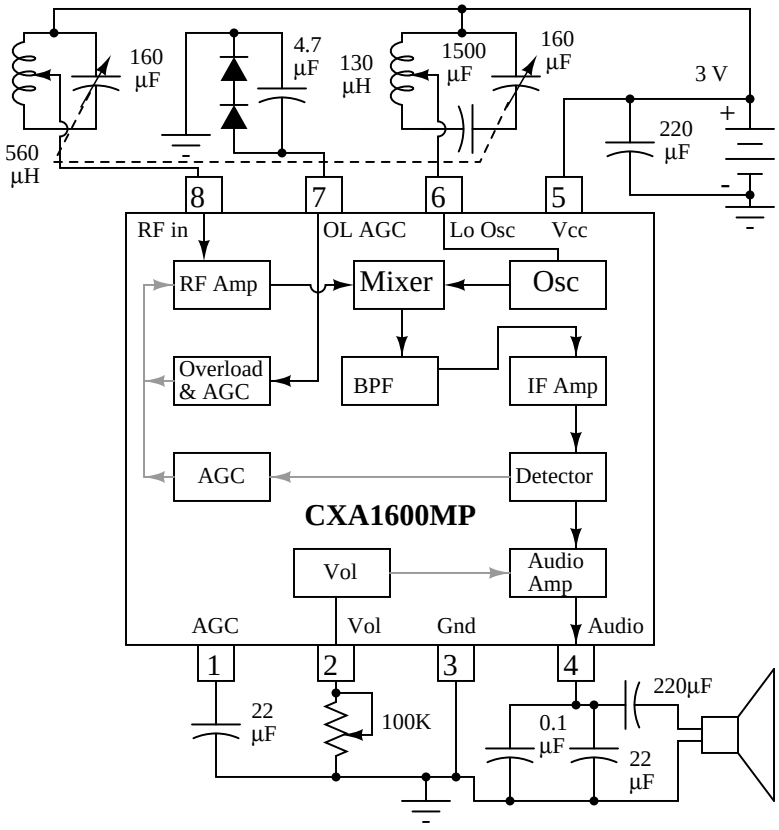


Figure 9.17: *Compact IC radio eliminates external IF filters. After Sony [4]*

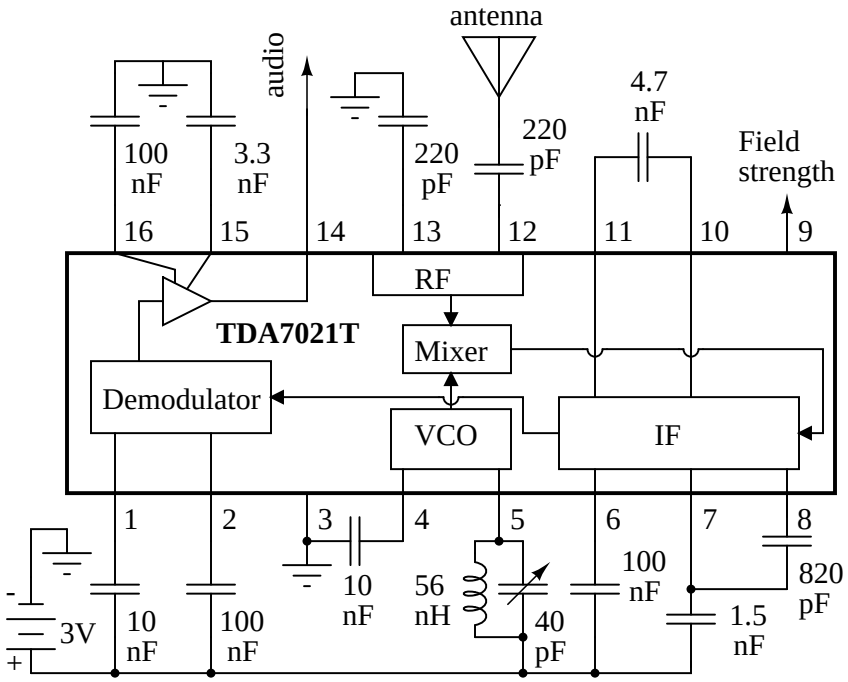


Figure 9.18: *IC FM radio, signal strength circuit not shown. After NXP Wireless Figure 5. [7]*

Figure 9.19 is an example of a common-base (CB) RF amplifier. It is a good illustration because it looks like a CB for lack of a bias network. Since there is no bias, this is a class C amplifier. The transistor conducts for less than 180^o of the input signal because at least 0.7 V bias would be required for 180^o class B. The common-base configuration has higher power gain at high RF frequencies than common-emitter. This is a power amplifier (3/4 W) as opposed to a small signal amplifier. The input and output π-networks match the emitter and collector to the 50 Ω input and output coaxial terminations, respectively. The output π-network also helps filter harmonics generated by the class C amplifier. Though, more sections would likely be required by modern radiated emissions standards.

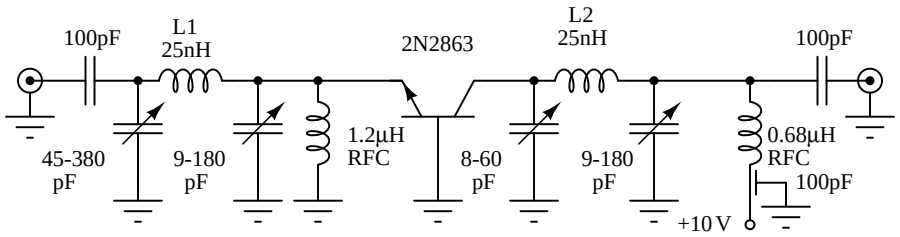


Figure 9.19: *Class C common-base 750 mW RF power amplifier. L1 = #10 Cu wire 1/2 turn, 5/8 in. ID by 3/4 in. high. L2 = #14 tinned Cu wire 1 1/2 turns, 1/2 in. ID by 1/3 in. spacing. After Texas Instruments [5]*

An example of a high gain common-base RF amplifier is shown in Figure 9.20. The common-base circuit can be pushed to a higher frequency than other configurations. This is a common base configuration because the transistor bases are grounded for AC by 1000 pF capacitors. The capacitors are necessary (unlike the class C, Figure 9.19) to allow the 1KΩ-4KΩ voltage divider to bias the transistor base for class A operation. The 500Ω resistors are emitter bias resistors. They stablize the collector current. The 850Ω resistors are collector DC loads. The three stage amplifier provides an overall gain of 38 dB at 100 MHz with a 9 MHz bandwidth.

A cascode amplifier has a wide bandwdth like a common-base amplifier and a moderately high input impedance like a common emitter arrangement. The biasing for this cascode amplifier (Figure 9.21) is worked out in an example problem (page 246).

This circuit (Figure 9.21) is simulated in the “Cascode” section of the BJT chapter (page 219). Use RF or microwave transistors for best high frequency response.

The PIN diodes are arranged in a π-attenuator network. The anti-series diodes cancel some harmonic distortion compared with a single series diode. The fixed 1.25 V supply forward biases the parallel diodes, which not only conducting DC current from ground via the resistors, but also, conduct RF to ground through the diodes' capacitors. The control voltage $V_{control}$, increases current through the parallel diodes as it increases. This decreases the resistance and attenuation, passing more RF from input to output. Attenuation is about 3 dB at $V_{control}$= 5 V. Attenuation is 40 dB at $V_{control}$= 1 V with flat frequency response to 2 gHz. At $V_{control}$= 0.5 V, attenuation is 80 dB at 10 MHz. However, the frequency response varies too much to use. [1]

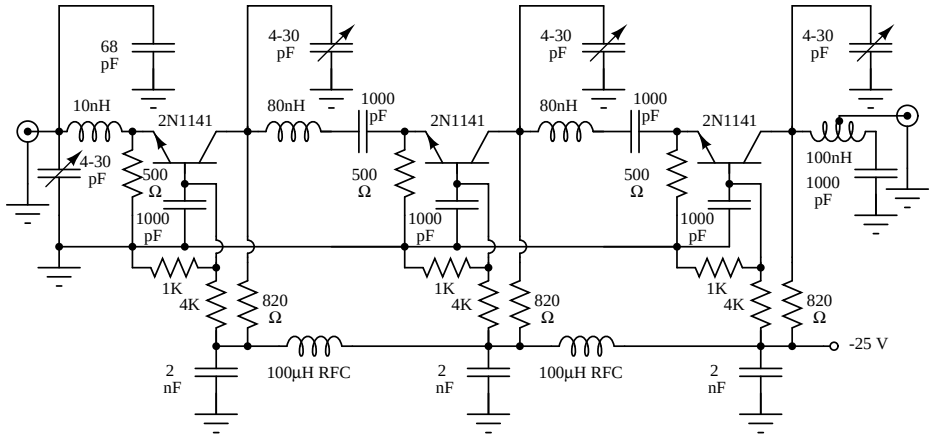


Figure 9.20: *Class A common-base small-signal high gain amplifier. After Texas Instruments [6]*

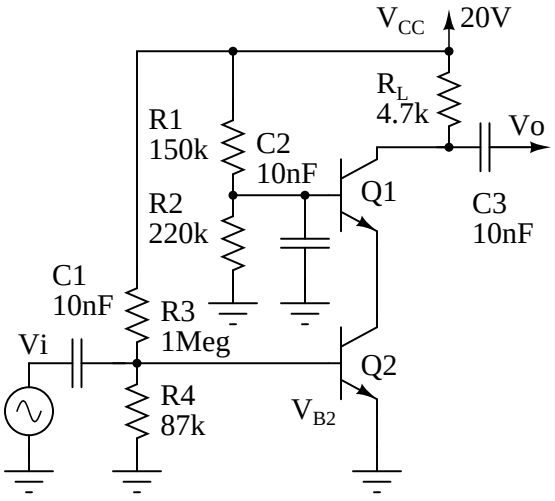


Figure 9.21: *Class A cascode small-signal high gain amplifier.*

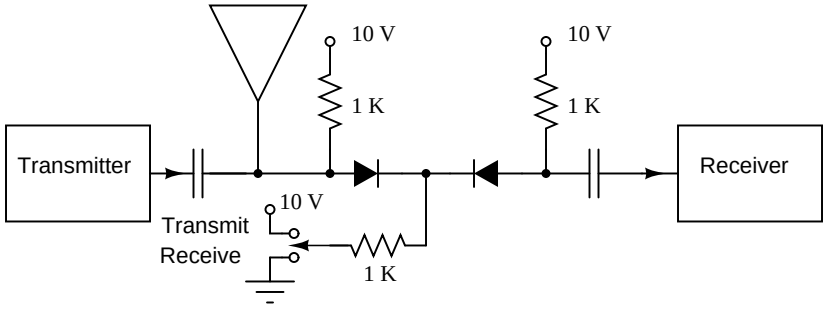


Figure 9.22: *PIN diode T/R switch disconnects receiver from antenna during transmit.*

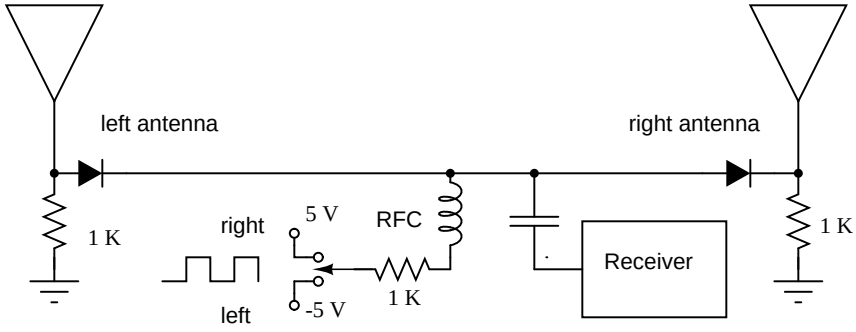


Figure 9.23: *PIN diode antenna switch for direction finder receiver.*

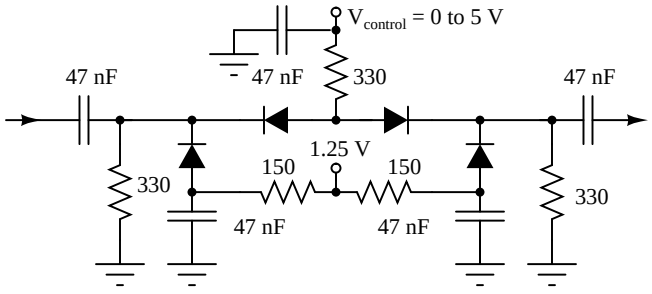


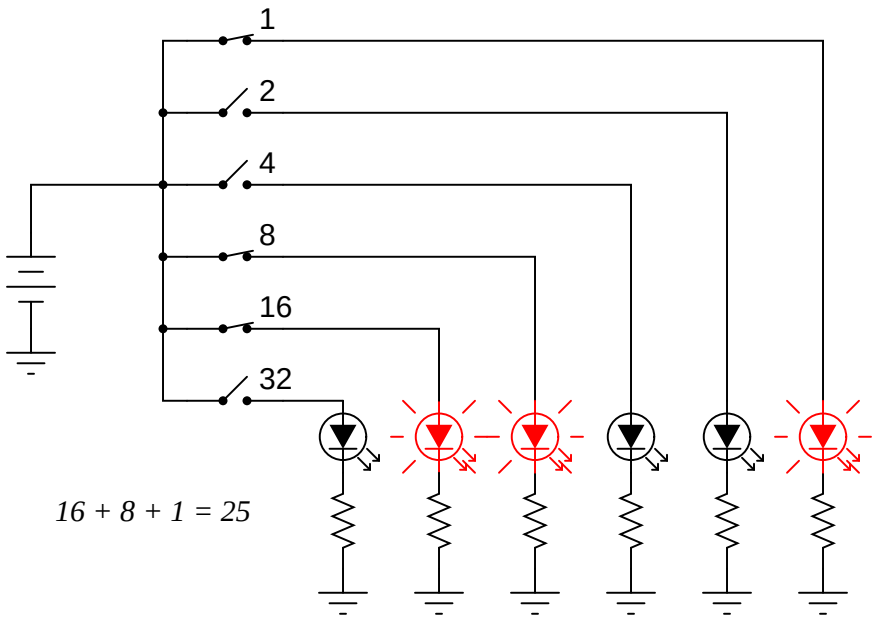
Figure 9.24: *PIN diode attenuator: PIN diodes function as voltage variable resistors. After Lin [1].*

9.7 Computational circuits

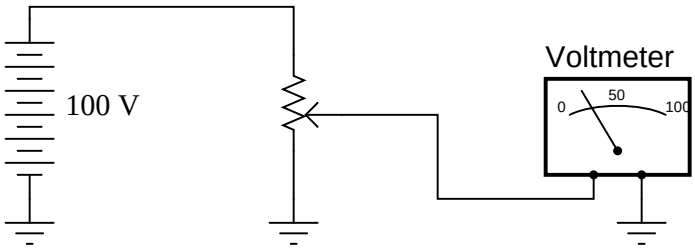
When someone mentions the word ”computer,” a digital device is what usually comes to mind. Digital circuits represent numerical quantities in *binary* format: patterns of 1’s and 0’s represented by a multitude of transistor circuits operating in saturated or cutoff states. However, analog circuitry may 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:



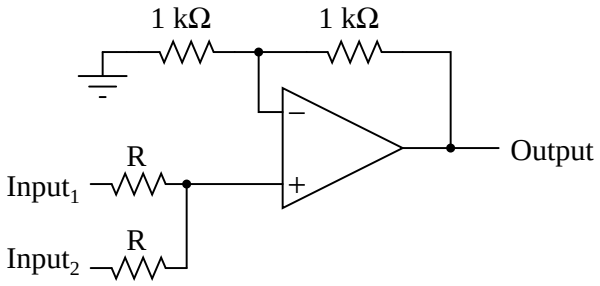
An analog circuit representing the number 25:



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.

