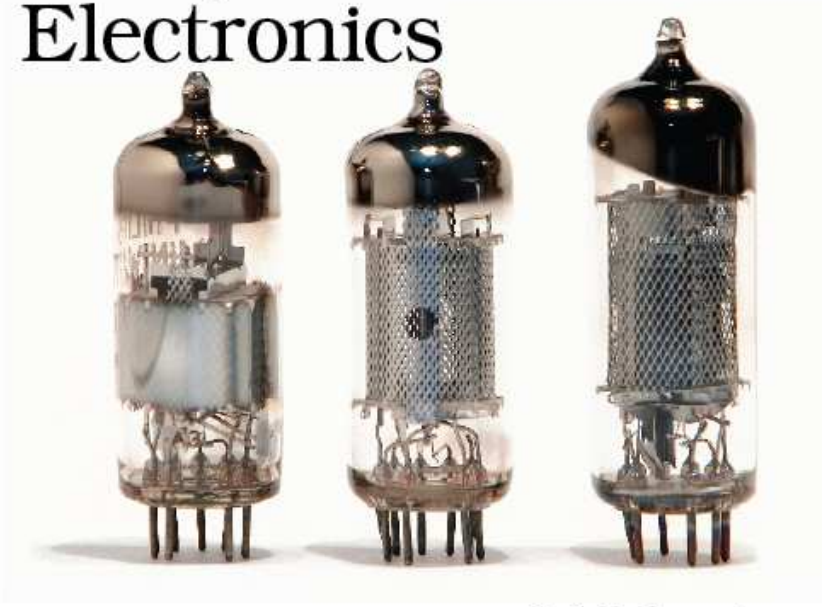
The Pragmatic Programmers

A Peek at Computer Electronics



Caleb Tennis

Things You Should Know

The *Things You Should Know* Series

This series is a little different from our usual books. The *Things You Should Know* series highlights interesting topics in technology and science that you should know about. Maybe you took these courses in school, and promptly forgot about them. Or maybe you've always been curious but never had the opportunity to learn more.

Now you can. With these titles, you can quickly become familiar with (or remind yourself of) an interesting topic area. We hope it gives you something to talk about at the next cocktail party, or brown-bag lunch at work, or user's group meeting. It might even further inspire you to delve into the topic more deeply.

In either case, we sincerely hope you enjoy the show. Thanks,

► **Andy Hunt**

Things You Should Know

A Peek at Computer Electronics

Caleb Tennis

The Pragmatic Bookshelf
Raleigh, North Carolina Dallas, Texas

Our Pragmatic courses, workshops, and other products can help you and your team create better software and have more fun. For more information, as well as the latest Pragmatic titles, please visit us at

http://www.pragmaticprogrammer.com

P1.2 printing, November 2007

Version: 2009-9-21

Contents

Chapter 1

Introduction

Let's face it—we take electronics for granted. All of our modern conveniences, from dishwashers to MP3 players, have some internal electronic components. These electronics are created with the intent to make our everyday lives easier.

So many of the things we take for granted everyday relies on some form of electronics. Without electronics, it would be impossible to enjoy so many of the modern conveniences we have come to rely on. Of course, they don't always work correctly 100% of the time. When your cell phone gets no signal or when your portable music player locks up in the middle of a song, the enamor for electronics goes away completely. However, their ubiquity cannot be overlooked.

And yet, with all of the conveniences and frustrations that electronics provide us, very few of us have any understanding as to what exactly make the whole thing work. Certainly, we're all aware of the terms voltage, current, electrons, and things like AC and DC, but for many of us the understanding of what those things really are stops short of just some vague notions. The vacuum tube, one of the more important electronics inventions, is shown on the cover of this book. And while most of us may know of the term "vacuum tube", very few of us know what it does or how it works.

This book is designed to help explain the core concepts of electronics, specifically targeted towards readers interested in computer technology. The main focus of this book is to give you an understanding what's really going on behind the scenes and how this makes the computer work. The idea is to give an inside view to people who already have an appreciation for computers. This isn't an introductory look at computers, but instead a look at how they tick. Of course, to get there a good

portion of the book focuses just on basic electronics and electricity, from how it gets to your house to how it works within the computer itself.

Of course, trying to tackle every topic in great detail is simply impossible, and it was not the goal in writing this book. There are many other good books which specialize in explaining various aspects of electronics and computer electronics. This book was meant to give some insight into the various aspects of the computer that most of us work with everyday, while trying to stay fresh and interesting as the material moves along. Unfortunately the details in some areas are not covered as well as some readers may like. I encourage you to give feedback through the publisher's website to tell what areas you would like to see covered in more detail. They may be included in future revisions of the book.

I hope you enjoy it. Furthermore, I hope you come away with a greater understanding and appreciation for all things electronic.

1.1 The disclaimer

Throughout the book, I make reference to values that are conventionally used throughout the United States. For example, I may refer to electrical power being distributed at 60 Hertz. This is not the case in many other parts of the world, where electrical standards differ. I tried my best to explain other common scenarios that are used in other parts of the world. In some cases, however, it's not easy to generalize these things.

Similarly, the nomenclature for electrical standards used in the book are the ones commonly used in the US. The same naming schemes and conventions may not be used in the same way throughout the rest of the world.

You may find terminology in this book that, if you already know about the concept, may seem illogical. For example, when talking about AC waveforms I sometimes refer to it as an AC Voltage. The direct meaning of Alternating Current Voltage doesn't make sense, but the logical concept of an alternating voltage does. I consider this notation similar to referring to an ATM as an ATM Machine. It's simply the convention that is used most commonly when teaching about the concepts.

Sometimes in order to help explain a concept I use an example and a picture that help to describe what's going on. On the surface the

description is logical, but the underlying physics may actually explain something different. For example, the description of electron flow is described somewhat in terms of atom-to-atom jumping by electrons though the actual physics is a bit different. My goal is to use the more simplified approach in the explanation. After reading the text, I highly recommend a visit to the website http://amasci.com/miscon/eleca.html which has a list of popular misconceptions about electricity.

In some instances the dates of historic events are different based on the source. When unable to find multiple reliable sources, I tried generalizing the date to a time period. Even in the case of multiple source verification, sometimes it's still possible to be incorrect at pin-pointing an exact date.

I welcome your errata and suggestions as to making the book a better resource for people wanting to learn about the topics contained inside.

1.2 Notation

In dealing with very large and very small numbers, we sometimes use the concept of scientific notation throughout the book. This means that instead of writing a number like 5000000, we would write it as 5 x 10^6, or simply 5e6. Similarly, 2.4e-7 would be scientific notation for 0.00000024.

Sometimes to deal with large and small values, we use *SI* prefixes, which come from the International System of Units[1]. For example, instead of writing 0.003 amps we write 3 milliamps, or simply 3 mA.

1.3 Organization

This book is divided into three major sections:

Electronic Fundamentals

In the first section of the book,*Basic Electricity*, we take the atomic fundamentals and expand them into the concepts needed to understand electricity at its basic level.

1. see http://en.wikipedia.org/wiki/SI_prefix for the list of prefixes

In *Electrical Power*, we look at the history of the development of electricity for the use of providing energy and powering electro-mechanical devices.

Next, in *Making Waves* we stop to analyze and study one of the most important concepts in electricity: the wave.

Finally, in *The Power Supply* we bring all of the previous concepts together to take a look at a computer power supply and how it performs its tasks of *rectification* and *providing DC power*.

Microprocessor Technology

In the section on microprocessors, we discuss the theory needed to understand how the processor works.

First, we talk about *Semiconductors*. In this section we study the history of the semiconductor and the physics behind how semiconductors work.

Next, we put the knowledge of semiconductors together to look at *Transistors*. Since the transistor is so important to microprocessors it is only fitting to take a look at their history and how they are created.

In the *Processor* section, we put transistors together to create an entire processor.

Finally, in *The Motherboard*, we study how the processor works and all of the peripheral components the processor may need in order to do its work.

Peripheral Technology

In the final section of the book, we look at peripherals of the computer, how they work, and a look at the electronics functionality that they provide. In *Data Storage*, we examine technologies such as RAM, hard disk drives, and flash memory. In the section on *Networking* we discuss the various types of networking technology, and the electronics concepts behind them. For *External Devices* we look at the peripheral technology of things that are external to the main computer box. This includes videos monitors, keyboards and mice, serial and parallel ports, and USB. Finally, in *Wireless* we look at the ideas behind wireless communications and how it relates to the computing world.

Finally, in the appendix of the book, *The Low Level* we have a refresher as to how electricity is formed at the atomic level, for anyone who might

want to a quick refresher. Some readers may enjoy starting the book with the appendix to help remember just how the electricity is formed at the atomic level.