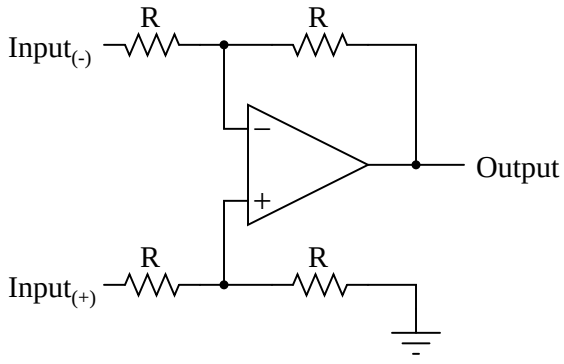
In applications where precision is not critical, analog computational circuits are very practical and elegant. Shown here are a few op-amp circuits for performing analog computation:

Analog summer (adder) circuit



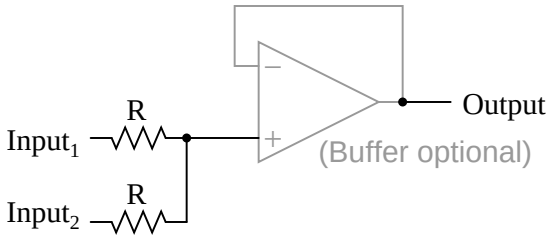
$$\text{Output} = \text{Input}_1 + \text{Input}_2$$

Analog subtractor circuit



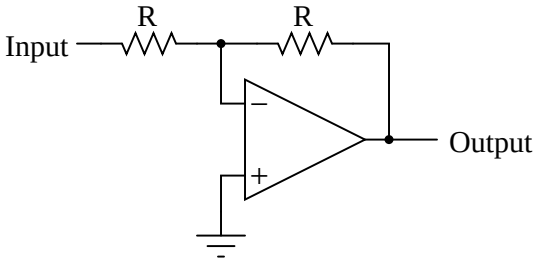
$$\text{Output} = \text{Input}_{(+)} - \text{Input}_{(-)}$$

Analog averager circuit



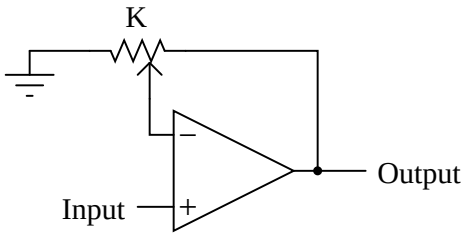
$$\text{Output} = \frac{\text{Input}_1 + \text{Input}_2}{2}$$

Analog inverter (sign reverser) circuit



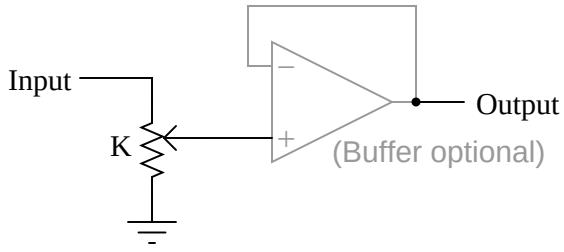
$$\text{Output} = -\text{Input}$$

Analog "multiply-by-constant" circuit



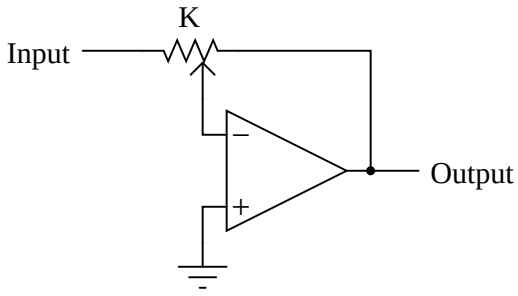
$$\text{Output} = (K)(\text{Input})$$

Analog "divide-by-constant" circuit



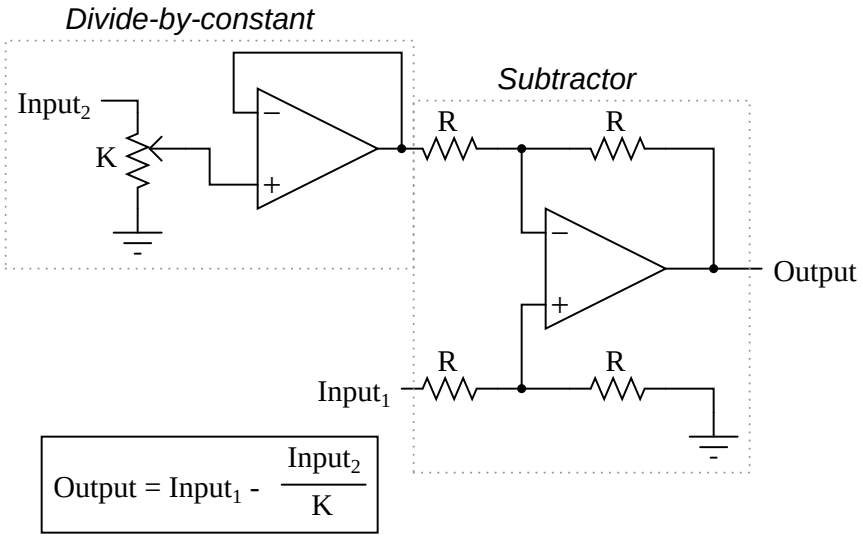
$$\text{Output} = \frac{\text{Input}}{K}$$

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*:

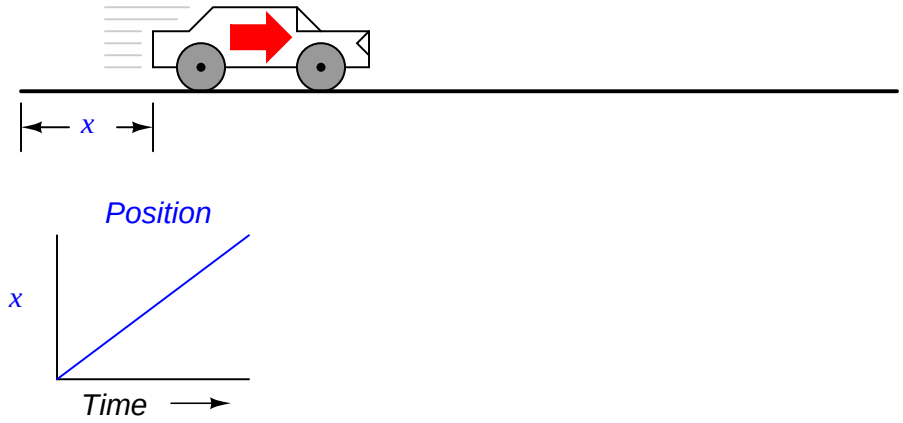
> *"A note of encouragement is offered to certain readers: integral calculus is one of the mathematical disciplines that operational [amplifier] circuitry exploits and, in the process, rather demolishes as a barrier to understanding."* (pg. 4)

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 not a Computer Trademark!":

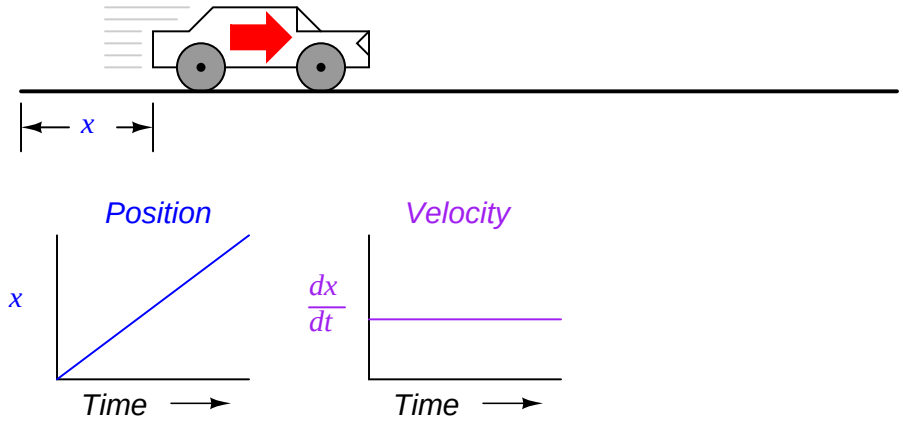
> *"Creating a real physical entity (a circuit) governed by a particular set of equations and interacting with it provides unique insight into those mathematical statements. There is no better way to develop a "gut feel" for the interplay between physics and mathematics than to experience such an interaction. 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):



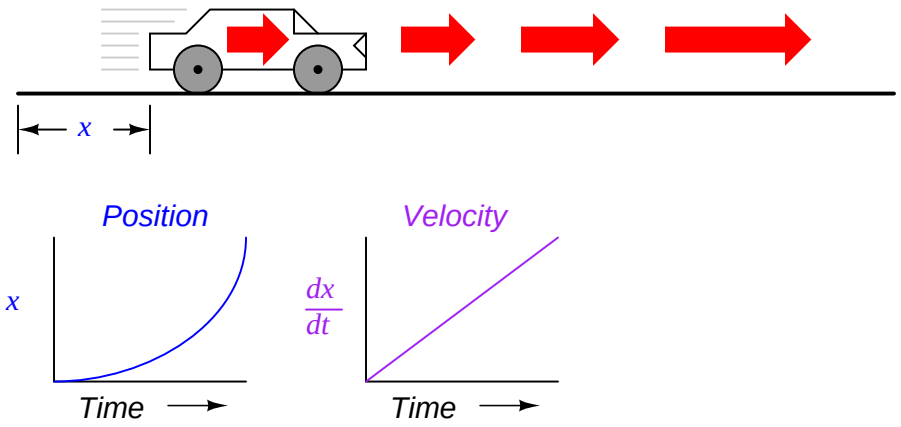
If we were to calculate the *derivative* of the car's position with respect to time (that is, determine the rate-of-change of the car's position with respect to time), we would arrive at a quantity representing the car's velocity. The differentiation function is represented by the fractional notation d/d, so when differentiating position (x) with respect to time (t), we denote the result (the derivative) as dx/dt:



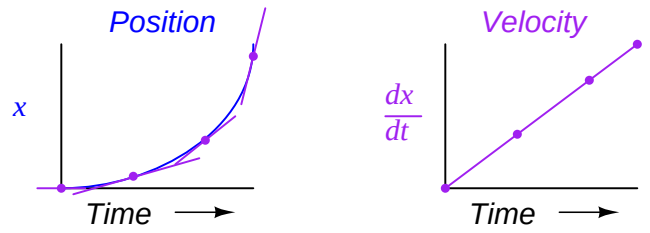
For a linear graph of x over time, the derivate of position (dx/dt), otherwise and more commonly known as *velocity*, will be a flat line, unchanging in value. The derivative of a mathematical function may be graphically understood as its *slope* when plotted on a graph,

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.

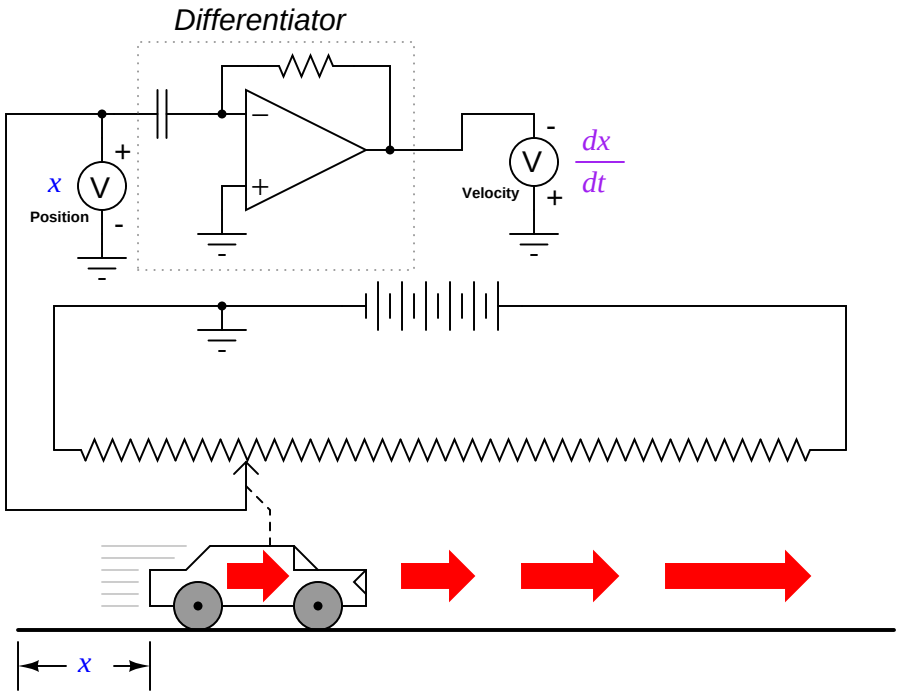
Now, suppose the distance traveled by the car increased exponentially over time: that is, it began its travel in slow movements, but covered more additional distance with each passing period in time. We would then see that the derivative of position (dx/dt), otherwise known as velocity (v), would not 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:



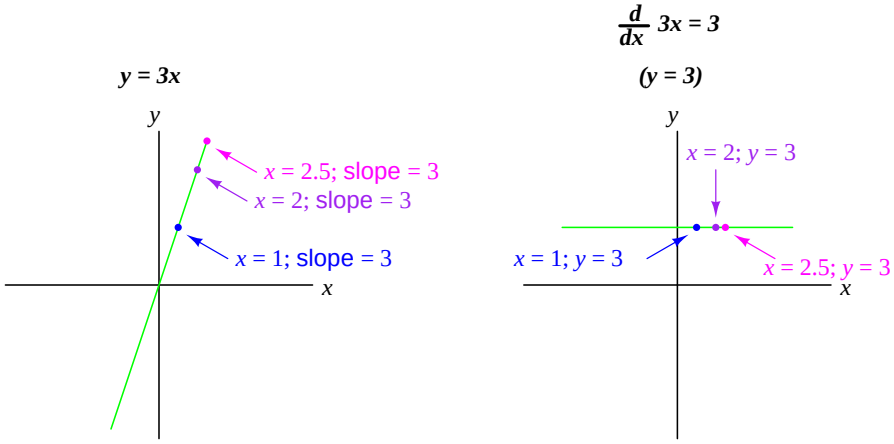
What does this have to do with analog electronic circuits? Well, if we were to have an analog voltage signal represent the car's position (think of a huge potentiometer whose wiper was attached to the car, generating a voltage proportional to the car's position), we could connect a differentiator circuit 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:



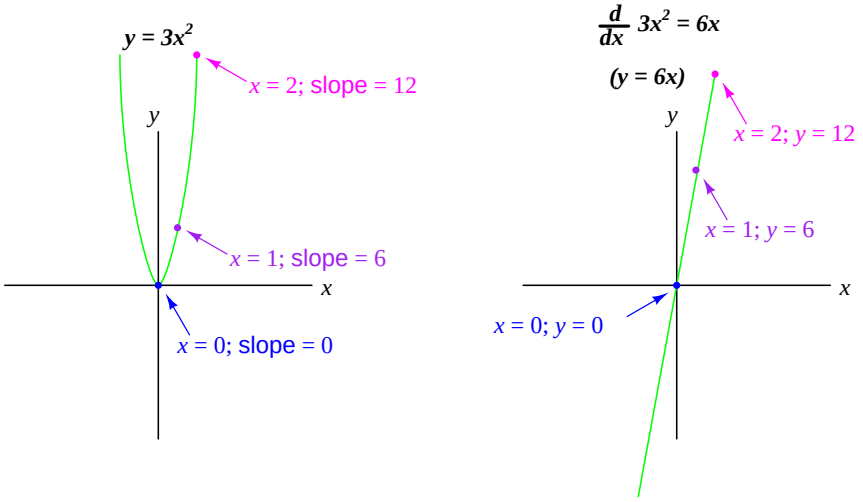
Recall from the last chapter that a differentiator circuit outputs a voltage proportional to the input voltage's *rate-of-change over time* (d/dt). Thus, if the input voltage is changing over time at a constant *rate*, the output voltage will be at a constant value. If the car moves in such a way that 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:



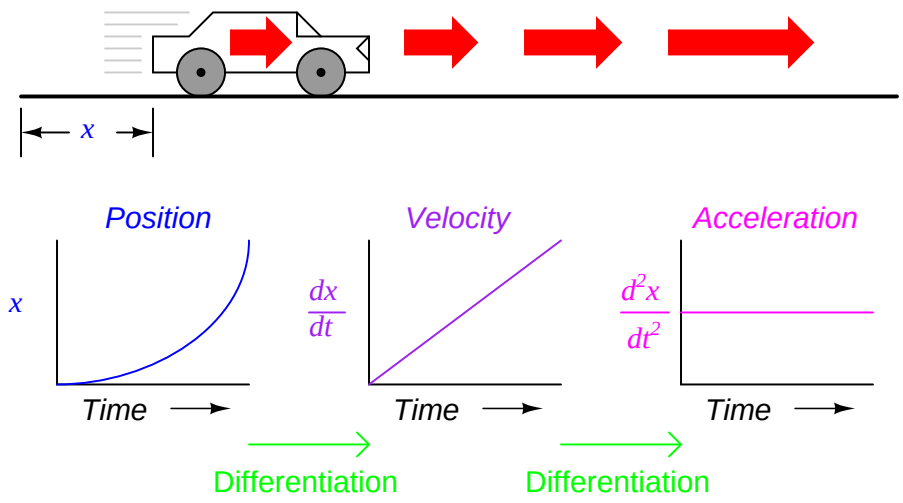
Nonlinear functions such as $y = 3x^2$ may also be differentiated by symbolic means. In this case, the derivative of $y = 3x^2$ with respect to x is $6x$:



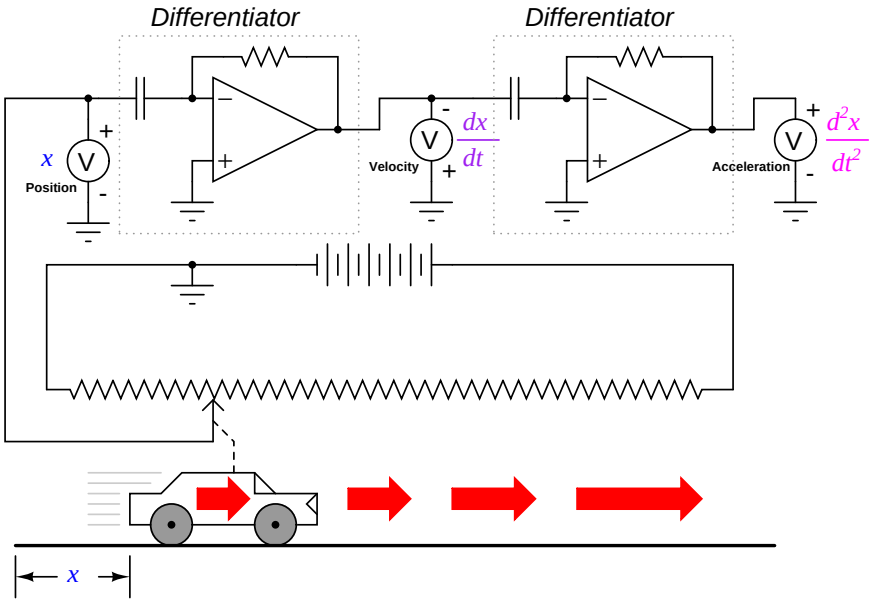
In real life, though, we often cannot describe the behavior of any physical event by a simple equation like $y = 3x$, and so symbolic differentiation of the type learned by calculus students may be impossible to apply to a physical measurement. If someone wished to determine the derivative of our hypothetical 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 them-

selves for inspection and analysis.

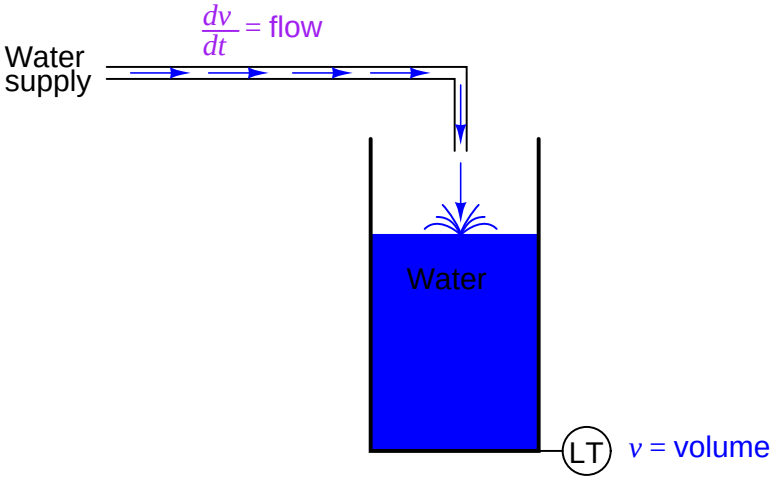
We may take the principle of differentiation one step further by applying it to the velocity signal using another differentiator circuit. In other words, use it to calculate the rate-of-change of velocity, which we know is the rate-of-change of position. What practical measure would we arrive at if we 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:



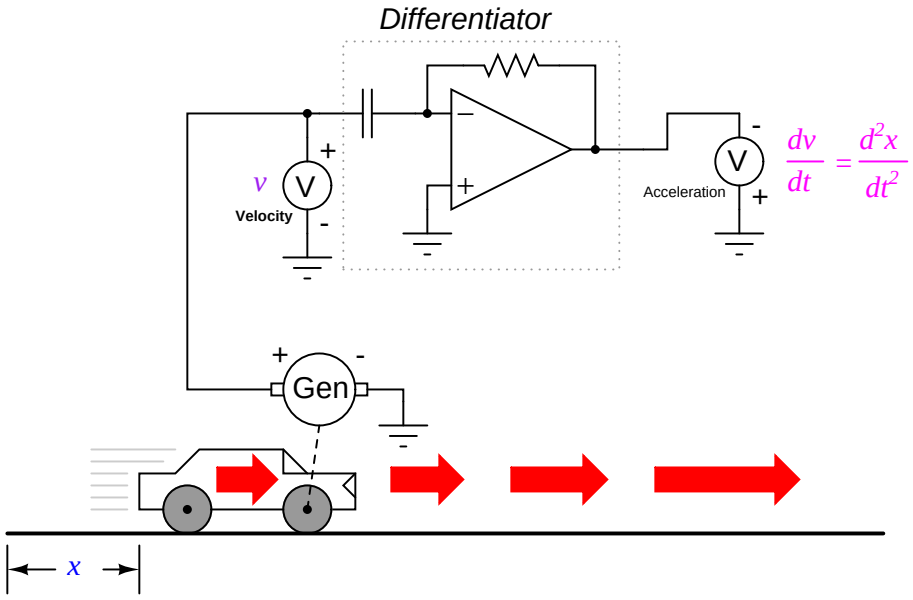
Deriving velocity from position, and acceleration from velocity, we see the principle of differentiation very clearly illustrated. These are not the only physical measurements related to each other in this way, but they are, perhaps, the most common. Another example of calculus in action is the relationship between liquid flow (q) and liquid volume (v) accumulated in a vessel over time:



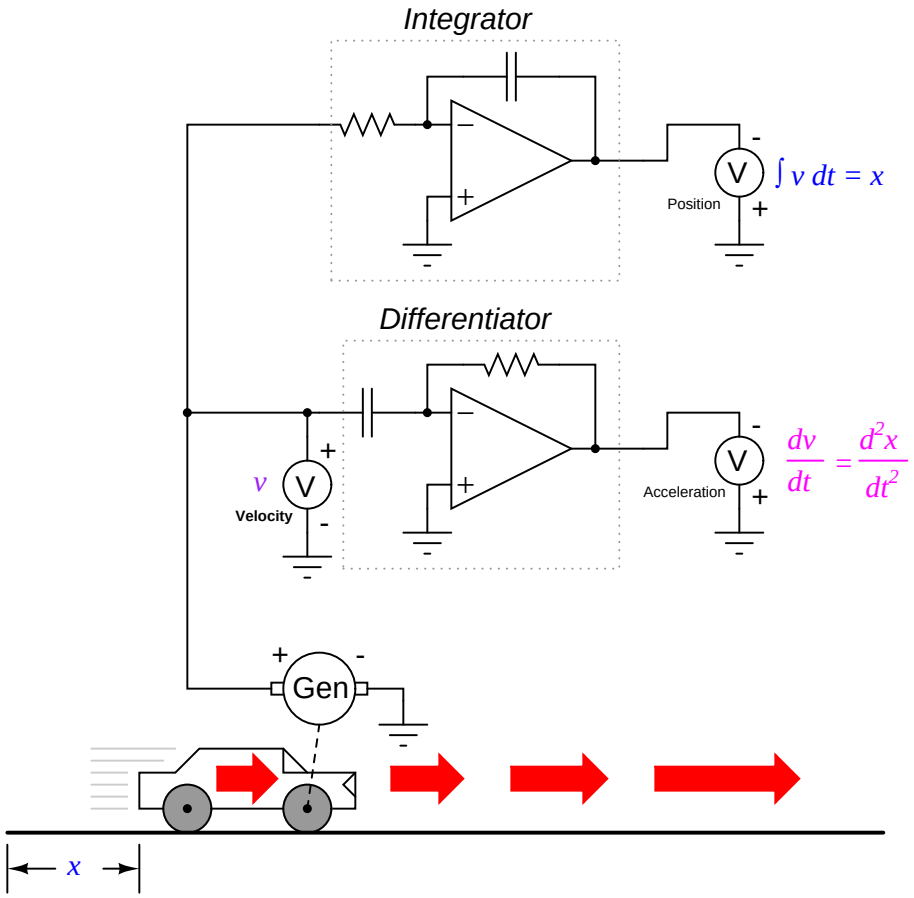
A "Level Transmitter" device mounted on a water storage tank provides a signal directly proportional to water level in the tank, which – if the tank is of constant cross-sectional area throughout its height – directly equates water volume stored. If we were to take this volume signal and differentiate 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:



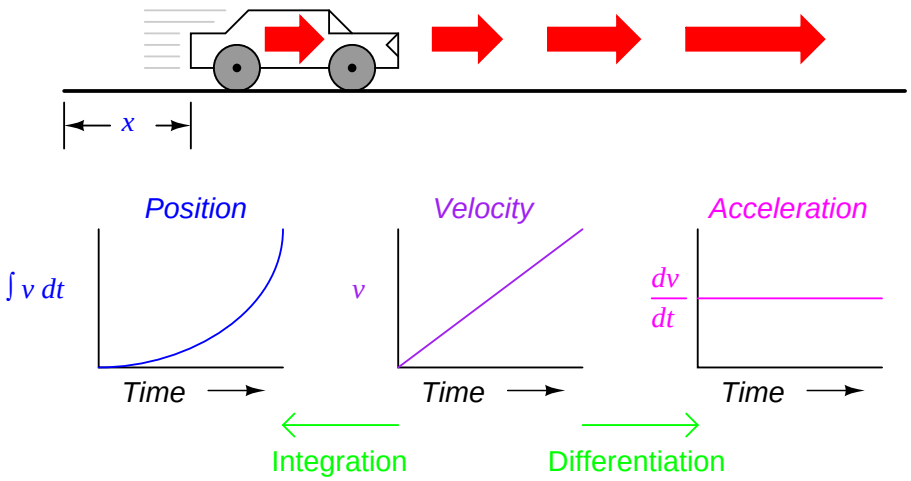
By its very nature, the tachogenerator differentiates the car's position with respect to time, generating a voltage proportional to how rapidly the wheel's angular position changes over time. This provides us with a raw signal already representative of velocity, with only a single step of differentiation 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:



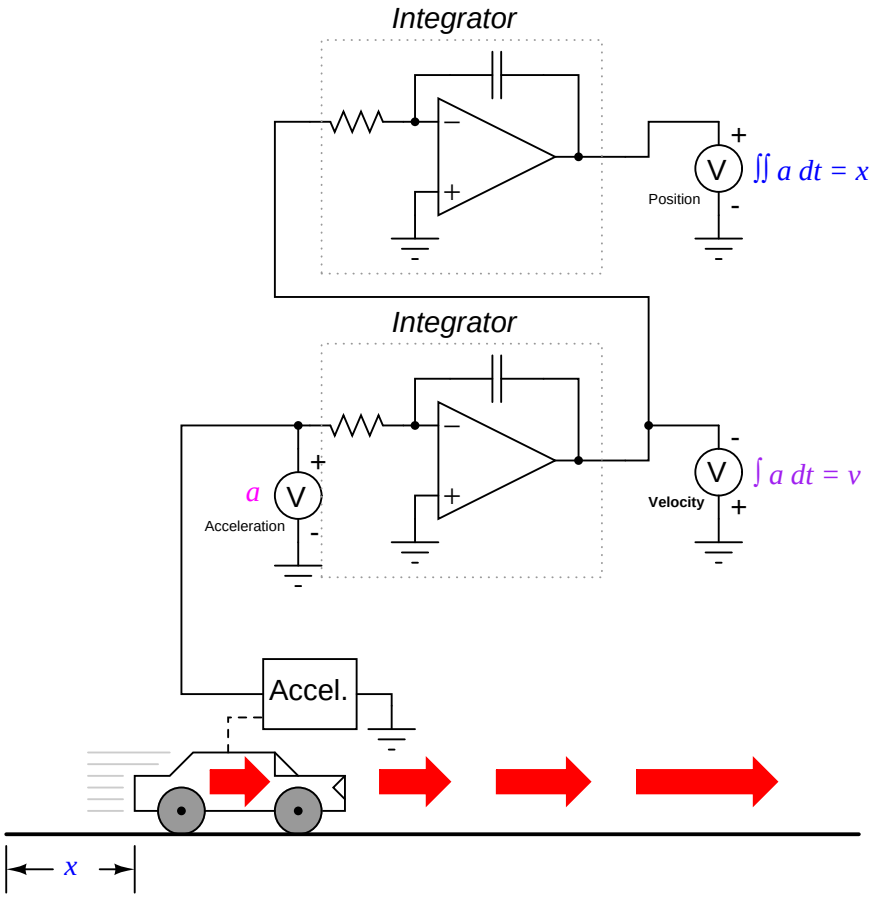
Recall from the last chapter that an integrator circuit outputs a voltage whose rate-of-change over time is proportional to the input voltage's magnitude. Thus, given a constant input voltage, the output voltage will *change* at a constant *rate*. If the car travels at a constant velocity (constant voltage 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.

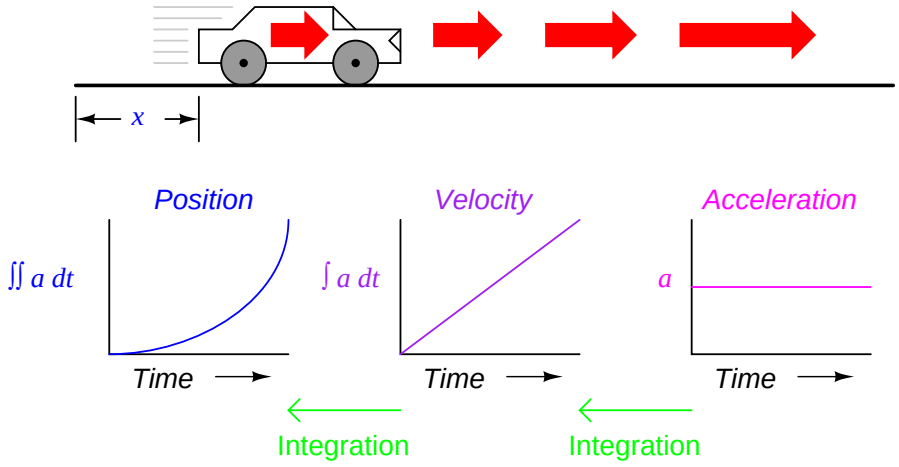
So, we may express three measures of the car's motion (position, velocity, and acceleration) in terms of velocity (v) just as easily as we could in terms of position (x):



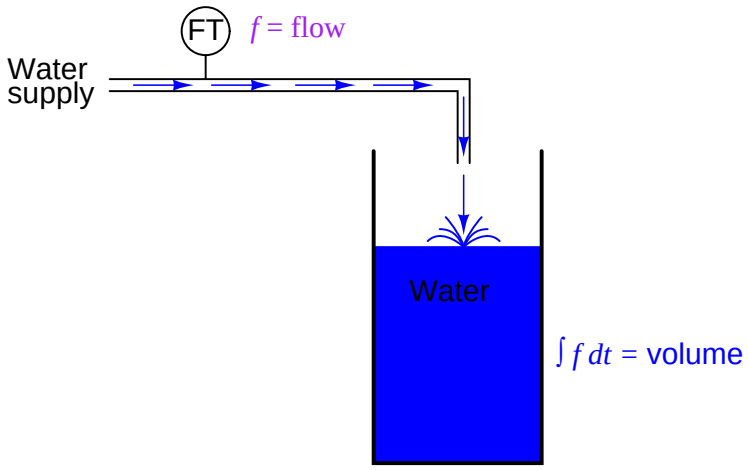
If we had an accelerometer attached to the car, generating a signal proportional to the rate of acceleration or deceleration, we could (hypothetically) obtain a velocity signal with one step of integration, and a position signal with a second step of integration:



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:

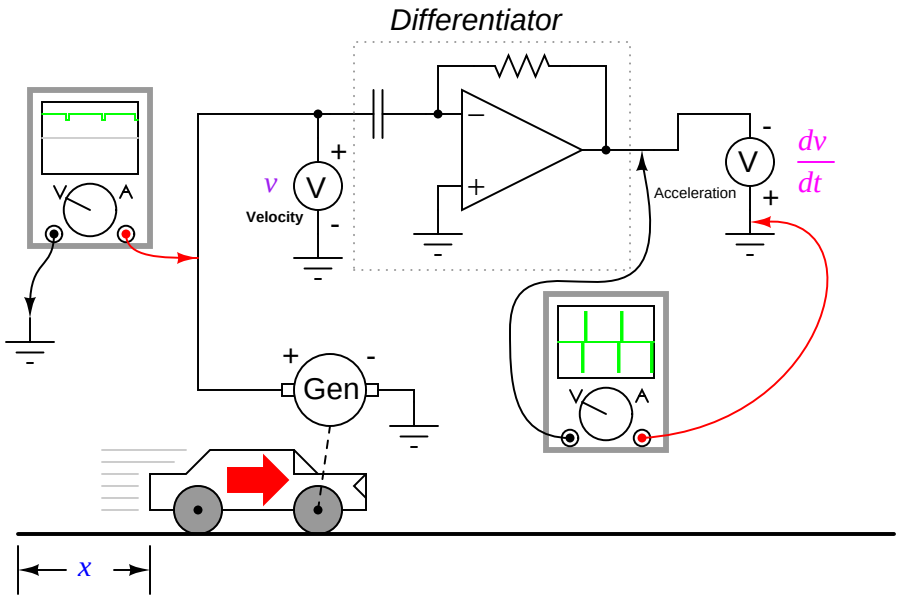


If we were to use a "Flow Transmitter" device to measure water flow, then by time-integration we could calculate the volume of water accumulated in the tank over time. Although it is theoretically possible to use a capacitive op-amp integrator circuit to derive a volume signal from a flow signal, mechanical 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.