Part I

Electronic Fundamentals

Chapter 2

Basic Electricity

We are all familiar with the aspects of electricity seen in daily life, such as lightning, batteries, and home appliances. But what is similar to all of these with respects to electricity? The answer lies in their atoms.

2.1 What is electricity?

Every material, be it solid, liquid, or gas contains two basic sub-atomic particles that house a fundamental property known as *electrical charge*. These particles are the *proton* and the *electron*. The proton and electron each contain the same amount of electrical charge, however their type of charge is exactly opposite of each other. We distinguish the two by defining the proton's charge as positive and the electron's charge as negative. Electricity is simply the movement (or "flow") of this electrical charge.

These equal and opposite charges are simply facets of nature, and are indicative of many other paired characteristics of the physical world. For example, Sir Isaac Newton's famous "third law" tells us that every action has an equal an opposite reaction. Magnets, as another example, have two poles that tend to attract or repel other magnetic poles. It is opposing properties such as these that tend to provide the balance and stability of most natural processes.

One fundamental aspect of charge carrying particles like the proton and electron is that opposite charges attract and like charges repel each other. This means that protons and electrons tend to pair up and stay connected with each other. We don't witness electricity in most materials we see because they are electrically neutral; that is, the number of

protons and electrons is equal. The electrical charges cancel each other out.

In order to use the attraction force that exists between two opposite charges we first must work to separate them. When the neutral balance is changed, the resulting imbalance creates electricity. For instance, a household battery makes electricity through a chemical process that separates protons and electrons in a special type of fluid. The battery builds up electrons at one terminal, marked with a -, and protons at the other terminal, marked with a +.

Let's take a closer look at the battery to try and understand what is really happening.

Fundamental Terms

When the protons and electrons become separated and migrate to the two terminals of the battery, a *voltage* is created. Voltage is an electrical potential. This means that it provides, potentially, the ability to create electricity.

After the buildup of electrical potential at the two terminals of the battery, the next step is to connect up some kind of device that will utilize the generated electricity. When the device connects to the two terminals of the battery, the separated protons and electrons are given a path over which they can rejoin back as pairs. During this rejoining process, electrical charges move from one terminal of the battery to the other. This moving electrical charge is known as *current*.

In reality, the moving electrical charge we know as electricity is only the result of moving electrons. In most cases, protons tend to stay where they are; it's the electrons that flow and create electrical current. So when the device is connected to the battery, the electrons from the negative terminal flow into the device and towards the positive terminal of the battery to rejoin with the protons.

If the chemical separation process in the battery ceases, eventually all of the electrons would rejoin with all of the protons and there would be no more voltage at the battery's terminals. This means there would be no electrons available to rejoin with the protons, and thus no more electricity.

From the battery perspective, electricity generation is a simple process! But, before we continue on, let's look at some of the terminology surrounding these two fundamental electricity terms: current and voltage.

Current

Current is moving charge, typically electrons. And just as the amount of water flowing in a river can be measured, so can the amount of flowing electrons through a medium. To make this measurement, we simply pick a reference point and count the number of electrons that flow past that point over time.

The standard measure of electrical current is the Ampere, often referred to just as "amp". It is equal to 6.24e18 (that's 6 quintillion!) electrons flowing past a reference point in 1 second. The amp is named after André-Marie Ampère, a French physicist credited with the discovery of electromagnetism.

Many times the term amp is abbreviated as just a capital A. For example, instead of seeing "5 amps" it may be more common to see "5A". This is especially true when SI prefixes are used, such as writing 5mA instead of 5 milliamps.

Finally, the terminology of current is often abbreviated with the letter I (probably because the letter C had already been used as an abbreviation for charge). Electrical schematics that need to show the presence of current in a portion of a circuit will often use the letter I as a symbol for current.

Voltage

Voltage is defined as the difference in electrical potential between two points in an electrical circuit. It is a measure of the electrical energy difference that would cause a current to flow between those two points. Sometimes voltage is referred to as the *electro-motive force*, since it loosely can be thought of as the force that pushes electrons through a circuit.

In reality, voltage is the result of an *electric field*, which is the force field that exists around electric charges causing them to attract or repel other charges, thus exerting forces on these other charges. While the actual study of electric fields is a bit beyond the topics of this book, just remember that they are the result of the interaction between charged particles.

Voltage is measured in terms of Volts, named after Alessandro Volta who first invented the Voltaic pile (the first modern battery). It is often abbreviated as an uppercase V.

2.2 Conductors and Insulators

Electrical current can travel through just about any material. Every material has an electrical property known as *conductivity* that describes its relative ability to conduct electrical current. Copper has a large conductivity, meaning it conducts electrical current quite well. Glass has a low conductivity, meaning it does not allow electrical current to flow through it very easily.

Materials with a high conductivity are known simply as *conductors*. Materials with a low conductivity are known as *insulators*, because they tend to block the flow of current.

While conductivity is a material property, the overall geometry of the material is also important in determining its current carrying capabilities. The combination of the material's conductivity and its shape and size is known as *conductance*. However, in the world of electricity, conductance is not an often used term. Its reciprocal, *resistance* is used instead.

Resistance

If you hover your finger near the surface of the microprocessor in your computer you probably notice that it generates heat. This heat indicates that work is being done by the electrical current flowing through the processor. The generated heat comes from the resistance of the material due to the fact that it's opposing the flow of current.

Resistance provides a direct relationship between current and voltage. Remember, voltage is (roughly) the force that causes current flow. If you can generate a certain amount of voltage across a material, then a certain amount of current will flow. The relationship between the two is governed by the resistance of the material.

As an electrical property, resistance is measured in ohms, named after Georg Ohm, a German physicist. Ohms are typically abbreviated with an uppercase Greek Omega (Ω).

The relationship of current, voltage, and resistance is described by Ohm's Law in Figure 2.1, on the following page. In simple terms, Ohm's law says that voltage and current are directly related by a factor called resistance. The relationship is linear. This means that if you double the voltage across a material, for example, you likewise will double the current.

Volts	=	Current	*	Resistance

Current	=	Volts	/	Resistance

Figure 2.1: Ohm's Law

2.3 Understanding Current Flow

Let's take a quick recap of what we have learned:

- Electrical current is the flow of charge (usually electrons).
- Electrical current flows as the result of the force created by a voltage.
- The amount of electrical current that flows is based on the resistance of the material it's flowing through.

Current Loops

It's not necessarily obvious, but current flow happens in a loop. If we want current to flow through a piece of wire, we have to somehow come up with a voltage to cause that to happen. Once we do that, every electron that comes in one end of the wire means that one electron has to leave the other end. This electron has to have a place to go. The voltage source supplying electrons to make the electrical current also receives electrons back at the other side.

Voltage Sources

Basically, a voltage source is an electrical "pump" that cycles current. The implication of this is that a voltage source has two sides, a side that lets electrons leave and a side that recollects electrons. When we talk about a voltage created by a voltage source, the voltage is really just the electrical potential difference between the two sides of the source.

Electrical Power

All of this talk of voltage and current would be remiss if it didn't actually do anything useful for us. Whenever current flows through some medium, it transfers energy into that medium. In an earlier example we discussed the heat coming from a microprocessor. That heat stems from the current flowing through the processor.

Electrical energy can be converted into a number of forms, such as heat, light, or motion. In the case of the microprocessor, the generated heat is an undesired byproduct of the current flowing through it and requires external intervention to help dissipate the heat away from the processor so as not to cause damage. A desired conversion can be seen in a light bulb, which converts electrical energy into light.

Electrical power is simply a measure of the amount of work (that is, energy transfer) done by electrical current.

Electrical power is measured in watts, named after James Watt, a Scottish engineer who is credited with the start of the Industrial Revolution through design improvements to the steam engine. The watt is abbreviated as an uppercase W.

The DC electrical power law is shown in Figure 2.2, on the next page. Mathematically, electrical power is the product of the voltage across a material and the amount of current flowing into that material. For example, if a 9V battery creates 0.001A of current in a circuit, then overall it is creating 0.009W of power.

2.4 Making use of electricity

We've identified that some materials are better than others at carrying electricity. For fun, let's try a few experiments. In order to make some electricity, we're going to need a source of voltage. Since we're already familiar with the battery as a voltage source we'll use it for our experiments. For our purposes, we'll utilize a 9V battery.

How batteries work - in depth

Batteries create their output voltage through a chemical reaction. Most commonly this is a *galvanic reaction*. This happens when two different metals are put into an *electrolyte*, which is a special type of charged solution.