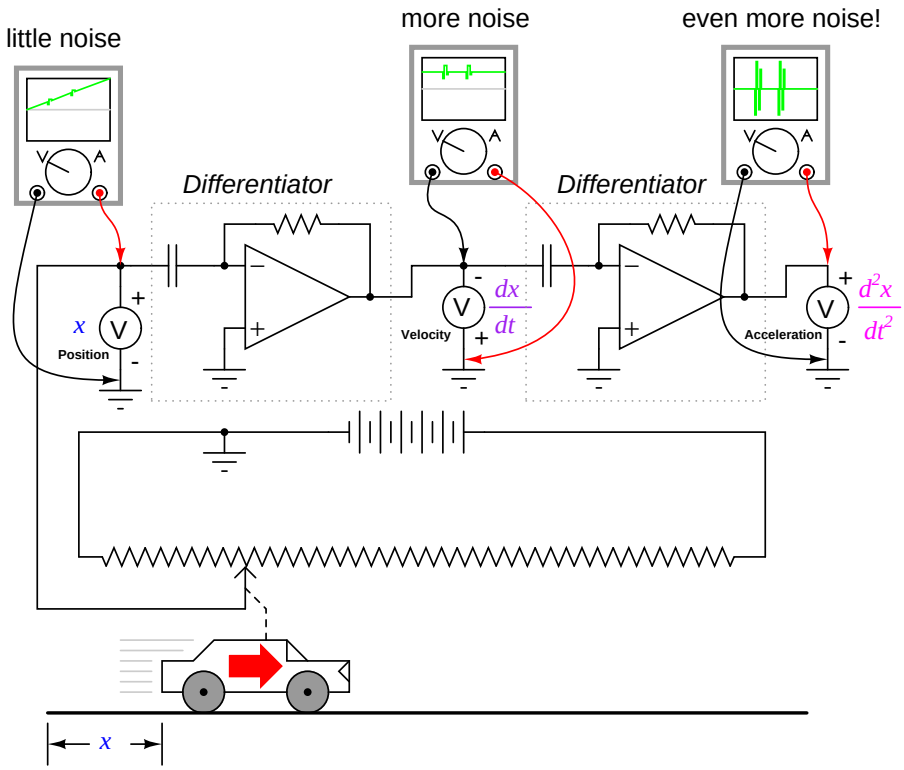
Having said this, there are caveats to the using calculus techniques to derive one type of measurement from another. Differentiation has the undesirable tendency of amplifying "noise" found in the measured variable, since the noise will typically appear as frequencies much higher than the measured variable, 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:

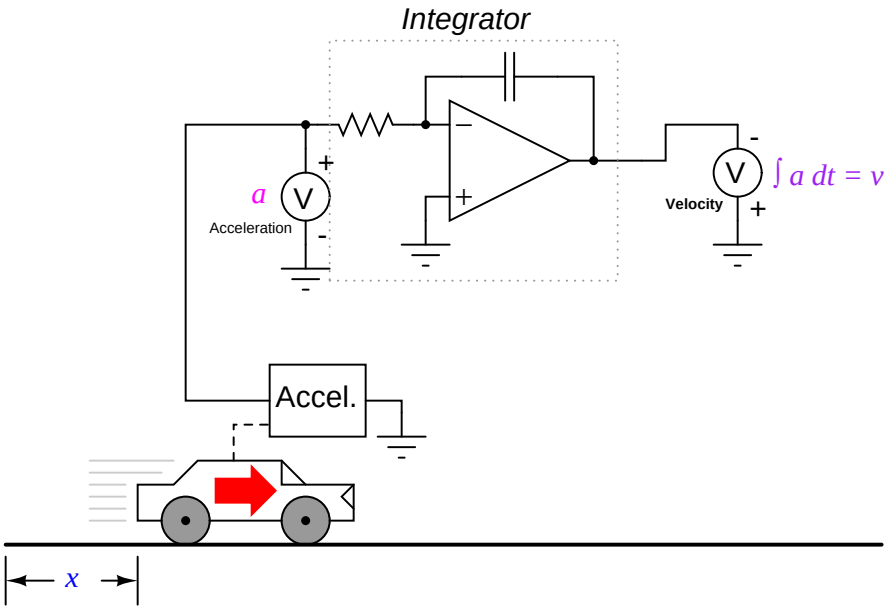


Noise voltage present in a signal to be differentiated need not be of significant amplitude to cause trouble: all that is required is that the noise profile have fast rise or fall times. In other words, any electrical noise with a high dv/dt component will be problematic when differentiated, even if 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:

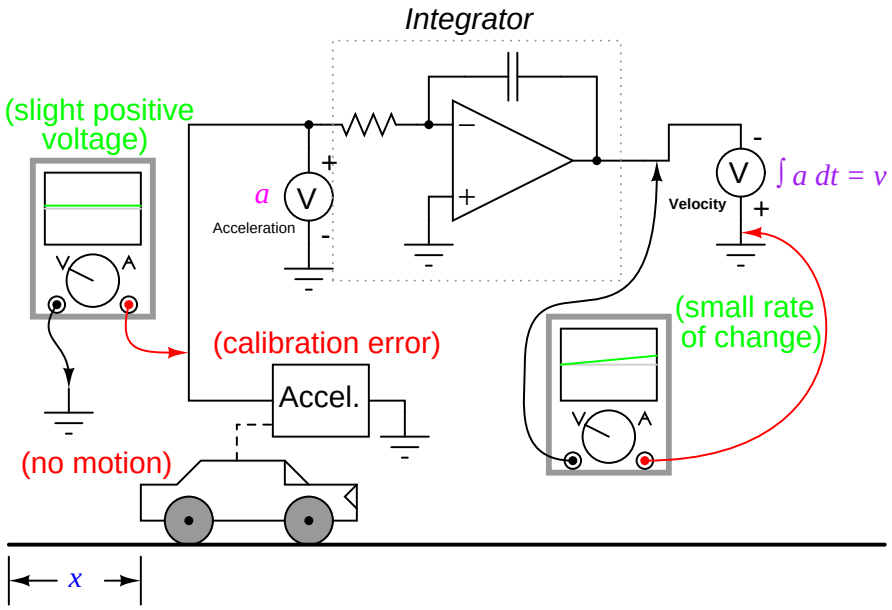


Integration does not suffer from this problem, because integrators act as low-pass filters, attenuating high-frequency input signals. In effect, all the high and low peaks resulting from noise on the signal become averaged together over time, for a diminished net result. One might suppose, then, that 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:



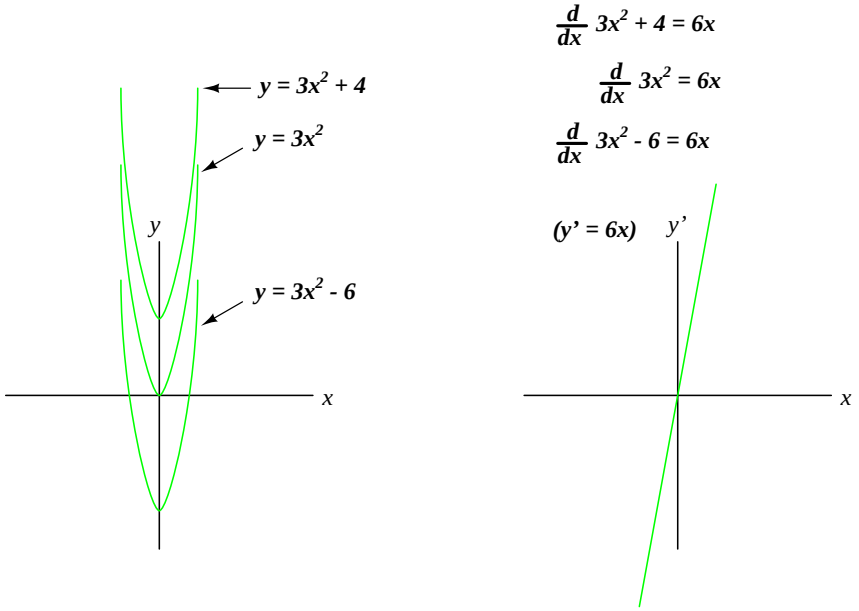
Unfortunately, following this methodology might lead us into other difficulties, one being a common artifact of analog integrator circuits known as *drift*. All op-amps have some amount of input bias current, and this current will tend to cause a charge to accumulate on the capacitor in addition to whatever 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.

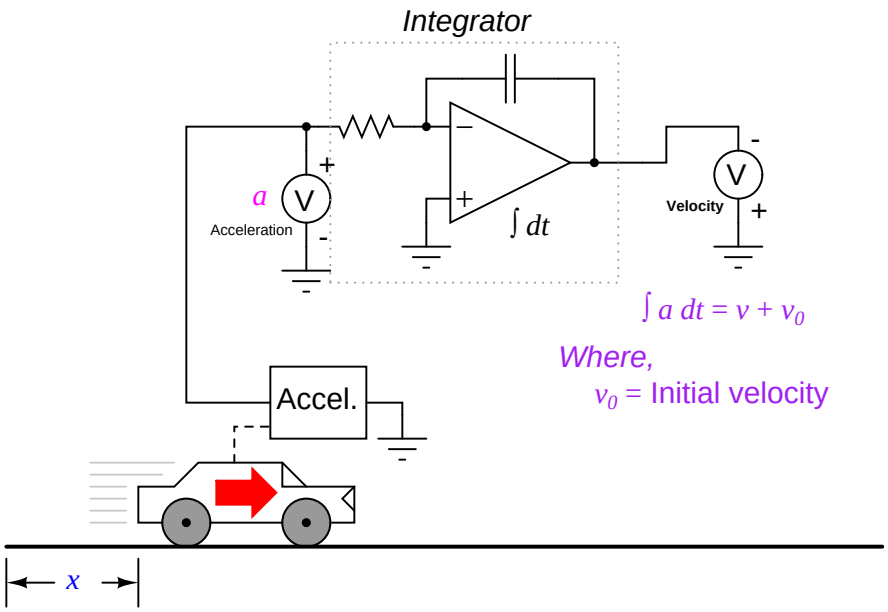
So far, the only integration errors discussed have been artificial in nature: originating from imperfections in the circuitry and sensors. There also exists a source of error inherent to the process of integration itself, and that is the *unknown constant* problem. Beginning calculus students learn that 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.

9.8 Measurement circuits – INCOMPLETE

Figure 9.25 shows a photodiode amplifier for measuring low levels of light. Best sensitivity and bandwidth are obtained with a *transimpedance amplifier*, a current to voltage amplifier, instead of a conventional operational amplifier. The photodiode remains reverse biased for lowest diode capacitance, hence wider bandwidth, and lower noise. The feedback resistor sets the "gain", the current to voltage amplification factor. Typical values are 1 to 10 Meg Ω. Higher values yield higher gain. A capacitor of a few pF may be required to compensate for photodiode capacitance, and prevents instability at the high gain. The wiring at the summing node must be as compact as possible. This point is sensitive to circuit board contaminants and must be thoroughly cleaned. The most sensitive amplifiers contain the photodiode and amplifier within a hybrid microcircuit package or single die.

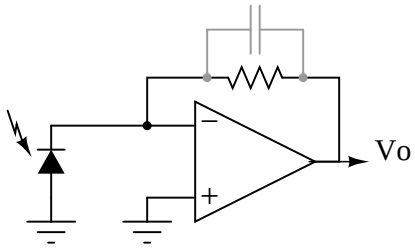


Figure 9.25: *Photodiode amplifier.*

9.9 Control circuits – PENDING

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).

Bill Marsden (April 2008) Author of "ElectroStatic Discharge" section.

Bibliography

[1] Chin-Leong Lim, Lim Yeam Ch'ng, Goh Swee Chye, "Diode Quad Is Foundation For PIN Diode Attenuator," Microwaves & RF, May 2006, at http://www.mwrf.com/Articles/Index.cfm?Ad=1&ArticleID=12523

[2] "Transistor Audio and Radio Circuits," TP1399, 2nd Ed., pp 39-40, Mullard, London, 1972.

[3] "AM Receiver Circuit TCA440," Analog Data Manual, 2nd Ed., pp 14-20 to 14-26, Signetics, 1982.

[4] Sony "8-pin Single-Chip AM Radio with Builot-in Power Amplifier," pp 5, at http://www.datasheetcatalog.com/datasheets_pdf/C/X/A/1/CXA1600.shtml

[5] Texas Instruments "Solid State Communications," pp 318, McGraw-Hill, N.Y., 1966.

[6] Texas Instruments "Transistor Circuit Design," pp 290, McGraw-Hill, N.Y., 1963.

[7] "Datasheet TDA7021T", STR-NXP Wireless, at http://www.nxp.com/acrobat_download/datasheets/TDA7021T_CNV_2.pdf

[8] Donald L. Stoner, L. A. Earnshaw, "The Transistor Radio Handbook," pp 76, Editors and Eenineers, Sumerland, CA, 1963.

[9] ,Charles Wenzel, "Crystal Radio Circuits," at http://www.techlib.com/electronics/crystal.html.

Chapter 10

ACTIVE FILTERS

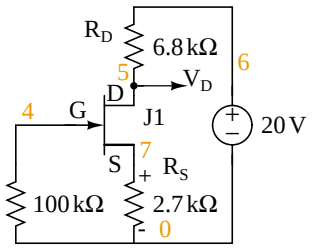
Contents

10.1 Two pole active filters 469

*** PENDING ***

10.1 Two pole active filters

Figure 10.1



I_D = 1mA

V_D = 13V±2V

R_D = (20V-13V)/1mA = 7k (6.8k)

R_S = 2.64V/1mA = 2.64k (2.7k)

Figure 10.1: *(a) 10Khz Low-pass filter. (b) 100Hz cutoff high-pass filter*

Test

$$C1 = \frac{R1 + R2}{\sqrt{2}R1R2)\omega_C} \qquad C1 = \frac{R1 + R2}{\sqrt{3}R1R2)\omega_C}$$

$$C2 = \frac{\sqrt{2}}{(R1 + R2)\omega_C} \qquad C2 = \frac{\sqrt{3}}{(R1 + R2)\omega_C}$$

Butterworth Linear Phase

Chapter 11

DC MOTOR DRIVES

Contents

11.1 Pulse Width Modulation 471

***** INCOMPLETE *****

11.1 Pulse Width Modulation

Pulse Width Modulation (PWM) uses digital signals to control power applications, as well as being fairly easy to convert back to analog with a minimum of hardware.

Analog systems, such as linear power supplies, tend to generate a lot of heat since they are basically variable resistors carrying a lot of current. Digital systems don't generally generate as much heat. Almost all the heat generated by a switching device is during the transition (which is done quickly), while the device is neither on nor off, but in between. This is because power follows the following formula:

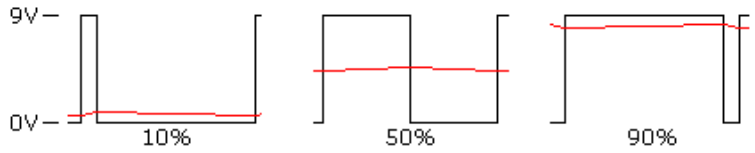
P = E I, or Watts = Voltage X Current

If either voltage or current is near zero then power will be near zero. PWM takes full advantage of this fact.

PWM can have many of the characteristics of an analog control system, in that the digital signal can be free wheeling. PWM does not have to capture data, although there are exceptions to this with higher end controllers.

One of the parameters of any square wave is duty cycle. Most square waves are 50%, this is the norm when discussing them, but they don't have to be symmetrical. The ON time can be varied completely between signal being off to being fully on, 0% to 100%, and all ranges between.