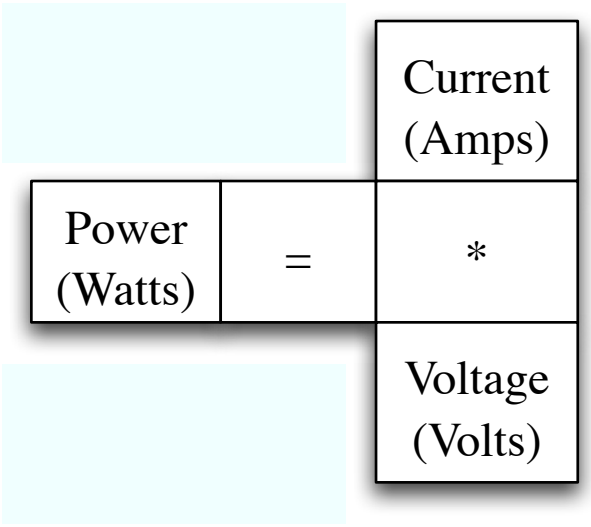


Figure 2.2: DC Electrical Power

The most common battery type uses electrodes made of zinc and copper. Both electrode types, when placed in the electrolyte solution, tend to lose electrons into the solution. The rate at which they lose electrons is different because they are different metals. If a wire is connected between the two electrodes, the excess electrons created by the material losing electrons faster are transferred over to the other metal by the wire.

This reaction cannot take place forever, because the charged particles that get transferred into the solution as a result of this process causes the corrosion of one of the electrodes and plating on the other electrode which reduces their ability to continue the reaction. This is what causes batteries to lose their ability to generate voltage over time.

Open Circuits

If we examine the battery in its normal state - that is, with nothing connected to the terminals, we would find that there is a voltage between the two terminals. This is highlighted in Figure 2.3, on the following page.

We can examine the battery using Ohm's Law. Remember, the battery's voltage creates current. In this case, the battery wants to push elec-

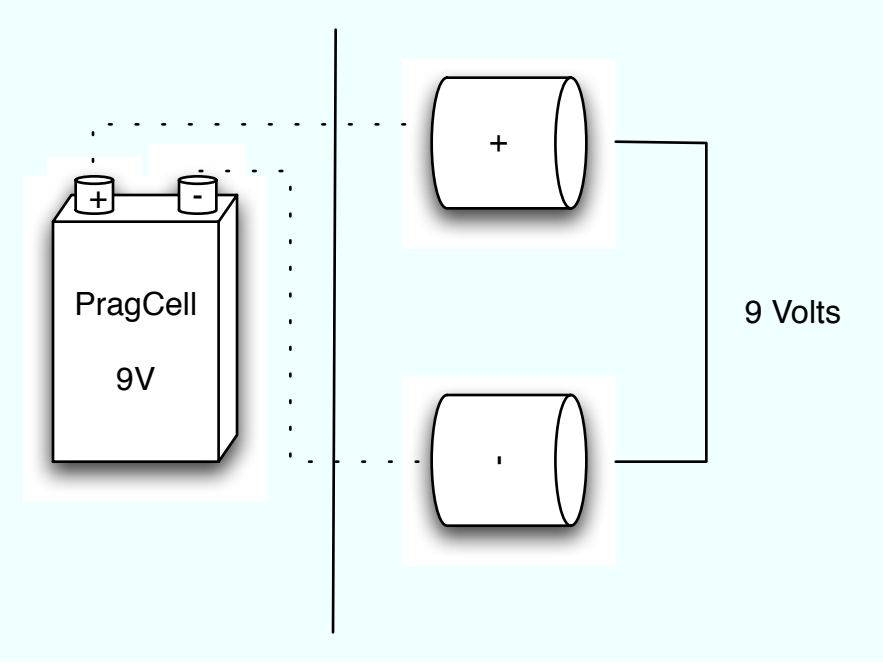


Figure 2.3: Voltage between two terminals of a Battery

trons out one terminal, through the air, and into the other terminal. How much current it is capable of moving in this fashion is based on the resistance of the air. A nominal value of the resistance of air is about 100 Megohms. Using Ohm's law, (it's back in Figure 2.1, on page 18), we see that this means that for the 9 volt battery only 0.00000009 amps, or 90 nanoamps, of current flows through the air. This is an extremely small amount, and is negligible for all practical purposes.

This condition — where there is a voltage but negligible current flow is called an *open circuit*. There's simply no place for current to flow. The resistance between the battery terminals is too high.

Since insulators like air and glass have such high resistances, we tend to think of their resistance as infinite. This means that the presence of a voltage across an insulator would cause no current flow. While there's no such thing as a perfect insulator (one with infinite resistance), for the purposes of this book we'll just consider all good insulators to be perfect.

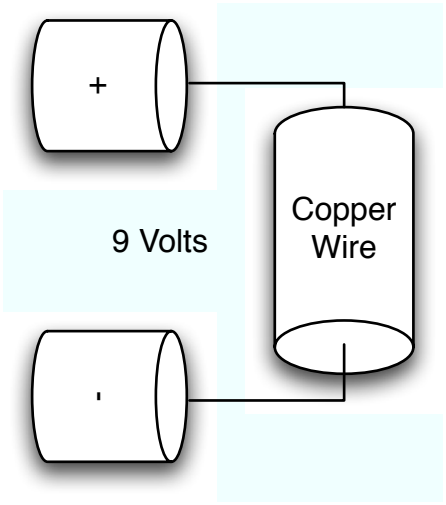


Figure 2.4: Battery Terminals with a Copper Wire

Short Circuits

Next, let's try putting a piece of copper wire between the battery terminals, like in Figure 2.4. The battery creates the exact same voltage as in the previous example, except this time it now has a piece of wire in which to pass current.

We can analyze the effect again using Ohm's Law. This small piece of copper wire has a resistance of around 0.001 Ohms. With a 9 volt battery, this means that we would have 9000 amps of current flowing through the piece of wire. This is an extremely large amount of current.

While the equation holds true, the logic isn't practical. It isn't possible for our little 9 volt battery to create 9000 amps. A typical 9 volt battery is only capable of producing around 15mA (0.015A) of current. If we try to force it to produce more, like we are with this piece of copper wire, the chemical reaction in the battery won't be able to keep up with the proton and electron separation needed to maintain 9 volts at the terminals. As a result, the voltage at the battery terminals will drop. We have created a *short circuit*

Because copper and other metals are such good conductors, and have very low resistances, we tend to like to think of them as perfect conductors, that is, conductors who have a resistance of 0. This isn't true

in all cases. Copper wire many miles in length (power lines, for example) does *not* have negligible resistance. But for the purposes of this book, we can consider good conductors, like copper wire, to be perfect. Because of this, we can ignore the resistance of wire within electrical circuits.

Actual Circuits

Finally, let's look at an in between case. Say we wanted to connect up something to the battery, such as a small light like in Figure 2.5, on the next page. In this case, we can ignore the effects of the wire we used to connect up the light—remember, it has negligible resistance. The light, however, does have a resistance—5000 Ohms. This means that, via Ohm's Law, our circuit is flowing 1.8mA of current (9V / 5000 Ohm = 1.8 mA). Furthermore, from the DC power law (Figure 2.2, on page 20) we can see that the light is receiving 9.8mW of power (9V * 1.8mA). This electrical power directly correlates into how bright the light shines.

On the right side of Figure 2.5, on the next page is the circuit model corresponding to the battery and light. DC voltage sources, such as batteries, are shown as a row of bars, alternating in size. A + sign highlights which end of the terminal is positive.

Anything in the circuit with non-negligible resistance, such as a light, is shown using a zigzag pattern. This pattern simply indicates to us that the object in the circuit has some form of resistance that we may need to take into account. The resistance value, in Ohms, is generally displayed next to the symbol.

Current Conventions

Electrons flow from more negative voltage to more positive voltage as shown in Figure 2.8, on page 26. However, a single electron doesn't directly travel between the two sides of the voltage source. Since all materials have electrons in them, these electrons also make up the current flow in the material. That is, when a voltage is presented across a material and current begins to flow, what happens is that one electron leaves the material and flows into the positive terminal of the voltage. This empty space, called a *hole*, is quickly filled in by another nearby electron. This process continues across the whole material until a hole exists close enough to the negative voltage terminal that a new electron can flow into the material.

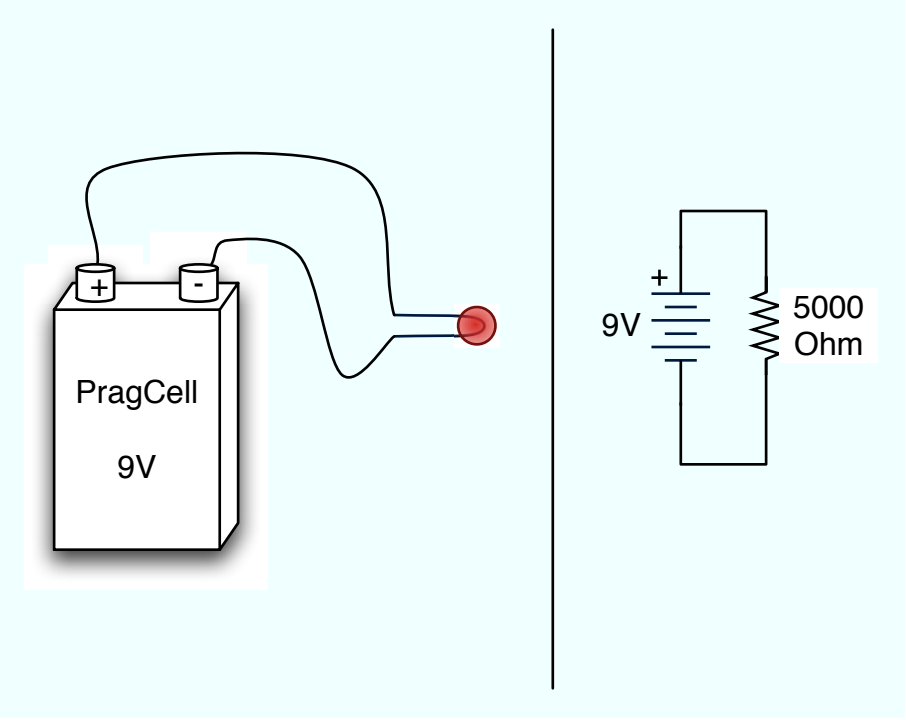
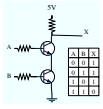


Figure 2.5: Battery Terminals with a Light

As electrons move in one direction, the holes they leave behind can be viewed as moving in the opposite direction as shown in Figure 2.9, on page 27.

Common electrical convention is to use hole current as the positive direction when discussing current flow. In general, hole current and electron current are really the same thing, just in opposite directions like in Figure 2.10, on page 27.

The reason for the convention of referring to hole current as the positive flow direction is to match current flow with the direction from higher to lower voltage. Since water flows from a higher pressure to a lower pressure, a natural analog is to have current flow from a higher voltage to a lower voltage. This technique also ensures some of the mathematical values calculate the correct way instead of having to remember to multiply them by -1.



The Buzz...

What's the difference between all these batteries?

See Figure 2.6 for an overview of common household battery voltages and current capabilities.

On an interesting note, all of the common household batteries with the exception of the 9V operate at the same voltage level (1.5V). The main difference between the batteries, however, is their current capacity (measured in milliamp-hours). If it wasn't for the physical limitations in making them fit, you could easily interchange batteries from one type to another and still have the same overall voltage level in your device. But the amount of current that the batteries could produce would be changed and as a result, the device may not have enough power to operate it properly.

Often, more than one battery is used in an application. The batteries can be chained together in two ways, either in series or in parallel. In series, the total voltage is increased while in parallel the total amount of current is increased. This is shown in Figure 2.7, on the next page.

Size	Voltage	Capacity (mAh)
9V	9	625
AAA	1.5	1250
AA	1.5	2850
C	1.5	8350
D	1.5	20500

Figure 2.6: Battery Capacity Table

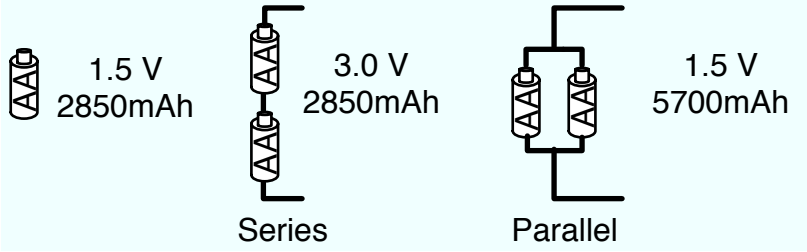


Figure 2.7: Batteries in series and in parallel

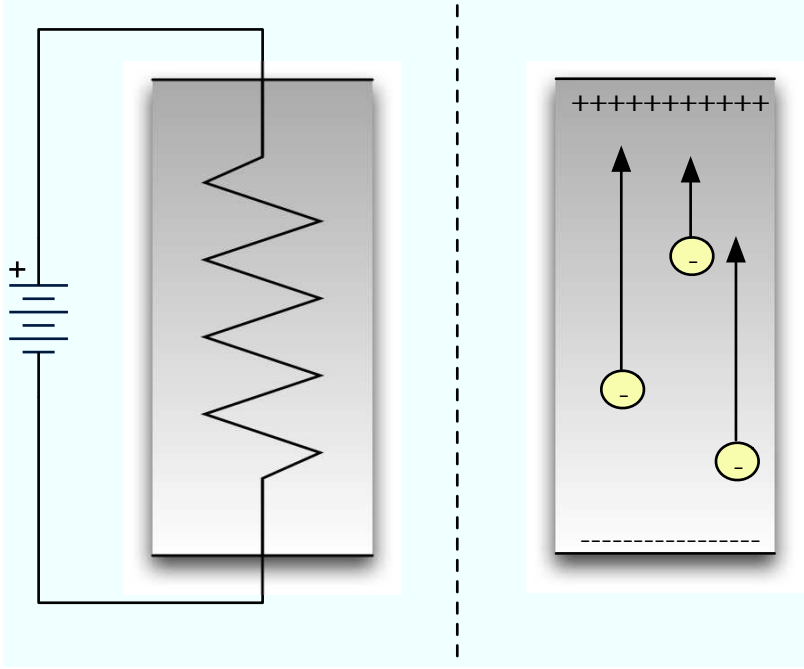


Figure 2.8: Electron Current Flow

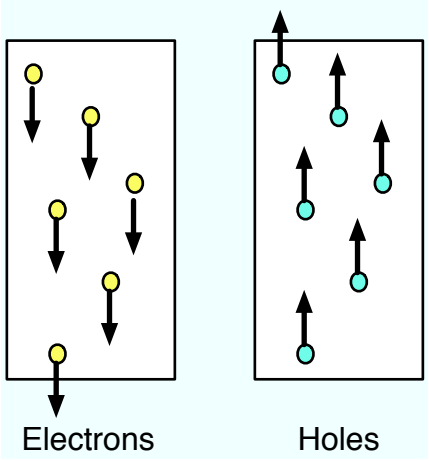


Figure 2.9: Electron and Hole Flow

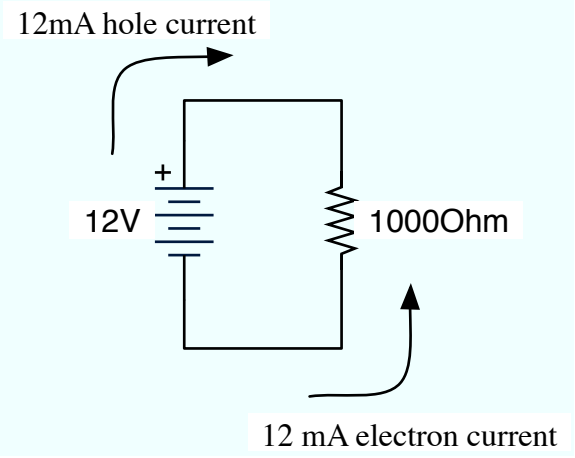


Figure 2.10: Hole And Electron Current Flow

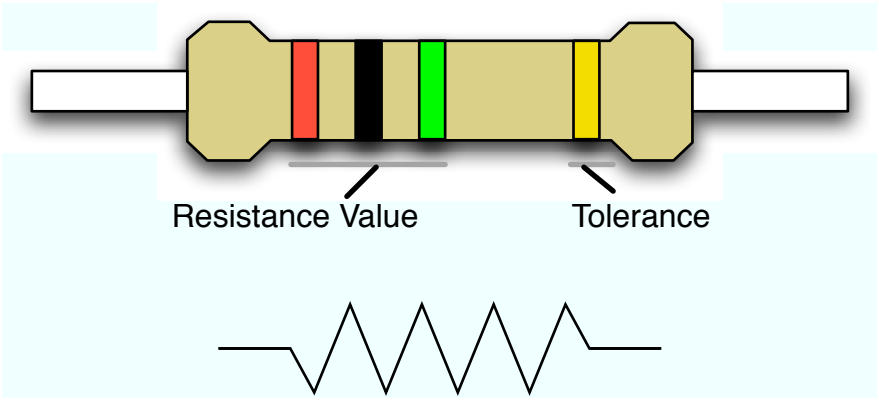


Figure 2.11: A resistor

This convention can be a little confusing, because we're not directly following the flow of electrons but instead following the flow of the holes left behind by the electrons. The important thing to remember is that electrical current, by normal convention, flows from positive voltage to negative voltage.