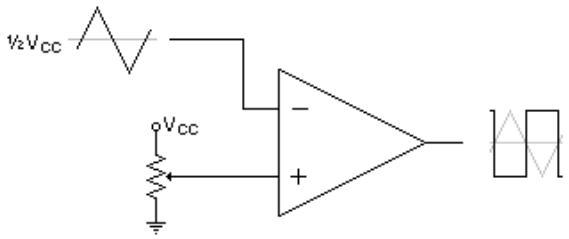
Shown below are examples of a 10%, 50%, and 90% duty cycle. While the frequency is the same for each, this is not a requirement.



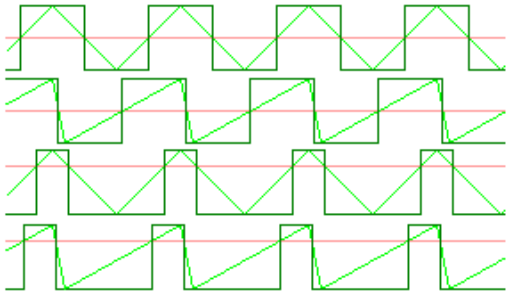
Examples of PWM Waveforms

The reason PWM is popular is simple. Many loads, such as resistors, integrate the power into a number matching the percentage. Conversion into its analog equivalent value is straightforward. LEDs are very nonlinear in their response to current, give an LED half its rated current you you still get more than half the light the LED can produce. With PWM the light level produced by the LED is very linear. Motors, which will be covered later, are also very responsive to PWM.

One of several ways PWM can be produced is by using a sawtooth waveform and a comparator. As shown below the sawtooth (or triangle wave) need not be symmetrical, but linearity of the waveform is important. The frequency of the sawtooth waveform is the sampling rate for the signal.



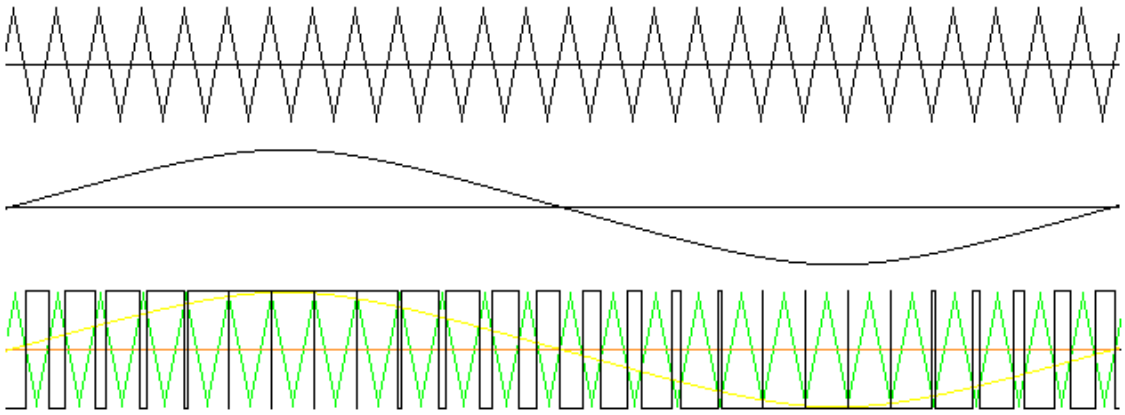
PWM Modulator



Why Ramp Symmetry Doesn't Matter

If there isn't any computation involved PWM can be fast. The limiting factor is the comparators frequency response. This may not be an issue since quite a few of the uses are fairly low speed. Some microcontrollers have PWM built in, and can record or create signals on demand.

Uses for PWM vary widely. It is the heart of Class D audio amplifiers, by increasing the voltages you increase the maximum output, and by selecting a frequency beyond human hearing (typically 44Khz) PWM can be used. The speakers do not respond to the high frequency, but duplicates the low frequency, which is the audio signal. Higher sampling rates can be used for even better fidelity, and 100Khz or much higher is not unheard of.

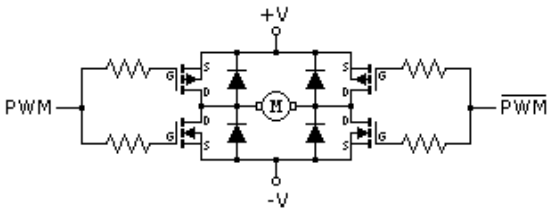


How an Audio Signal is modulated with PWM

Another popular application is motor speed control. Motors as a class require very high currents to operate. Being able to vary their speed with PWM increases the efficiency of the total system by quite a bit. PWM is more effective at controlling motor speeds at low RPM than linear methods.

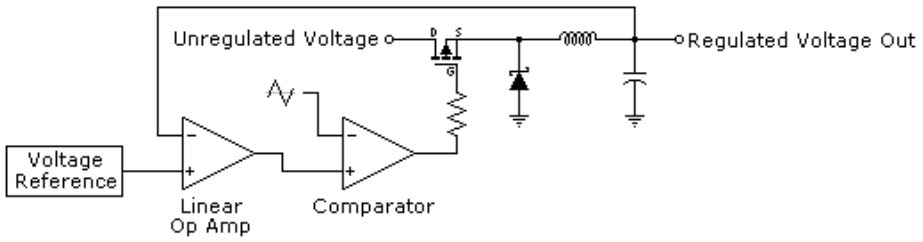
PWM is often used in conjunction with an H-Bridge. This configuration is so named because it resembles the letter H, and allows the effective voltage across the load to be doubled, since the power supply can be switched across both sides of the load. In the case of inductive loads, such as motors, diodes are used to suppress inductive spikes, which may damage the transistors. The inductance in a motor also tends to reject the high frequency component of the waveform. This configuration can also be used with speakers for Class D audio amps.

While basically accurate, this schematic of an H-Bridge has one serious flaw, it is possible while transitioning between the MOSFETs that both transistors on top and bottom will be on simultaneously, and will take the full brunt of what the power supply can provide. This condition is referred to as *shoot through*, and can happen with any type of transistor used in a H-Bridge. If the power supply is powerful enough the transistors will not survive. It is handled by using drivers in front of the transistors that allow one to turn off before allowing the other to turn on.



A simplified H Bridge

Switching Mode Power Supplies (SMPS) can also use PWM, although other methods also exist. Adding topologies that use the stored power in both inductors and capacitors after the main switching components can boost the efficiencies for these devices quite high, exceeding 90% in some cases. Below is an example of such a configuration.



Example of SMPS using PWM

Efficiency in this case is measured as wattage. If you have a SMPS with 90% efficiency, and it converts 12VDC to 5VDC at 10 Amps, the 12V side will be pulling approximately 4.6 Amps. The 10% (5 watts) not accounted for will show up as waste heat. While being slightly noisier, this type of regulator will run much cooler than its linear counterpart.

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.

Bill Marsden (February 2010) Author of "Pulse Width Modulation" section.

Chapter 12

INVERTERS AND AC MOTOR DRIVES

***** PENDING *****

Chapter 13

ELECTRON TUBES

Contents

13.1 Introduction 477
13.2 Early tube history 478
13.3 The triode 481
13.4 The tetrode 483
13.5 Beam power tubes 484
13.6 The pentode 485
13.7 Combination tubes 486
13.8 Tube parameters 489
13.9 Ionization (gas-filled) tubes 491
13.10Display tubes 495
13.11Microwave tubes 498
13.12Tubes versus Semiconductors 501

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

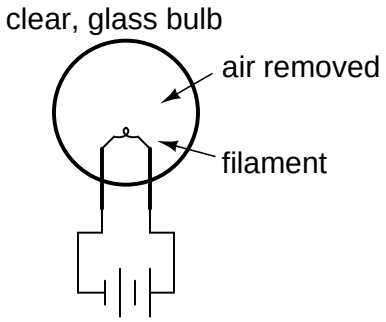
the field-effect transistor. In either case, a relatively small electric signal controls a relatively large electric current. This is the essence of the word ”electronic,” so as to distinguish it from ”electric,” which has more to do with how electron flow is regulated by Ohm’s Law and the physical attributes 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.

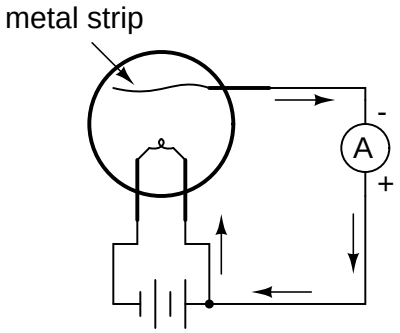
13.2 Early tube history

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).

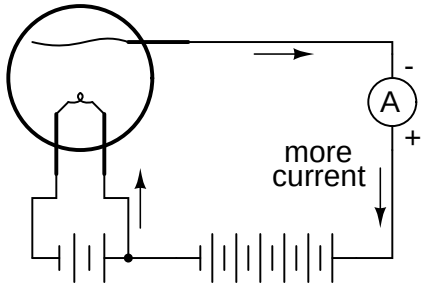
Edison was able to achieve his success by placing his filament (made of carbonized sewing thread) inside of a clear glass bulb from which the air had been forcibly removed. In this vacuum, the filament could glow at white-hot temperatures without being consumed by combustion:



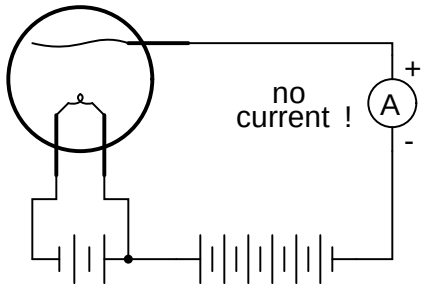
In the course of his experimentation (sometime around 1883), Edison placed a strip of metal inside of an evacuated (vacuum) glass bulb along with the filament. Between this metal strip and one of the filament connections he attached a sensitive ammeter. What he found was that electrons would flow through 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

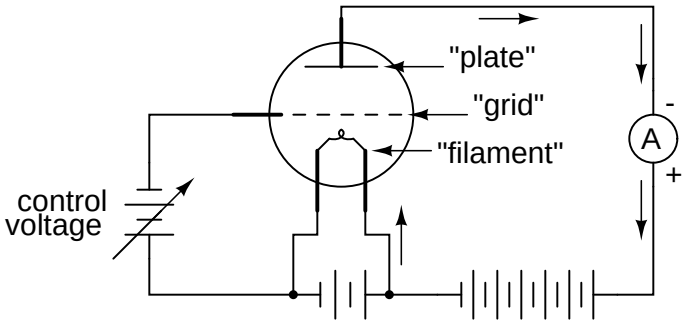


plate current can be controlled by the application of a small control voltage between the grid and filament!

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.

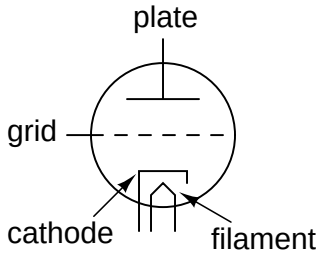
Calling his invention the "Audion," he vigorously applied it to the development of communications technology. In 1912 he sold the rights to his Audion tube as a telephone signal amplifier to the American Telephone and Telegraph Company (AT and T), which made long-distance telephone communication practical. 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.

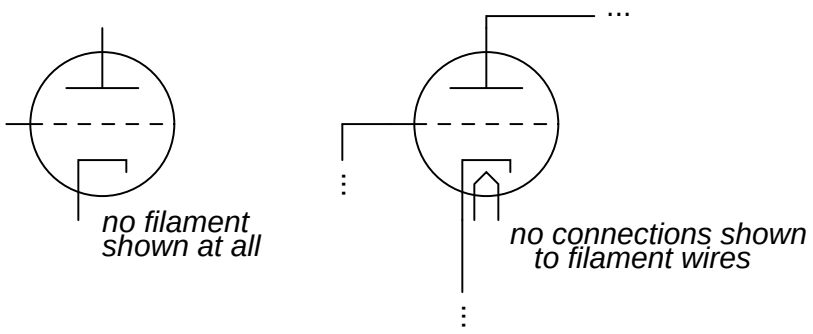
This refinement was necessary in order to avoid some undesired effects of an incandescent filament as an electron emitter. First, a filament experiences a voltage drop along its length, as current overcomes the resistance of the filament material and dissipates heat energy. This meant that the voltage 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:



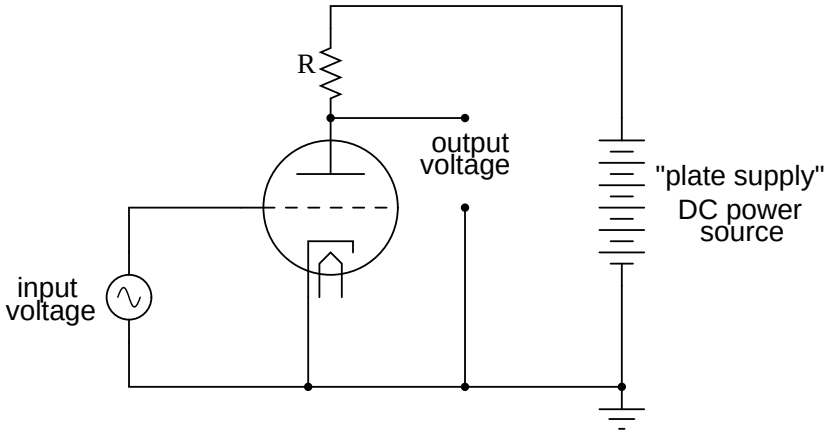
Since the filament is necessary for all but a few types of vacuum tubes, it is often omitted in

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:

Triode amplifier circuit



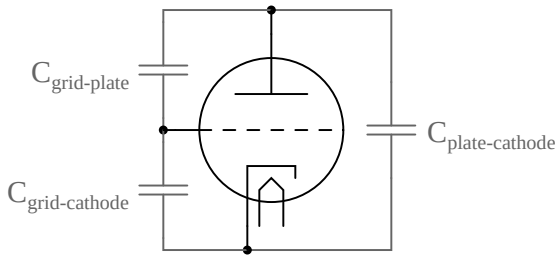
The low-voltage AC signal connected between the grid and cathode alternately suppresses, then enhances the electron flow between cathode and plate. This causes a change in voltage on the output of the circuit (between plate and cathode). The AC voltage and current magnitudes on the tube's grid are generally 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:



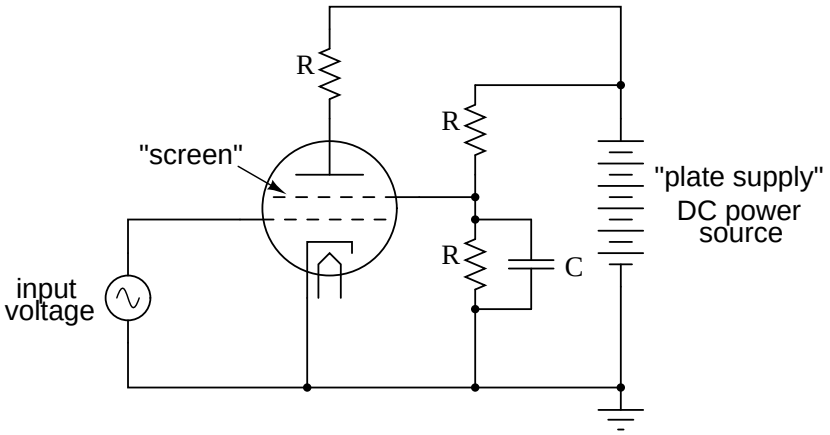
Now, this stray capacitance is quite small, and the reactive impedances usually high. Usually, that is, unless radio frequencies are being dealt with. As we saw with De Forest's Audion tube, radio was probably the prime application for this new technology, so these "tiny" capacitances became more than 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:

Tetrode amplifier circuit



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.

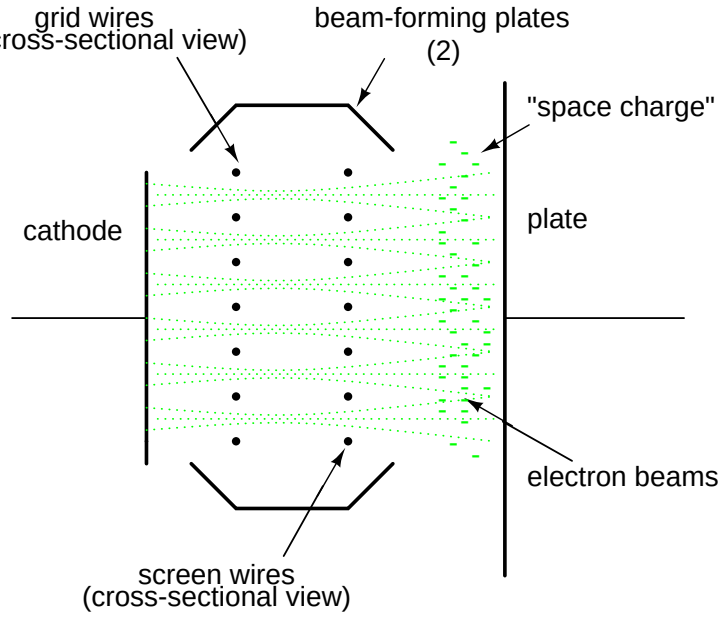
Despite the advantages realized by the addition of a screen, there were some disadvantages as well. The most significant disadvantage was related to something known as *secondary emission*. When electrons from the cathode strike the plate at high velocity, they can cause free electrons to be jarred loose 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

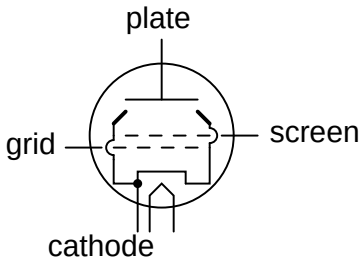
cathode to plate.