2.5 Electrical Components

There are three basic components used in the electronics world: the resistor, capacitor, and inductor.

Resistors

A resistor is simply a device that restricts the flow of current. Anything in a circuit that has resistance is a type of resistor. For example, the light in Figure 2.5, on page 24 is being utilized as a resistor.

A resistor is also an actual electrical component, as shown in Figure 2.11. Resistors are very common in electrical circuits as they provide a way to control voltages and currents. Resistors are used to divide voltages into smaller values or to limit the amount of current that can flow into a particular part of a circuit.

Resistors have colored stripes on them that represent their resistance value. They also have a colored stripe that represents a *tolerance* value.

Color	Value	Multiplier	Tolerance
Black	0	1Ω	
Brown	1	10Ω	±1%
Red	2	100Ω	±2%
Orange	3	1kΩ	
Yellow	4	10kΩ	
Green	5	100kΩ	±0.5%
Blue	6	1MΩ	±0.25%
Violet	7	10MΩ	±0.1%
Gray	8	100MΩ	±0.05%
White	9		
Gold			±5%
Silver			±10%

Figure 2.12: A resistor color code chart

Three or four colored stripes in close proximity designate the resistance value. The first two or three bands represent a numerical value with the last band representing a multiplier of that value. In the example figure, the resistor coloring of red-black-green signifies 2-0-5 which represents 20e5, or 2000000 ohms.

A separate lone band represents the tolerance. A gold colored tolerance band signifies a 5% tolerance level, meaning that the actual resistance value of this resistor is within 5% of the stated value, or between 1900000 and 2100000 ohms.

Capacitors

A *capacitor* is a device that can store electrical charge. Inside a capacitor are two metal plates, each connected to one of the capacitor's two terminals. Between these plates is a special insulator known as a *dielectric*. The model of a capacitor is shown in Figure 2.13, on page 31.

The use of the insulating dielectric makes it possible for charge to accumulate on the plates. For example, when a capacitor is connected to a battery, electrons redistribute themselves from the positive side of the capacitor to the negative side. This means that the negative side of the capacitor is negatively charged and the positive side of the capacitor is positively charged. This process is known as "charging the capacitor" and is shown in Figure 2.14, on the following page.

Eventually the capacitor becomes fully charged, like in Figure 2.15, on page 32. The electrical charge imbalance that has built up on the capacitor has created its own voltage, and the voltage of the battery no longer has the strength to overcome it. The battery cannot shuffle any more electrons around on the capacitor.

At this point we can disconnect the battery from the capacitor. But when we do, an interesting thing happens: the electrons on the capacitor plates stay put. The electrons on the negative plate want desperately to rejoin with their holes left on the positive plate, but the dielectric separating them makes that very difficult to do. There's no path to rejoin. Instead, the separated charge has created a voltage across the two terminals of the capacitor.

The charged capacitor is much like our battery in that it has a voltage across the two terminals and can act as a current source. However, the capacitor has no way to sustain this voltage once the electrons begin to flow and leave the negative terminal. The capacitor discharges rapidly, the voltage drops, and eventually the capacitor is completely discharged. Undisturbed, though, the capacitor ideally will store its charge forever. No capacitor is perfect, however, and over time some of the charge leaks out due to the *parasitic resistance* of the insulation materials used in the capacitors construction. The amount of time a capacitor stores its charge can range from very short (microseconds) to very long (many minutes).

The amount of charge a capacitor can hold is measured by its *capacitance*. The unit of capacitance is the *Farad*, abbreviated with a capital F. The Farad is named after Michael Faraday, a physicist who performed much of the initial research into electromagnetism.

Inductors

Another commonly used electrical component is an *inductor*. Like the capacitor, the inductor stores energy. Whereas the capacitor stored

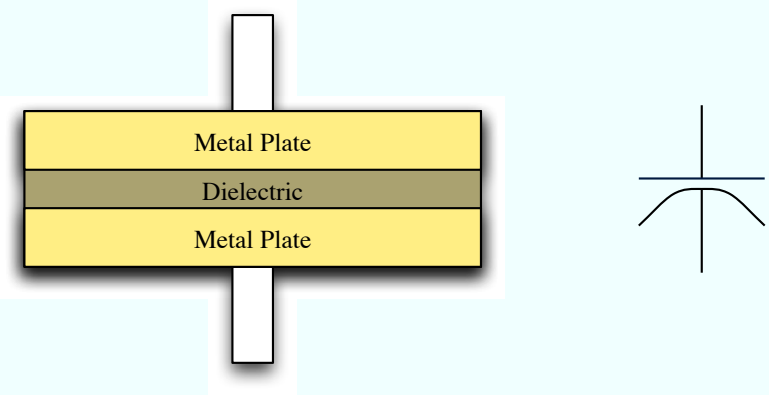


Figure 2.13: A capacitor

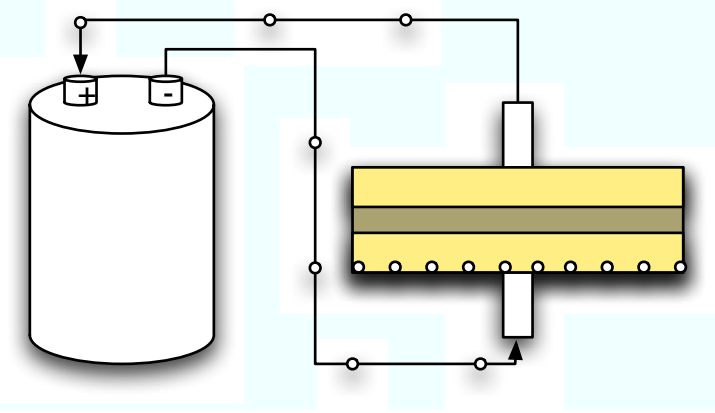


Figure 2.14: A capacitor charging

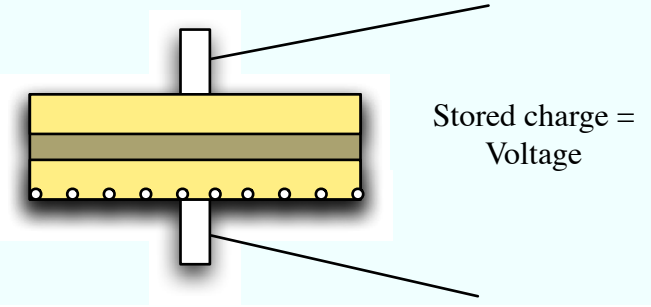


Figure 2.15: A charged capacitor

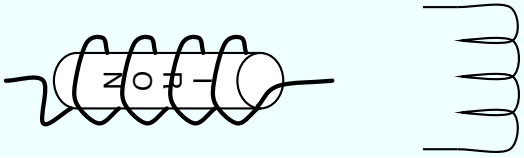


Figure 2.16: An inductor with an iron core

electrical charge, the inductor stores energy in a magnetic field (the same type of field created by a bar magnet).

An inductor is nothing more than a coiled piece of wire. When constant electrical current flows through the coil, it acts just like a piece of wire. However, when the current flowing through the coil changes over time, it creates a magnetic field inside of the coil. This magnetic field stores energy from the current. When the current in the wire goes away, the magnetic energy that had been stored turns back into current and attempts to continue to flow.

By placing a piece of iron in the inductor coil, we can create a *core* for the inductor. This piece of iron helps to guide the magnetic field and strengthen it, allowing for a larger *inductance*. The number of coils of wire in the inductor also correlate to the strength of the inductor.

The unit for inductance is the *Henry*, named after American scientist Joseph Henry, another research pioneer in the world of electromagnetism.

Mechanical Comparison

In the mechanical world, energy is utilized either in kinetic (moving) form or potential form. For example, a spring at rest has no energy. As you push the ends of a spring together, you are putting kinetic energy into the spring. Once you have the spring completely compressed, it now has stopped moving and the energy is now in its potential form. Once you release the spring, the potential energy converts back to kinetic energy and the spring expands. Over time, some of the energy is lost by friction. The spring may lose some of its energy via friction to the air, to your hands, and to anything else it comes into contact with.

The same is true in the electrical world. The resistor represents the friction component. The inductor and the capacitors represent the ability to take kinetic energy, in the form of electrical current, and store it as potential energy. In the capacitor, the potential energy is stored in an electric field. In the inductor, it's stored in a magnetic field. The stored potential energy can then later be released back into electrical energy.

Faith is like electricity. You can't see it, but you can see the light.

► Author Unknown

Chapter 3

Electrical Power

One of the most pervasive forms of electricity involved in our lives everyday is the electrical power distribution system.