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:

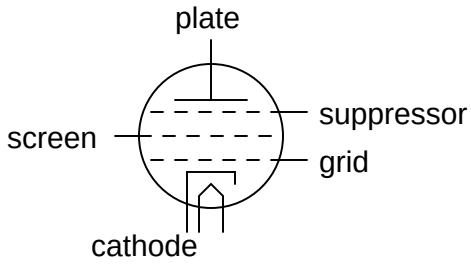
The "Beam power" tetrode tube



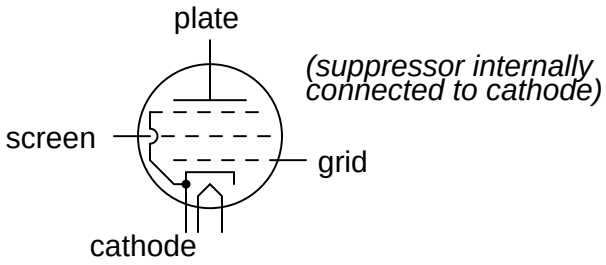
13.6 The pentode

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.

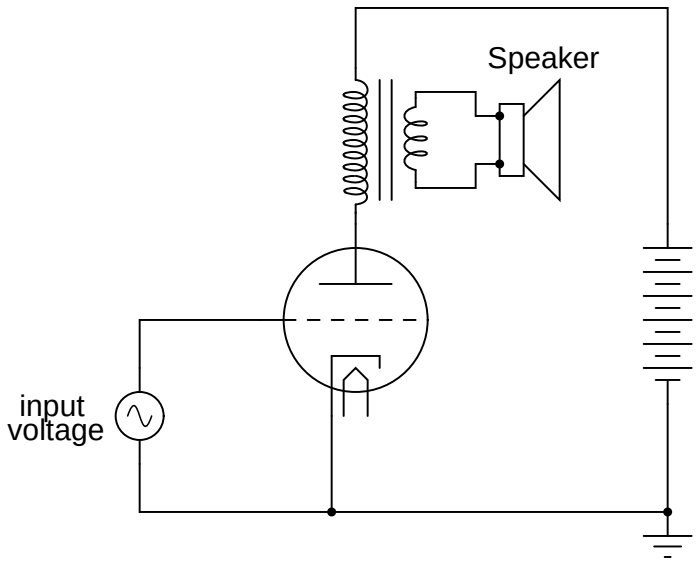
Of course, it would have been quite impossible to combine thousands of tube elements into a single tube envelope the way that thousands of transistors can be etched onto a single piece of silicon, but engineers still did their best to push the limits of tube miniaturization and consolidation. Some of 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 superhetero-

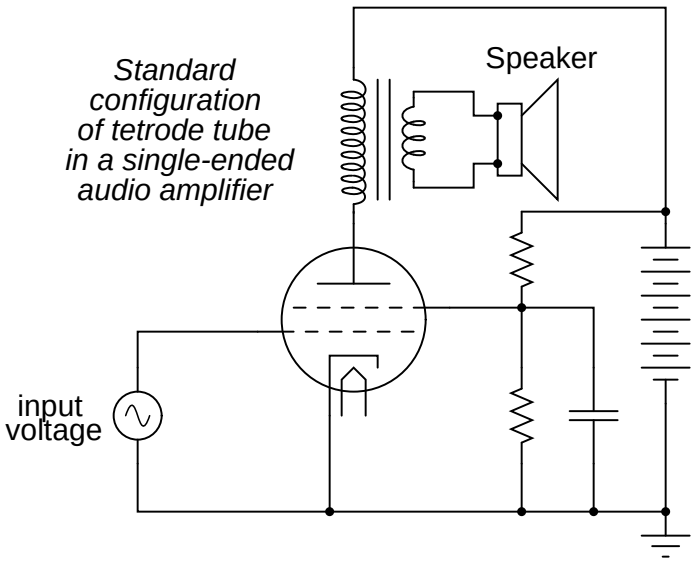
dyne radio designs. These tubes contained seven elements: 5 grids plus a cathode and a plate. Two of the grids were normally reserved for signal input, the other three relegated to screening and suppression (performance-enhancing) functions. Combining the superheterodyne functions of oscillator and signal 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.

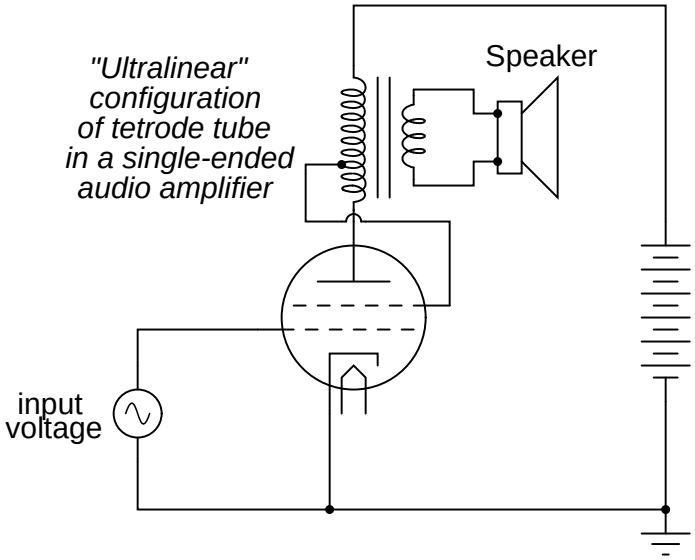
Enterprising electronics engineers also discovered ways to exploit such multi-variable capabilities of "lesser" tubes such as tetrodes and pentodes. One such way was the so-called *ultralinear* audio power amplifier, invented by a pair of engineers named Hafler and Keroes, utilizing a tetrode tube in 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:



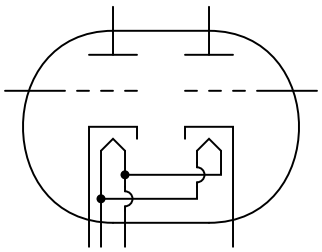
This signal feedback results in significant improvements in amplifier linearity (and consequently, distortion), so long as precautions are taken against "overpowering" the screen with

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*.

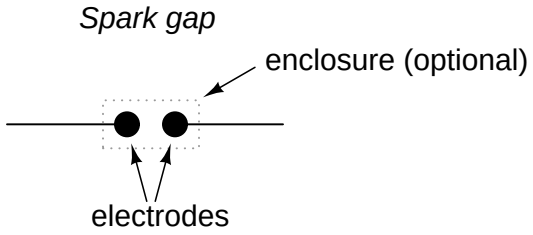
Ionized gas is not a perfect conductor. As such, the flow of electrons through ionized gas will tend to dissipate energy in the form of heat, thereby helping to keep the gas in a state

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.

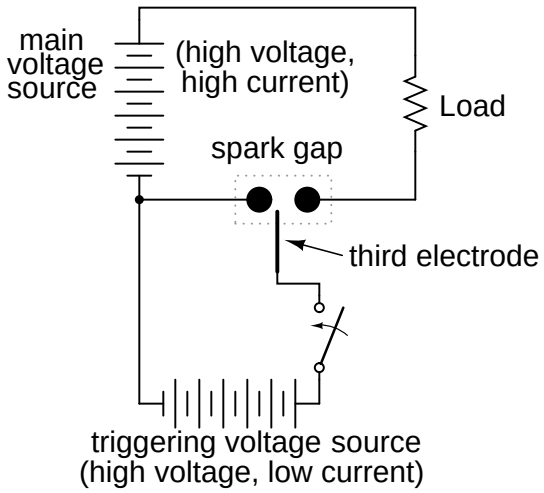


A prime application for spark gaps is in overvoltage protection. Engineered not to ionize, or ”break down” (begin conducting), with normal system voltage applied across the electrodes, the spark gap’s function is to conduct in the event of a significant increase in voltage. Once conducting, it will act 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:

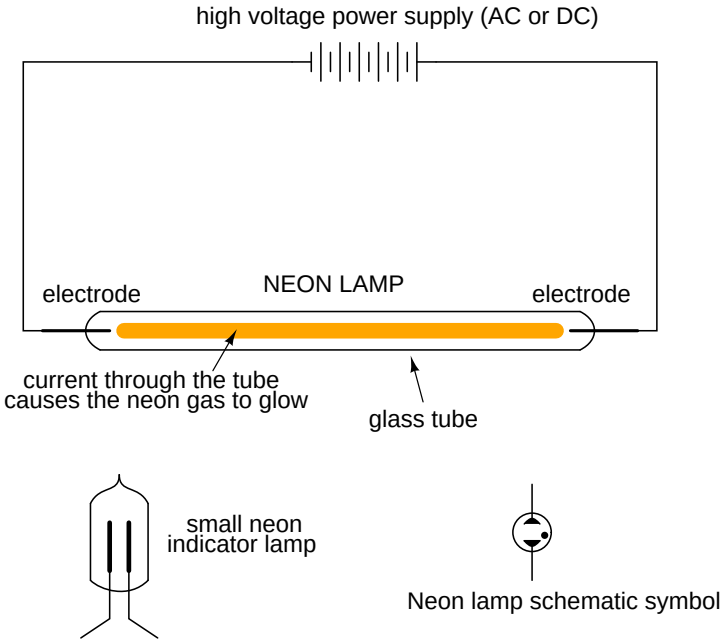
Triggered spark gap



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.

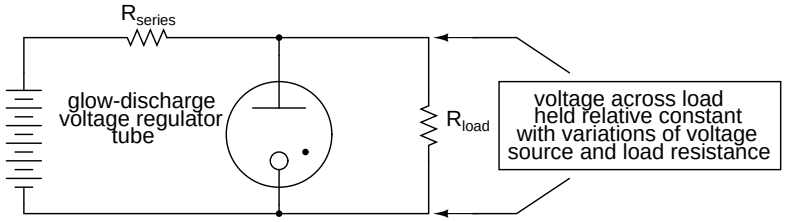
When the two main electrodes are placed in a sealed tube filled with a special gas, a *discharge tube* is formed. The most common type of discharge tube is the neon light, used popularly as a source of colorful illumination, the color of the light emitted being dependent on the type of gas filling the 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.

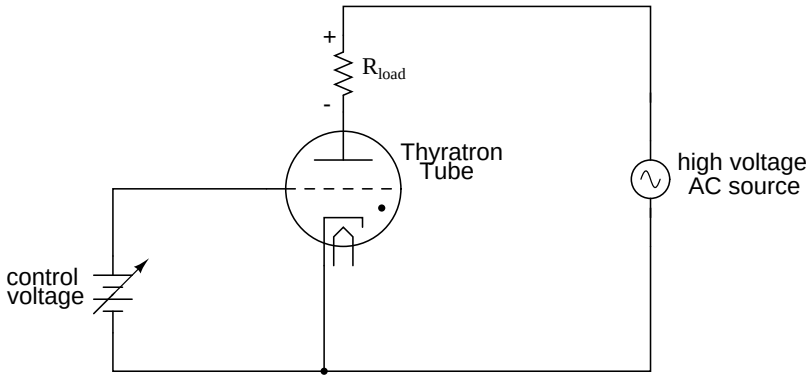


Please take note of the black dot found in the tube symbol shown above (and in the neon lamp symbol shown before that). That marker indicates the tube is gas-filled. It is a common marker used in all gas-filled tube symbols.

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*.



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.

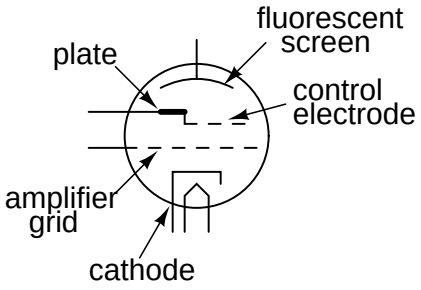
An interesting ”cousin” to the CRT is the *Cat-Eye* or *Magic-Eye* indicator tube. Essentially, this tube is a voltage-measuring device with a display resembling a glowing green ring. Electrons emitted by the cathode of this tube impinge on a fluorescent screen, causing the green-colored light to be emitted. The shape of the glow produced by the fluorescent screen varies as the amount of voltage applied to a grid changes:

"Cat-Eye" indicator tube displays

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:

"Cat-Eye" or "Magic-Eye" indicator tube

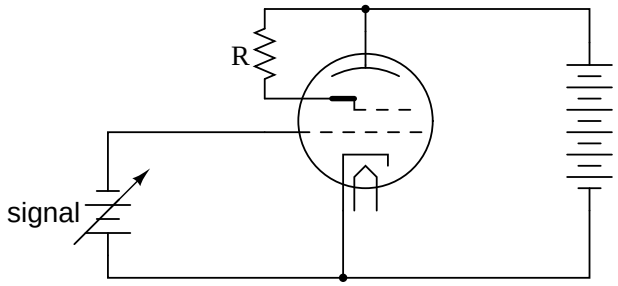


Here is a photograph of a cat-eye tube, showing the circular display region as well as the glass envelope, socket (black, at far end of tube), and some of its internal structure:

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:

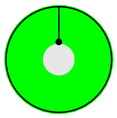
"Cat-Eye" indicator tube circuit



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).

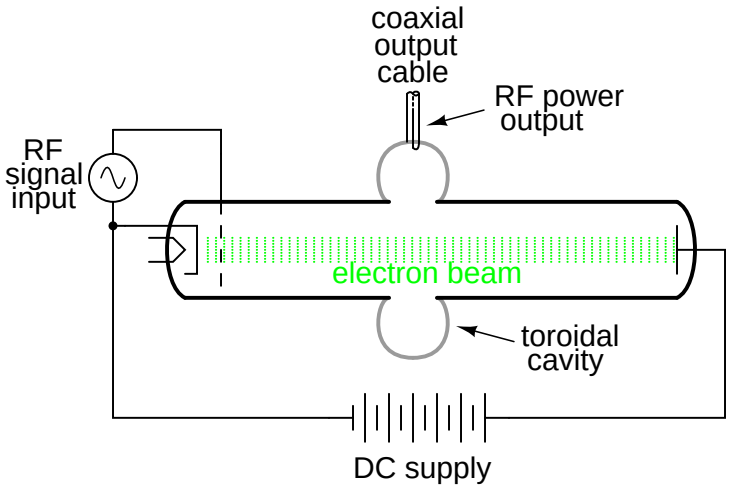
"Cat-eye" tubes were never accurate enough to be equipped with a graduated scale as is the case with CRT's and electromechanical meter movements, but they served well as null detectors in bridge circuits, and as signal strength indicators in radio tuning circuits. An unfortunate limitation to the "cat-eye" 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.