3.1 Some History

Mention the history of electricity and the first thing that comes to most people's minds is a kite, a key, and a guy named Ben Franklin. Informally, though, it goes back much further than that. The Greeks were said to have discovered static electricity by rubbing fur on other materials. An ancient device known as the Baghdad Battery was a primitive battery thought to have been used for electroplating. In fact, scientists were predicting the effects of electricity as early as the 1600s.

Ben Franklin's kite flying experiment of 1752 is not known to be a fact, but he did correlate the relationship between lightning and electricity. Following this, scientists began to seriously study the effects of electricity and began to formulate their theories and terminologies. In 1786, Luigi Galvani, an Italian medical professor, discovered that a metal knife touching the leg of a dead frog caused violent twitching. He proposed that the frog's leg must contain electricity.

In 1792, Alessandro Volta disagreed. He proposed that the discovery was centered around dissimilar metal of the knife. When moisture came between them, electricity was created. This discovery led Volta to invent the first modern electric battery, a galvanic cell.

The new discovery was revolutionary. Up until Volta's discovery, all electricity discoveries had centered around static electricity and discharged sparks. However, Volta showed that this new kind of electric-

ity, which flowed like water, could be made to travel from one place to another in a controllable way.

Magnetic Motion

Following Volta's development of the battery, which was suitable for laboratory study, scientists began down the long road of electrical discovery. In 1831, Londoner Michael Faraday discovered the next major breakthorough. He found that when a magnet was moved inside of a coil of wire, electricity was produced. Where Volta had created an electricity source via a chemical reaction, Faraday created his through mechanical motion.

Faraday's experiment was relatively simple in nature. He made a coil by wrapping wire around a paper cylinder (a simple inductor). He connected the coil to a galvanometer and observed it when moving a magnet back and forth between the cylinder. When the magnet was stationary, no current was created in the wire and thus no voltage was observed at the ends of the wire, as seen in Figure 3.1, on the following page. However, when the magnet was moving Faraday observed an induced current through the wire as seen in Figure 3.2, on the next page. Faraday's experiment was termed *electromagnetic induction*, since a magnet was inducing the electricity on the wire.

Power on a Bigger Scale

For years, scientists continued to improve on the theories and designs of Volta and Faraday. Practical ways of using Faraday's electrical generation methods were sought. Initial designs involved moving a coil of wire around inside of a magnet, like in Figure 3.3, on page 37. The rotation of the coil of wire through the presence of the magnetic field creates electromagnetic induction, just like what was observed by Faraday.

In the 1860s, Charles Wheatstone and William Cooke improved upon the design by adding magnets to the coil of wire. Further improvements by other scientists finally made the generation of electrical power viable. In the mid 1870s, street lights in some major cities were being illuminated by electric arcs created from these electrical power generation machines.

The Ultimate Power Battle

Soon, Thomas Edison, a prolific inventor, began thinking about uses for electricity. His creation of a small incandescent lamp in 1879 which

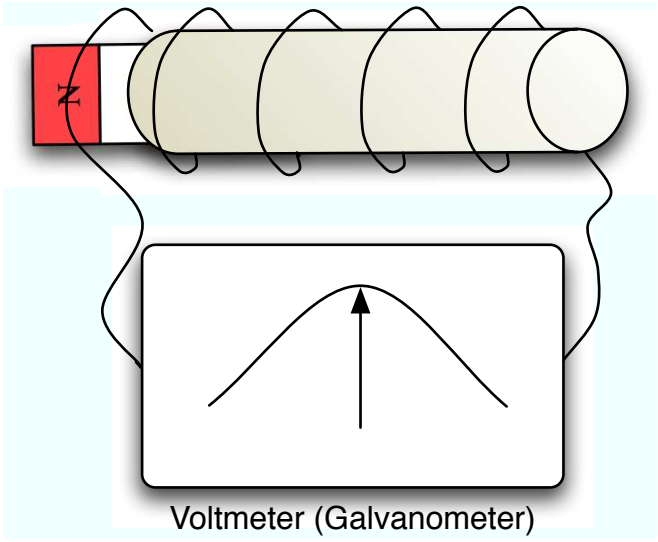


Figure 3.1: A stationary magnet inside of a coil of wire

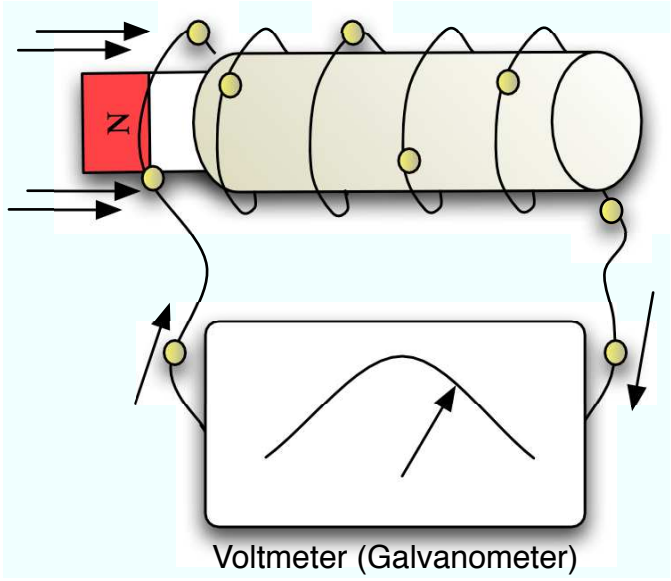


Figure 3.2: A moving magnet inside of a coil of wire

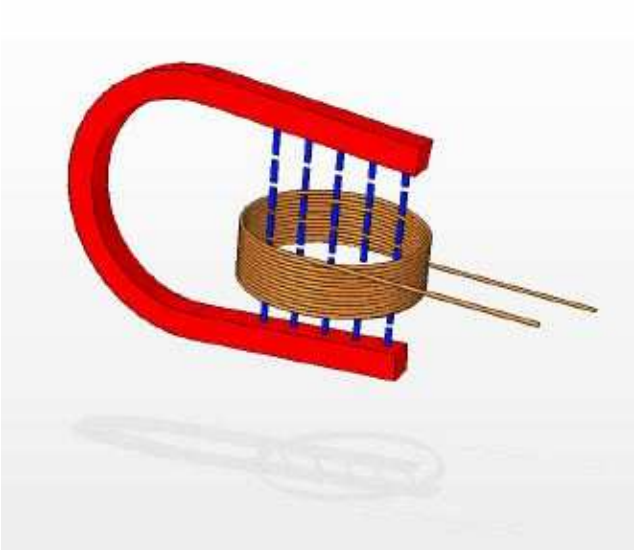


Figure 3.3: A horseshoe magnet with a perpendicular coil of wire

was suitable for indoor use led to his creation of a generation station in lower Manhattan, in New York City. By the mid 1880s, cities all over America yearned for their own electrical generation stations so they too could use Edison's incandescent light to illuminate the insides of their buildings.

Incandescent Light Bulb

The incandescent light bulb is very familiar to all of us. Inside of the glass bulb, an electric current is passed through a wire filament. This filament has an electrical resistance, meaning that the filament utilizes electrical power. In this case, the electrical power in the filament generates heat and causes the filament to glow white, generating light. The bulb's filament is surrounded by a vacuum or some inert gas to prevent the filament from oxidizing, reducing its usefulness. Early filaments were made from carbon, but modern light bulbs use tungsten filaments.

Incandescent light bulbs are notoriously energy inefficient; they waste about 98% of their power consumption to heat instead of light. The new trend in light bulb design seems to be moving to compact fluorescent designs which are more energy efficient, requiring only about 25% of

the energy as a similar incandescent bulb to generate the same amount of light.

Edison vs. Tesla

Using Faraday's principles of electromagnetic induction, Edison created a generator capable of producing DC, or direct current. One of Edison's employees, Nikola Tesla, a Croatian born inventor, had been working on a generation machine of his own that produced what Tesla called AC, or alternating current. The story between these two inventors is long and arduous, but nevertheless with different ideas and methodologies for electrical power generation design they soon parted ways.

George Westinghouse, another prolific inventor, saw the potential for electricity and created his own company. He purchased the rights to Tesla's invention and soon took on Edison in an epic battle to decide which machine was better capable of producing electric power.

3.2 AC versus DC

We'll get back to Edison and Westinghouse in a moment, but first let's take a look at their two competing concepts.

Electro-mechanical power generation

Whether we're dealing with AC or DC, electrical power generation as the result of some mechanical motion is generally handled by two principal components. The first, known as the *field* exists simply to create a magnetic field that we can use to later create the current. In Faraday's experiments, the field was created by the use of moveable magnets. Today, depending on the type of motor, the field can be created by either permanent magnets (magnetic materials like iron) or electromagnets.

The other needed part is the *armature*. The armature carries the current that is being generated. Faraday's armature was a stationary coil of wire, though generators may make use of moving or rotating wire coils.

Next, we'll look at a simple way of using a permanent magnet field along with a rotating coil armature to make electrical power.

AC Power Generation

To create AC power, we can start with the idea proposed by Faraday: a moving magnet and coil of wire produce electric potential. Similarly, a moving coil of wire in a magnetic field also produces electric potential.

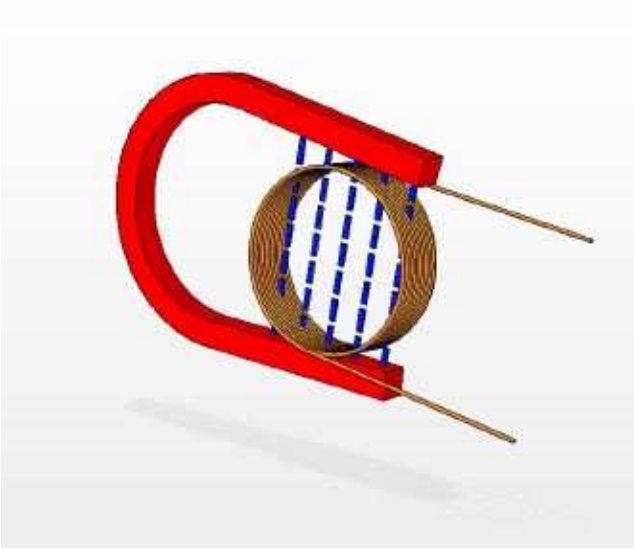


Figure 3.4: A horseshoe magnet with a parallel coil of wire

An example of this can be seen in Figure 3.3, on page 37. Note in this figure that the wire coil is oriented perpendicular to the magnet. In Figure 3.4, the coil of wire has been changed to be oriented parallel with the magnet.

In each figure, the blue lines represent the magnetic *flux* that is created by the permanent horseshoe magnet. In both figures the only thing that has changed is the orientation of the coil of wire with respect to the magnet.

If we were constantly to rotate this coil of wire, the induced voltage would look like Figure 3.5, on the next page. The voltage constantly cycles between some peak values, when the coil is perpendicular to the magnet. Along the way, when the coil is parallel to the magnet, the induced voltage is 0.

It's also very important to note that the coil must be rotating for this voltage to be induced. If at any time the rotation stops, even if the coil stays oriented perpendicular to the magnet, the induced voltage will drop to zero.

Finally, we need a way to get this induced voltage out of the ends of the

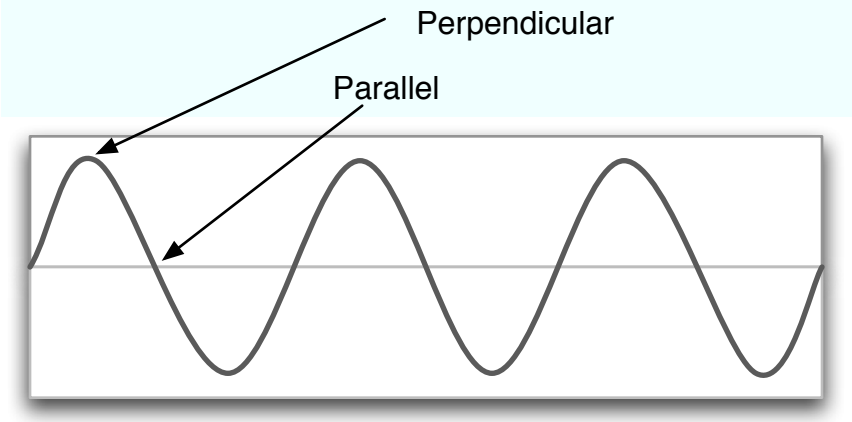


Figure 3.5: Induced Voltage on a Rotating Coil in a Magnet

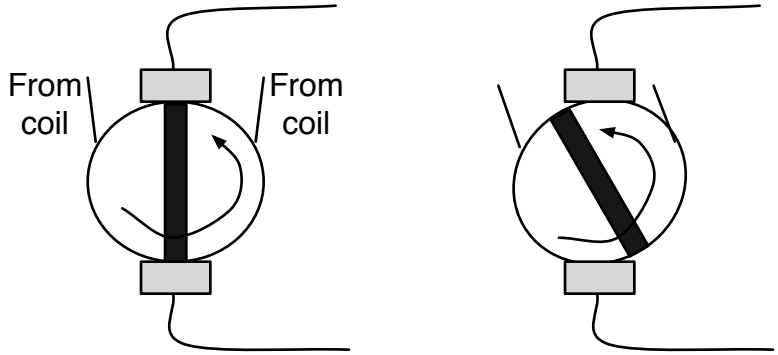
coil of wire and into something useful. The dilemma is that if the coil of wire is constantly rotating, it becomes difficult to connect the ends of the wire to anything practical since it would also have to rotate with the coil.

Slip Ring

The easiest fix for this is to use a device called a *slip ring* which is basically an electrical connector that can rotate. Internally, the slip ring is nothing more than a graphite brush that is in constant contact with a metal disk. As the disk turns, the brush is always in contact with it. This allows the current to constantly flow from the brush to the disk no matter if the disk is turning or not.

One downside to using slip rings is that their constant motion means there is some friction between the brushes and the metal rings. Over time, the brushes wear out and must be repaired or replaced. This means that there is some maintenance required for slip ring based devices.

Connecting the slip rings to the ends of the coil of wire allows the coil to continually rotate while allowing the wires coming out of the generator to remain stationary.



The two pieces of wire coming from the coil attach to each half of a round conductor. In the middle of the round conductor is an insulating piece that keeps the two halves separate. As the coil turns, the round conductor also turns. During its turning, it stays in contact with two fixed terminals.

Figure 3.6: A rotating commutator on a DC generator

DC Power Generation

DC power creation is somewhat similar to that of AC power. A coil of wire is rotated within the presence of a magnetic field. This in turn induces current in the wire. How that current is used, however, is different that with AC. Instead of using slip rings, like with AC, a DC generator has its wires attached to a *commutator*. The commutator is a type of rotating switch that allows the current flow to reverse direction in the wires. An example of a rotating commutator is shown in Figure 3.6.

What happens in the DC generator is the same as the AC generator for the first part of the cycle. As the commutator turns, the rotating coil of wire in the magnetic field induces a current in the wire and that current begins to flow. As the coil passes through the peak value, the current begins to come back down towards zero again. This is also exactly the same as AC. However, at the 180 degree point things change. With the DC commutator, at the 180 degree spot there is a small gap between the two sides of the commutator. As the commutator passes through this break, the current flow is zero. The inertia of the rotating part of the generator continues to spin and eventually the two metal pieces of the commutator are in contact again with the stationary pieces, but

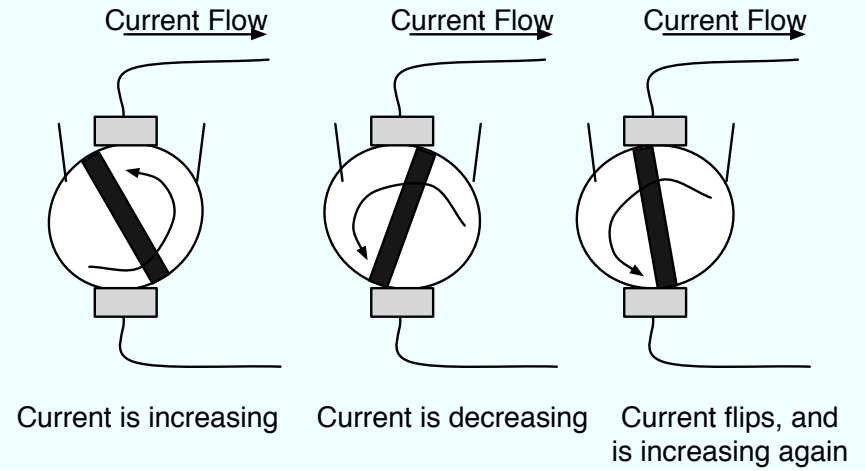


Figure 3.7: Current Induction in a DC generator

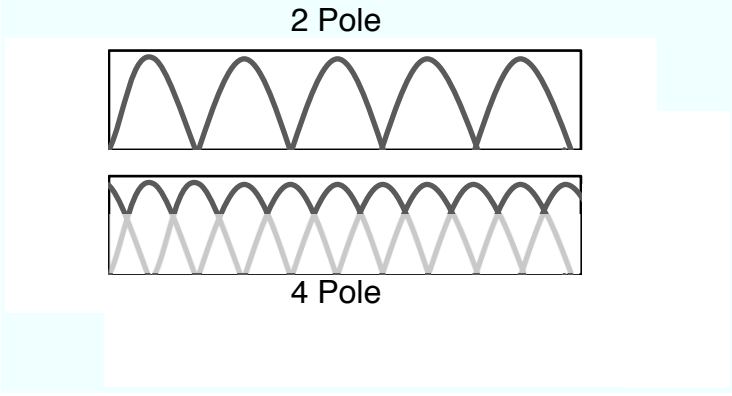


Figure 3.8: Induced DC Voltage on a Rotating Coil

each of the two parts are now touching the opposite pieces that they were touching earlier. However, as the rotor continues to spin through this section, current again flow in the direction it was flowing before.