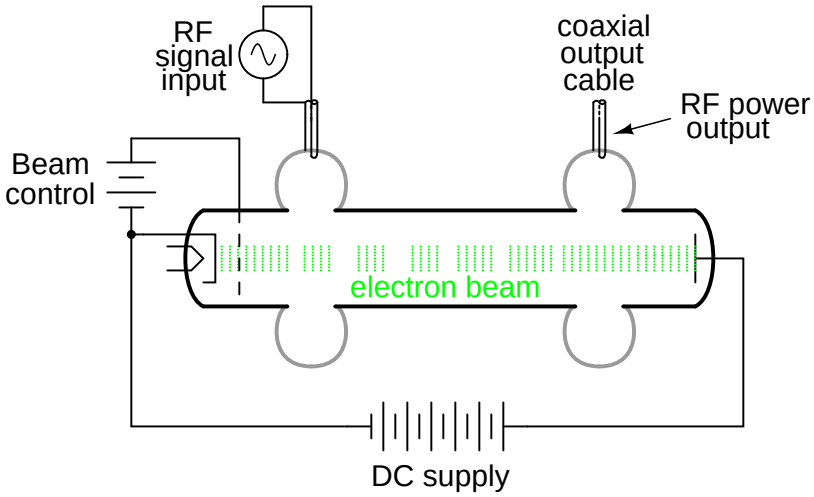
It was discovered in 1939 that a toroidal cavity made of conductive material called a *cavity resonator* surrounding an electron beam of oscillating intensity could extract power from the beam without actually intercepting the beam itself. The oscillating electric and magnetic fields associated with 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)



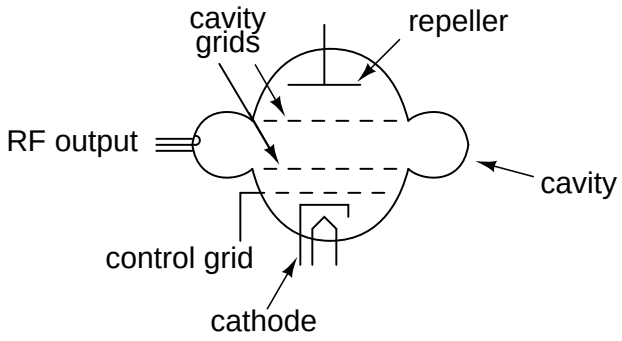
Two of the researchers instrumental in the initial development of the IOT, a pair of brothers named Sigurd and Russell Varian, added a second cavity resonator for signal input to the inductive output tube. This input resonator acted as a pair of inductive grids to alternately "bunch" and release packets 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



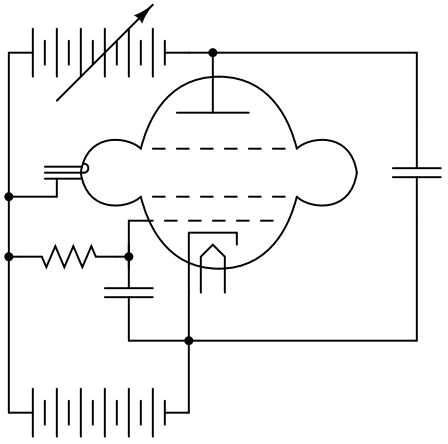
Another invention of the Varian brothers was the *reflex klystron* tube. In this tube, electrons emitted from the heated cathode travel through the cavity grids toward the repeller plate, then are repelled and returned back the way they came (hence the name *reflex*) through the cavity grids. Self-sustaining 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

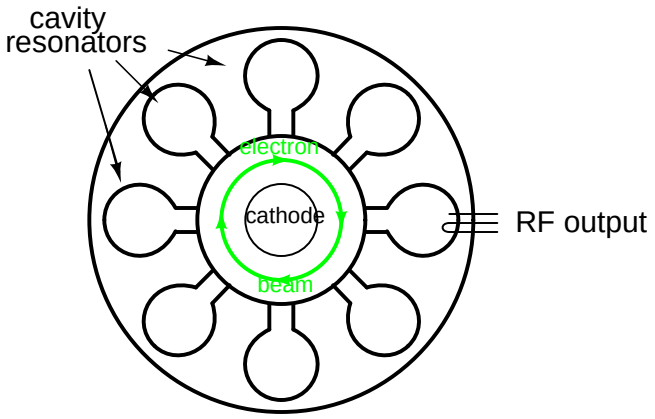


Initially developed as low-power devices whose output required further amplification for radio transmitter use, reflex klystron design was refined to the point where the tubes could serve as power devices in their own right. Reflex klystrons have since been superseded by semiconductor devices in the 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:

The magnetron tube



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. Once the author of the DSL said it was obsolete, I moved on to Creative Commons (CC BY License).

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 CC BY 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, to Richard Stallman for pioneering the concept of copyleft, and to Creative Commons for the CC BY License.

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.

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 for the CreativeCommons CC BY License:

```
Subject to the terms and conditions of this Public License, the Licensor
hereby grants You a worldwide, royalty-free, non-sublicensable,
non-exclusive, irrevocable license to exercise the Licensed Rights in the
Licensed Material to:
A. reproduce and Share the Licensed Material, in whole or in part; and
B. produce, reproduce, and Share Adapted Material

No warranties are given. The license may not give you all of the
permissions necessary for your intended use. For example, other rights
such as publicity, privacy, or moral rights may limit how you use the
material.
```

If you wish to modify this book in any way, you should 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 CC BY License:

```
You are free to:
(a) Share -- copy and redistribute the material in any medium
or format.
```

```
(b)  Adapt -- remix, transform, and build upon the material
for any purpose, even commercially.
(c)  Attribution -- You must give appropriate credit, provide
a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any
way that suggests the licensor endorses you or your use.
```

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 John Anhalt

- **Date(s) of contribution(s):** December 2008
- **Nature of contribution:** Updated lead-acid cell chemistry, Ch 11
- **Contact at:** `jpa@anhalt.org`

A-2.2.2 Benjamin Crowell, Ph.D.

- **Date(s) of contribution(s):** January 2001
- **Nature of contribution:** Suggestions on improving technical accuracy of electric field and charge explanations in the first two chapters.
- **Contact at:** `crowell01@lightandmatter.com`

A-2.2.3 Dennis Crunkilton

- **Date(s) of contribution(s):** January 2006 to present
- **Nature of contribution:** Mini table of contents, all chapters except appedicies; html, latex, ps, pdf; See Devel/tutorial.html; 01/2006.

- DC network analysis ch, Mesh current section, Mesh current by inspection, new material.i DC network analysis ch, Node voltage method, new section.
- Ch3, Added AFCI paragraphs after GFCI, 10/09/2007.
- **Contact at:** liecibiblio(at)gmail.com

A-2.2.4 Tony R. Kuphaldt

- **Date(s) of contribution(s):** 1996 to present
- **Nature of contribution:** Original author.
- **Contact at:** liec0@lycos.com

A-2.2.5 Ron LaPlante

- **Date(s) of contribution(s):** October 1998
- **Nature of contribution:** Helped create the ”table” concept for use in analysis of series and parallel circuits.

A-2.2.6 Davy Van Nieuwenborgh

- **Date(s) of contribution(s):** October 2006
- **Nature of contribution:** DC network analysis ch, Mesh current section, supplied solution to mesh problem, pointed out error in text.
- **Contact at:** Theoretical Computer Science laboratory, Department of Computer Science, Vrije Universiteit Brussel.

A-2.2.7 Ray A. Rayburn

- **Date(s) of contribution(s):** September 2009
- **Nature of contribution:** Nonapplicability of Maximum Power Transfer Theorem to Hi-Fi audio amplifier.
- **Contact at:** http://forum.allaboutcircuits.com/member.php?u=54720

A-2.2.8 Jason Starck

- **Date(s) of contribution(s):** June 2000
- **Nature of contribution:** HTML formatting, some error corrections.
- **Contact at:** jstarck@yhslug.tux.org

A-2.2.9 Warren Young

- **Date(s) of contribution(s):** August 2002
- **Nature of contribution:** Provided capacitor photographs for chapter 13.

A-2.2.10 someonesdad@allaboutcircuits.com

- **Date(s) of contribution(s):** November 2009
- **Nature of contribution:** Chapter 8, troublehooting tip end of Kelvin section.

A-2.2.11 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-2.2.12 Typo corrections and other ”minor” contributions

- *The students of Bellingham Technical College's Instrumentation program*.
- **anonymous** (July 2007) Ch 1, remove :registers. Ch 5, s/figures something/figures is something/. Ch 6 s/The current/The current. (September 2007) Ch 5, 8, 9, 10, 11, 12, 13, 15. Numerous typos, clarifications.
- **Tony Armstrong** (January 2003) Suggested diagram correction in ”Series and Parallel Combination Circuits” chapter.
- **James Boorn** (January 2001) Clarification on SPICE simulation.
- **Dejan Budimir** (January 2003) Clarification of Mesh Current method explanation.
- **Sridhar Chitta**, Assoc. Professor, Dept. of Instrumentation and Control Engg., Vignan Institute of Technology and Science, Deshmukhi Village, Pochampally Mandal, Nalgonda Distt, Andhra Pradesh, India (December 2005) Chapter 13: CAPACITORS, Clarification: s/note the direction of current/note the direction of electron current/, 2-places
- **Colin Creitz** (May 2007) Chapters: several, s/it's/its.
- **Larry Cramblett** (September 2004) Typographical error correction in ”Nonlinear conduction” section.
- **Brad Drum** (May 2006) Error correction in ”Superconductivity” section, Chapter 12: PHYSICS OF CONDUCTORS AND INSULATORS. Degrees are not used as a modifier with kelvin(s), 3 changes.
- **Jeff DeFreitas** (March 2006)Improve appearance: replace “/” and ”/” Chapters: A1, A2. Type errors Chapter 3: /am injurious spark/an injurious spark/, /in the even/inthe event/

- **Sean Donner** (December 2004) Typographical error correction in ”Voltage and current” section, Chapter 1: BASIC CONCEPTS OF ELECTRICITY,(by a the/ by the) (current of current/ of current).

  (January 2005), Typographical error correction in ”Fuses” section, Chapter 12: THE PHYSICS OF CONDUCTORS AND INSULATORS (Neither fuses nor circuit breakers were not designed to open / Neither fuses nor circuit breakers were designed to open).

  (January 2005), Typographical error correction in ”Factors Affecting Capacitance” section, Chapter 13: CAPACITORS, (greater plate area gives greater capacitance; less plate area gives less capacitance / greater plate area gives greater capacitance; less plate area gives less capacitance); ”Factors Affecting Capacitance” section, (thin layer if insulation/thin layer of insulation).

  (January 2005), Typographical error correction in ”Practical Considerations” section, Chapter 15: INDUCTORS, (there is not such thing / there is no such thing).

  (January 2005), Typographical error correction in ”Voltage and current calculations” section, Chapter 16: RC AND L/R TIME CONSTANTS (voltage in current / voltage and current).

- **Manuel Duarte** (August 2006): Ch: DC Metering Circuits ammeter images: 00163.eps, 00164.eps; Ch: RC and L/R Time Constants, simplified ln() equation images 10263.eps, 10264.eps, 10266.eps, 10276.eps.

- **Aaron Forster** (February 2003) Typographical error correction in ”Physics of Conductors and Insulators” chapter.

- **Bill Heath** (September-December 2002) Correction on illustration of atomic structure, and corrections of several typographical errors.

- **Stefan Kluehspies** (June 2003): Corrected spelling error in Andrew Tannenbaum’s name.

- **David M. St. Pierre** (November 2007): Corrected spelling error in Andrew Tanenbaum’s name (from the title page of his book).

- **Geoffrey Lessel**,Thompsons Station, TN (June 2005): Corrected typo error in Ch 1 ”If this charge (static electricity) is stationary, and you won’t realize–remove If; Ch 2 ”Ohm’s Law also make intuitive sense if you apply if to the water-and-pipe analogy.” s/if/it; Chapter 2 ”Ohm’s Law is not very useful for analyzing the behavior of components like these where resistance is varies with voltage and current.” remove ”is”; Ch 3 ”which halts fibrillation and and gives the heart a chance to recover.” double ”and”; Ch 3 ”To be safest, you should follow this procedure is checking, using, and then checking your meter.... s/is/of.

- **LouTheBlueGuru**, allaboutcircuits.com, July 2005 Typographical errors, in Ch 6 ”the current through R1 is half:” s/half/twice; ”current through R1 is still exactly twice that of R2” s/R3/R2

- **Norm Meyrowitz** , nkm, allaboutcircuits.com, July 2005 Typographical errors, in Ch 2.3 ”where we don’t know both voltage and resistance:” s/resistance/current

- **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.
- **Derek Terveer** (June 2006) Typographical errors, several in Ch 1,2,3.
- **Geoffrey Lessel** (June 2005) Typographical error, s/It discovered/It was discovered/ in Ch 1.
- **Austin@allaboutcircuits.com** (July 2007) Ch 2, units of mass, pound vs kilogram, near "units of pound" s/pound/kilogram/.
- **CATV@allaboutcircuits.com** (April 2007) Telephone ring voltage error, Ch 3.
- **line@allaboutcircuits.com** (June 2005) Typographical error correction in Volumes 1,2,3,5, various chapters ,(:s/visa-versa/vice versa/).
- **rob843@allaboutcircuits.com** (April 2007) Telephone ring voltage error, Ch 3.
- **bigtwenty@allaboutcircuits.com** (July 2007) Ch 4 near "different metric prefix", s/right to left/left to right/.
- **jut@allaboutcircuits.com** (September 2007) Ch 13 near s/if were we to/if we were to/, s/a capacitors/a capacitor.
- **rxtxau@allaboutcircuits.com** (October 2007) Ch 3, suggested, GFCI terminology, non-US usage.
- **Stacy Mckenna Seip** (November 2007) Ch 3 s/on hand/one hand, Ch 4 s/weight/weigh, Ch 8 s/weight/weigh, s/left their/left there, Ch 9 s/cannot spare/cannot afford/, Ch1 Clarification, static electricity.
- **Cory Benjamin** (November 2007) Ch 3 s/on hand/one hand.
- **Larry Weber** (Feb 2008) Ch 3 s/on hand/one hand.
- **trunks14@allaboutrcircuits.com** (Feb 2008) Ch 15 s/of of/of .
- **Greg Herrington** (Feb 2008) Ch 1, Clarification: no neutron in hydrogen atom.
- **mark44** (Feb 2008) Ch 1, s/naturaly/naturally/
- **Unregistered@allaboutcircuits.com** (February 2008) Ch 1, s/smokelsee/smokeless , s/ecconomic/economic/ .
- **Timothy Unregistered@allaboutcircuits.com** (Feb 2008) Changed default roman font to newcent.
- **Imranullah Syed** (Feb 2008) Suggested centering of uncaptioned schematics.
- **davidr@insyst_ltd.com** (april 2008) Ch 5, s/results/result 2plcs.

- **Professor Thom@allaboutcircuits.com** (Oct 2008) Ch 6, s/g/c near Ecd and near 00435.png, 2plcs.
- **John Schwab** (Dec 2008) Ch 1, Static Electricity, near Charles Coulomb: rearrangement of text segments.
- **Olivier Derewonko** (Dec 2008) Ch 4 s/orientation a voltage/orientation of a/. Ch2 s/flow though/flow through/. Ch Safe meter usage, REVIEW, s/,/./ . Ch5, s/is it/it is/.
- **dor@allaboutcircuits.com** (June 2009) Ch 1, s/nusiance/nuisance.
- **rspuzio@allaboutcircuits.com** (September 2009) Ch 8, s/logarithmic/nonlinear , 6-plcs.
- **David Lewis@allaboutcircuits.com** (September 2009) Ch 1, hide paragraph: Physical dimension also impacts conductivity. . . etc.
- **Walter Odington@allaboutcircuits.com** (January 2010) Ch 3, s/hydration another/hydration is another/ .
- **tone_b@allaboutcircuits.com** (January 2010) Ch 6 , s/must were/were/ .
- **Unregistered Guest@allaboutcircuits.com** (July 2010) Ch 1 , s/is is/it is/ .
- **Unregistered Guest@allaboutcircuits.com** (July 2010) Ch 5 , added I2 to image 00090.png .
- **Unregistered Guest@allaboutcircuits.com** (August 2010) Ch 1 , s/was one the/was one of the/ .
- **D. Crunkilton** (June 2011) hi.latex, header file; updated link to openbookproject.net .
- **Bob Arthur** (Jan 2012) images: 00046.eps, 00047.eps,00048.eps 00362.eps, graph line visibility fixed.
- **vspriyan@allaboutcircuits.com** (Jan 2013) Ch 10, Near: voltages divided by their s/currents/resistances/ .
- **Eugene Smirnoff** (Jan 2013) Ch1, s/an hypothetical/a hypothetical/ . Ch 2 s/An historic/A historic/ .
- **Gulliveig@allaboutcircuits.com** (Jan 2014) Ch4, s/significant digits/mantissa, s/1000/999/ .
- **slidercrank@allaboutcircuits.com** (Feb 2014) Ch6, s/both positive/both be positive/ .
- **Skfir@allaboutcircuits.com** (August 2015) Ch10, s/suppling/supplying/ .
- **John Wang** (Sept 2017) Ch2, s/points 1 and 4/points 1 and 6/, s/points 2 and 3/3 and 4/ .
- **Stewart Todd Morgan** (Feb 2020) Ch3, s+http://web.mit.edu/safety+https://ehs.mit.edu/workplace-safety-program/electr
- **DC** (Sept 2017) Ch12, Chi3; Reformated various tables to html/latex.
- **DC** (Nov 2021) ChA1, ChA2, ChA3; Replaced Design Science with Creative Commons CC BY license

Appendix A-3

CC BY License

A-3.1 Creative Commons Attribution 4.0 International (CC BY 4.0)

This is a human-readable summary of the full license code:
https://creativecommons.org/licenses/by/4.0/legalcode
This page was adapted from: https://creativecommons.org/licenses/by/4.0/

A-3.2 You are free to:

- **Share** – copy and redistribute the material in any medium or format
- **Adapt** – remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

A-3.3 Under the following terms:

- **Attribution** – You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **No additional restrictions** – You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits./item¿

A-3.4 Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

The full CC BY 4.0 license is available at Creative Commons:

https://creativecommons.org/licenses/by/4.0/legalcode

A-3.5 A simple explanation of your rights

This section was reused from Tony Kuphaldt, "Lessons In Industrial Instrumentation", Appendix B at:

https://ibiblio.org/kuphaldt/socratic/sinst/book/liii_0v2.pdf

This is an "open-source" textbook, which means the entirety of it is freely available for public perusal, reproduction, distribution, and even modification. All digital "source" files comprising this textbook reside at the following website:

https://www.ibiblio.org/kuphaldt/electricCircuits

The Creative Commons Attribution license grants you (the recipient), as well as anyone who might receive my work from you, the right to freely use it. This license also grants you (and others) the right to modify my work, so long as you properly credit my original authorship and declare these same rights (to my original work) for your own readers. My work is copyrighted under United States law, but this license grants everyone else in the world certain freedoms not customarily available under full copyright. This means no one needs to ask my permission, or pay any royalties to me, in order to read, copy, distribute, publish, or otherwise use this book.

If you choose to modify my work, you will have created what legal professionals refer to as a *derivative work*. The Creative Commons license broadly groups derivative works under the term *adapted material*. In simple terms, the fundamental restriction placed on you when you do this is you must properly credit me for the portions of your adaptation that are my original work. Otherwise, you may treat your adaptation the same way you would treat a completely original work of your own. This means you are legally permitted to enjoy full copyright protection for your adaptation, up to and including exclusive rights of reproduction and distribution. In other words, this license does *not* bind your derivative work under the same terms and conditions I used to release my original work, although it does require you notify your readers of this free source text and the License under which it is offered.

The practical upshot of this is you may modify my work and re-publish it as you would any other book, with the full legal right to demand royalties, restrict distributions, etc. This does not compromise the freedom of my original work, because that is still available to everyone under the terms and conditions of the Attribution license[1]. It does, however, protect the investment(s) you make in creating the adaptation by allowing you to release the adaptation under whatever terms you see fit (so long as those terms comply with current intellectual property laws, of course).

In summary, the following "legalese"[2] is actually a very good thing for you, the reader of my book. It grants you permission to do so much more with this text than what you would be legally allowed to do with any other (traditionally copyrighted) book. It also opens the door to open collaborative development, so it might grow into something far better than what I alone could create.

[1] You *cannot* pass my original work to anyone else under different terms or conditions than the Attribution license. That is called *sublicensing*, and the Attribution license forbids it. In fact, any re-distribution of my original work must come with a notice to the Attribution license, so anyone receiving the book through you knows their rights.

[2] "legalese" at https://creativecommons.org/licenses/by/4.0/legalcode

Index

α ratio, 210, 264
β ratio, 186, 264
10-50 milliamp signal, 377
4-20 milliamp signal, 377
4-layer diode, 320
741 operational amplifier, 359

A-weighted dB scale, 14
A/D converter, 363
AC-DC power supply schematic, 330
Active device, 3
Active mode, transistor, 183
Alpha ratio, 210, 264
Amplification, definition, 3
Amplifier, differential, 355
Amplifier, inverting, 369
Amplifier, noninverting, 369
Amplifier, single-ended, 355
Analog-to-digital converter, 363
Angular Momentum quantum number, 33
Anti-static foam, 285
Antilogarithm, 10
Artifact, measurement, 460
Astable, 389
Attenuator, 16
Attenuator, bridged T, 21
Attenuator, coaxial, 23
Attenuator, L, 21
Attenuator, PI, 20
Attenuator, rf, 23
Attenuator, T, 19
Avalanche photodiode, 153
Averager, 378

Band, electron, 47
Bandwidth, amplifier, 256
Bardeen, John, 60, 65
Beam power tube, 484
Bel, 8
Beta ratio, 186, 264
Beta ratio, bipolar transistor, 489
Beta variations, 187
Bias current, op-amp, 396
Bias, BJT, base, 235
Bias, BJT, calculations, 235
Bias, BJT, collector-feedback, 236
Bias, BJT, emitter, 237
Bias, BJT, voltage divider, 243
Bias, diode, 98
Bias, transistor, 195, 222
Bilateral, 292
Bipolar-mode MOSFET, 312
Bistable, 387
Brattain, Walter, 60, 65
Breakdown, diode, 102
Breakdown, transistor, 325
Breakover, thyristor, 325
Bridge rectifier circuit, 111
Bridge rectifier circuit, polyphase, 111
Bypass capacitor, 261

Calculus, 358, 384, 448
Capacitance, diode, 108
Capacitor, bypass, 261
Capacitor, coupling, 231
Capacitor, op-amp compensation, 402
Cascode amplifier, 218
Cat-Eye tube, 496
Cathode, 481
Cathode Ray Tube, 495
Center-tap rectifier circuit, 109
Characteristic curves, transistor, 186, 291
Check valve, 98
Clamper circuit, 121

Class A amplifier operation, 223
Class AB amplifier operation, 224
Class B amplifier operation, 223
Class C amplifier operation, 225
Class D amplifier operation, 225
Class, amplifier operation, 222
Clipper circuit, 117
clipper, zener diode, 142
CMRR, 391
Cockcroft-Walton, voltage multiplier, 128
Coherent light, 151
Cold-cathode tube, 492
COMFET, 312
Common-base amplifier, 210
Common-collector amplifier, 202
Common-emitter amplifier, 189
Common-mode rejection ratio, 391
Common-mode voltage, 391
Commutating diode, 130, 131
Commutation, 131
Commutation time, diode, 108
Commutation, forced, 348, 349
Commutation, natural, 331, 349
Comparator, 361
Compensation capacitor, op-amp, 402
Conduction band, 48
Conductivity-Modulated Field-Effect Transistor, 312
Constant-current diode, 162
Controlled rectifier, 335
Conventional flow, 98
Cooper pair, 80
Coupling capacitor, 231
Coupling loop, resonator, 498, 501
Critical rate of voltage rise, 325, 327
Crossover distortion, 504
Crowbar, 330
CRT, 495
Crystal radio, 434
Current mirror, 264
Current source, 184, 376
Current sourcing vs. sinking, 267
Current, diode leakage, 108
Current-limiting diode, 162
Current-regulating diode, 162
Curve, characteristic, 186, 291
Cutoff voltage, 283
Cutoff, transistor, 177, 183
Czochralski process, silicon , 75

Darlington pair, 209
Datasheet, component, 107
dB, 8
dB, absolute power measurements, 15, 16
dB, sound measurements, 14
dBA, 14
dBk, 16
dBm, 15
dBW, 16
DC restorer circuit, 121
Decibel, 8
Decibels, attenuator, 17
Decineper, 13
Degenerative feedback, 256
Derivative, calculus, 449
DIAC, 326
Differential amplifier, 355
Differential pair, 406, 407
Differentiation, 358
Differentiation, calculus, 384, 448
Diode, 98
Diode check, meter function, 104, 180
Diode equation, the, 101
Diode junction capacitance, 108
Diode leakage current, 108
Diode PIV rating, 102
Diode tube, 481
Diode, 4-layer, 73
Diode, constant-current, 162
Diode, Esaki, 144
Diode, four-layer, 320
Diode, hot carrier, 143, 159
Diode, IMPATT, 160
Diode, laser, 150
Diode, light-activated, 151
Diode, light-emitting, 146
diode, MIIM, 85
diode, MIM, 163
Diode, pin, 160
Diode, PNPN, 320
Diode, schottky, 143
Diode, Shockley, 320

Diode, snap, 159
Diode, SPICE, 163
Diode, tunnel, 144
Diode, varactor, 159
Diode, varicap, 159
Diode, zener, 135
DIP, 359
Discharge tube, 493
Distortion, amplifier, 256
Distortion, crossover, 504
dn, 13
Double-layer tunneling transistor, 84
Drift, op-amp, 402
Dropout, thyristor, 325
Dual Inline Package, 359
Dual power supply, 355
Duty cycle, square wave, 362
Duty cycle, squarewave, 225

Edison effect, 479
Effect, Edison, 479
Electrode, cathode, 481
Electrode, grid, 480
Electrode, screen, 483
Electrode, suppressor, 485
Electron, 28
Electron flow, 98
Emitter follower, 205
Emitter-follower amplifier, 202
Equation, diode, 101
Equilibrium, 364
Esaki diode, 144
Exclusion principle, 36

Failure mode, zener diode, 136
Faraday's Law, 130, 131
Feedback, amplifier, 256
Feedback, negative, 364
Feedback, positive, 386
FET, field effect transistor, 65
Field effect transistor , 65
Firing, thyristor, 325
Flash converter, 363
Floating, 177, 327
Flow, electron vs. conventional, 98
Foam, anti-static, 285

Forced commutation, 348, 349
Forward bias, 98
Forward voltage, diode, 100
Four-layer diode, 320
Frequency response, op-amp, 402
Full-wave rectifier circuit, 109, 111

Gain, 6
Gain, AC versus DC, 7
Gate turn off switch, 73
Gate-Controlled Switch, 328
Gate-Turn-Off thyristor, 328
GCS, 328
Glow tube, 494
Grid, 480
Ground, 354
Ground, virtual, 369
GTO, 328
GTO, gate turn off switch, 73

Half-wave rectifier circuit, 108
Harmonic, 339
Harmonic, even vs. odd, 339
Harmonics and waveform symmetry, 339
Heptode, 486
hfe, 187
High temperature superconductors:, 82
Holding current, SCR, 330
hot carrier diode, 143
Hot-cathode tube, 492
Hybrid parameters, 187
Hysteresis, 387, 492

IC, 267
IGBT, 312, 351
IGFET, insulated gate field effect transistor, 70
IGT, 312, 351
IMPATT diode, 160
Inductive output tube, 498
Inert elements, 38
Input, inverting, 356
Input, noninverting, 356
Insulated gate field effect transistor, 70
Insulated-Gate Bipolar Transistor, 312, 351
Insulated-Gate Transistor, 312, 351

Integrated circuit, 267
Integration, calculus, 384, 448
Inverting amplifier, 192, 369
Inverting summer, 379
Ionization, 316, 491

JFET, junction field effect transistor, 65
Josephson junctions, 80
Josephson transistor, 81
Joule's Law, 11, 136
Junction capacitance, diode, 108

Kickback, inductive, 130
Kirchhoff's Current Law, 175
Kirchhoff's Voltage Law, 205
Klystron, 498

Laser diode, 150
Laser light, 151
Latch-up, 394
Latching, thyristor, 325
Leakage current, diode, 102, 108
LED, 146
Light-emitting diode, 146
Lilienfeld, Julius, 65
Load line, 226
Logarithm, 10

Magic-Eye tube, 496
Magnetic quantum number, 33
Magnetic tunnel junction, 88
Mechanics, quantum, 32
MESFET, metal semiconductor field effect transistor, 68
Metal oxide field effect transistor, 70
Mho, 294
Microphonics, electron tube, 503
MIIM, diode, 85
Miller effect, 277
MIM diode, 163
Monochromatic light, 151
MOS Controlled Thyristor, 350
MOS-gated thyristor, 350
MOSFET, metal oxide field effect transistor, 70
MTJ, magnetic tunnel junction, 88

Mu, tube amplification factor, 487
Multiplier circuit, diode, 123
Multiplier, frequency, varactor, 433

Natural commutation, 331, 349
Negative feedback, 256, 364
Negative resistance, 144
Neper, 13
Neutron, 28
Noble elements, 38
Noninverting amplifier, 369
Noninverting summer, 379
Number, quantum, 33

Offset null, op-amp, 395
Offset voltage, op-amp, 394
Ohmic region, JFET, 293
Op-amp, 262, 359
Operational amplifier, 262, 359
Orbital, electron, 35
Oscillator, 256
Oscillator, op-amp, 389
oscillator, phase shift, 433
Oscillator, relaxation, 317
Oscillator, voltage-controlled, 499
Over-unity machine, 5

Passive averager, 378
Passive device, 3
Pauli, exclusion principle, 36
PCB, 106
Peak detector, 115
Pentagrid tube, 486
Pentode tube, 304
Perpetual motion machine, 3
Phase shift, op-amp, 403
Photodiode, 151
Photodiode amplifier, 465
Photodiode, APD, 153
Photodiode, PIN, 153
PI-network , 17
PIN diode, 160
PIN, photodiode, 153
Pinch-off voltage, 283
PIV rating, diode, 102
Plasma, 316, 491

PNPN diode, 320
Polyphase bridge rectifier circuit, 111
Positive feedback, 256, 316, 386
Power supply schematic, AC-DC, 330
Principal quantum number, 33
Printed circuit board, 106
Process variable, 357
Programmable unijunction transistor, 344
Proton, 28
Pulse-width modulation, 362
Push-pull amplifier, 223
PWM, 362

Quantum dot, 86
Quantum dot transistor, 86
Quantum mechanics, 32
Quantum number, 33
Quantum physics, 28
quantum tunneling, 83
Quiescent, 226

Radio, crystal, 434
Rail voltage, 366
Rectifier, 98
Rectifier circuit, 108
Rectifier circuit, full-wave, 109, 111
Rectifier circuit, half-wave, 108
Rectifier, controlled, 335
Reference junction, thermocouple, 396
Reflex klystron, 499
Regenerative feedback, 256, 316
Regulator, voltage, 207
Relaxation oscillator, 317
Resistance, negative, 144
Resonant tunneling diode, 84
Restorer circuit, 121
Reverse bias, 98
Reverse recovery time, diode, 108
Reverse voltage rating, diode, 102
Rheostat, 188, 294
Richter scale, 9
Ripple voltage, 113
Runaway, thermal, 259

s,p,d,f subshell notation, 34
Saturable reactor, 3
Saturation voltage, 366
Saturation, transistor, 177, 183
Schottky diode, 143
SCR, 327, 495
SCR bridge rectifier, 335
SCR, silicon controlled rectifier, 73
Screen, 483
SCS, 348
Secondary emission, 484
Semiconductor, defined, 48
Sensitive gate, SCR, 330
Setpoint, 357
Shell, electron, 33
Shockley diode, 320
Shockley, William, 60, 65, 73
Siemens, 294, 377
Signal, 10-50 milliamp, 377
Signal, 4-20 milliamp, 377
Silicon controlled rectifier, 73
Silicon-controlled rectifier, 327, 495
Silicon-controlled switch, 348
Single-ended amplifier, 355
Sink, current, 267
Slicer circuit, 117
Slide rule, 10
Small-scale integration, 407
Snap diode, 159
Snubber, 131
Solar cell, 153
Solid-state, 2
Sound intensity measurement, 14
Spark gap, 492
SPICE, diode, 163
Spin quantum number, 33
Spintronics, 88
Split power supply, 355
SQUID:, 81
SSI, 407
Step recovery diode, 159
Subshell notation, 34
Subshell, electron, 34
Superconduction quantum interference device, 81
Superconductivity, 79
Superposition theorem, 231
Suppressor, 485

Switching time, diode, 108

T-network , 17
Tetrode tube, 304, 483
Theorem, Superposition, 231
Thermal runaway, BJT, 259
Thermal voltage, diode, 101
Thermocouple, 396
Three-phase bridge rectifier circuit, 111
Thyratron, 495
Thyratron tube, 318
Thyristor, 73, 492
Time, diode switching, 108
Totalizer, 385
Transconductance, 294, 377
Transconductance amplifier, 377
Transistor, field effect, 65
Transistor, insulated gate field effect, 70
Transistor, Josephson, 81
Transistor, metal oxide field effect, 70
Transistor, programmable unijunction, 344
Transistor, single electron, 86
Transistor, unijunction, 342
Triode tube, 304, 318, 481
Tube, discharge, 493
Tunnel diode, 144
Tunnel junction, magnetic, 88
tunneling, quantum, 83

Unijunction transistor, 342
Unipolar, conduction, 65
Unit, bel, 8
Unit, decineper, 13
Unit, mho, 294
Unit, neper, 13
Unit, siemens, 294, 377

Valence band, 48
Valence shell, 34
Valve, “check”, 98
Varactor diode, 159
Varicap diode, 159
VCO, 319
Virtual ground, 369
VMOS transistor, 312
Voltage buffer, 366
Voltage doubler circuit, 123
Voltage follower, 205, 366
Voltage multiplier circuit, 123
Voltage multiplier, Cockcroft-Walton, 128
Voltage regulator, 207
Voltage regulator tube, 494
Voltage rise, critical rate of, 325, 327
Voltage, bias, 195, 222
Voltage, common-mode, 391
Voltage, forward, 100
Voltage, op-amp output saturation, 366
Voltage, ripple, 113
Voltage-controlled oscillator, 319, 499
Volume units, 15
VU scale, 15

Waveform symmetry and harmonics, 339

Zener diode, 135
Zener diode failure mode, 136
Zener diode, clipper, 142

.