The end result is that a DC generator flips the flow of current when it would normally be negative in an AC generator. This means that the current flow is always in the same direction, even though it may vary a

little bit. In the graphics, we've looked at a DC generator that has two portions, known as a 2 *pole* generator. However, it's possible to break up the generator into more poles. Adding a second coil of wire to our rotating coil field would create a 4 pole motor. The output of a 2 pole and 4 pole generator is shown in Figure 3.8, on the previous page.

The 2 pole generator output, while technically DC, fluctuates a bit. The 4 pole generator, however, has a much more stable output. A 6-pole or 8-pole motor would have an even better output yet. However, additional poles require a more complex and thus more expensive commutator.

Motors and Generators

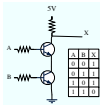
So far we've seen how generators work. Generators take mechanical motion and turn it into electricity. However, the opposite transformation, going from electricity into mechanical motion, is also commonly desired. This is what a motor does. Motors and generators are basically the same thing, except that power flow goes from electrical to mechanical in a motor and mechanical to electrical in a generator.

The same principles we've looked at for generators apply to motors as well, particularly for DC motors. However, many AC motors today instead use a fixed armature with a rotating set of magnets (electro or permanent). Current flowing into the fixed armature (also known as the *stator*, since it's stationary) creates a magnetic field that induces current in conductors in the inner rotating part (known as the *rotor*). The induced current in the rotor creates another magnetic field that counteracts to the original magnetic field in the stator. The two magnetic fields oppose each other, causing the rotor to turn.

3.3 And the winner is...

The battle of Edison and Westinghouse came down to politics and practicality. Early on, Edison's DC power reigned supreme. He had control of the distribution and held patents from which he was obtaining license revenue. He also had political clout and was a very outspoken opponent of AC power. DC power generation worked well to power incandescent lights, which were about the only things needing electrical power at the time.

It quickly became apparent that the downside to Edison's DC power was in the distribution. A low DC voltage was impractical to transmit across



The Buzz...

DC generator sparking

DC generators use commutators to get the DC power out of the generator. These commutators utilize brushes that transfer the power from the rotating part of the generator to the stationary part. The brushes consist of small "fingers" that contact the rotating shaft.

Sometimes one or more of the fingers may temporarily lose contact with the rotating shaft for a moment. In normal operation it's okay as the rest of the brush is still in contact with the shaft. But when the generated voltage becomes very high (near 1000V) it can cause a spark between a brush finger and the shaft when contact is lost. This sparking isn't good for the brushes or the DC motor, which means that it becomes expensive to utilize a DC motor at high voltages (near 1000V).

long distances because long power lines did not have neglibile resistance. High DC voltages were difficult to generate because of sparking that would occur in the armature of the generator. DC power was also very difficult to change to higher or lower voltages. Whatever voltage the generator created was what you had to work with.

The solution proposed by Edison was to have generation facilities near by the places where it would be utilized. Each needed voltage would be transferred on a separate wire. But having a generation station every few miles as well as having to run many different wires to each site turned out to be costly and impractical.

Westinghouse's AC power handled higher voltages much more readily. The higher AC voltage was easier to transmit over longer distances because the amount of electrical power loss in the power lines was minimal compared to the transmitted voltages.

However, the main advantage of this form of distribution was the easy ability to transform AC power from higher voltages to lower ones (using a *transformer*). This meant that high voltage AC power could be generated at a centralized station and distributed over long distances, being transformed down to lower voltage AC power at its destination. This

situation was very advantageous.

Power Line Loss

Earlier we discussed how the resistance of copper wire is negligible in most circuits. However, when we start talking power lines, which are thick cables strung over very long distances, the resistance of the wire becomes a factor. This means that as we're using the electrical power lines to send current from a generating station to customers, some of the power we generate is lost due to the resistance of the power lines.

The end customer that is being served by the power line has a certain power consumption need that we are trying to address. Not only do we have to create the power needed by the end customer but also the power that will be lost as heat in the transmission lines. Since customers don't directly pay for the power that's lost in the lines, we want to minimize that loss.

The power we are generating is a product of the voltage and current we are producing (see the Power Law graphic, Figure 2.2, on page 20). So, for a fixed power requirement, if we were to increase our generated voltage it would reduce the amount of current that would have to flow into the wire. Ohm's Law tells us that less current flowing through the resistive wire means a smaller voltage drop across the wire which means less power is lost in the wire. This means more power can be delivered to the final destination. This is highlighted in Figure 3.9, on the next page.

Edison fought hard against Westinghouse's AC power. His easiest target was the lethality of the high voltages that would be sent over power distribution lines. He demonstrated the devastating effects that Westinghouse's high voltage AC distribution would have on animals, including an elephant. And though he was against capital punishment, he created the first electric chair for New York to show how much deadlier AC was than DC.

Edison was correct in that AC power can be deadlier than DC at similar voltage levels because the frequencies of the voltages can interfere with the beating of the heart. However, at high voltages both AC and DC power are deadly. His demonstrations were more propaganda effort than real actual science.

Note that in both cases, the power plant is generating the same amount of power: 10kW. But the bottom power plant is doing so at a higher voltage, and has a result more power is being distributed to the end customer and not being lost in the power lines.

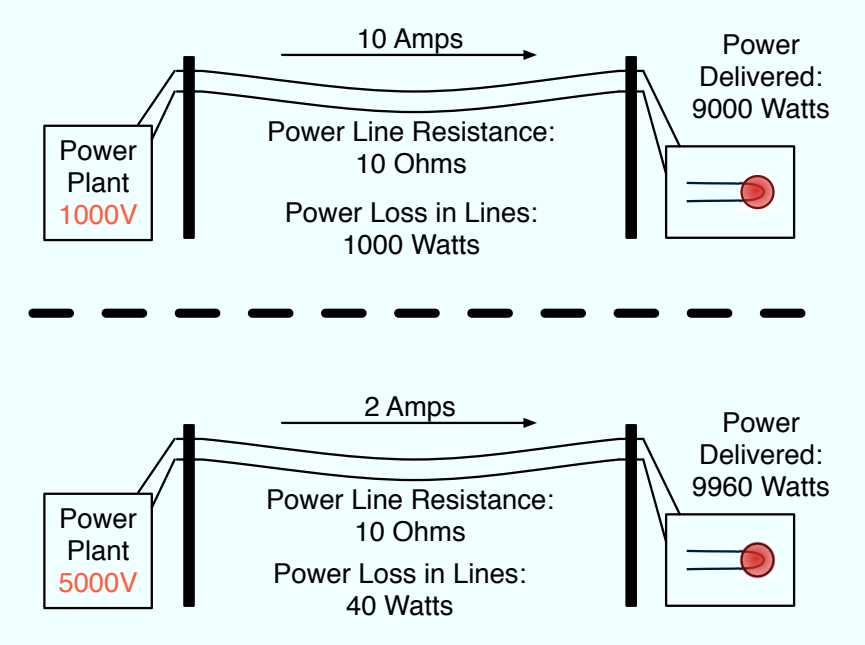


Figure 3.9: Power Line Loss

The Power of Water

It was becoming apparent that AC was superior to DC, but Edison was still a very prominent figure in the world of electricity and wouldn't let his DC power concepts and inventions die silently. The final tipping point came when Westinghouse and Tesla won a contract to create a power generation station at Niagara Falls. The system worked wonderfully with the power being distributed to Buffalo, New York at a distance of over 20 miles. This station proved the viability and safety of AC power generation and distribution.

Today, Niagara Falls is still one of the largest electrical power generation stations in the United States.

Because the advantages of AC over DC seemingly outweighed the disadvantages, it became the practical standard for power generation and

distribution—one we still use today.

3.4 AC Power Fundamentals

The terminology of *A*lternating *C*urrent and *D*irect *C*urrent is widely used, but a bit of a misnomer. When we mention AC, for example, we're referring to the alternating current that passes through a fixed resistance. This then implies that there must be a related alternating voltage that creates the alternating current. Similarly with DC, for a fixed resistance there is a steady voltage that must be supplying the current.

It's very common to refer to voltages as DC or AC. For example, the electrical power that enters your house does so by an AC voltage. The alternating voltage creates an alternating current, which is what the AC is really referring to.

Both AC and DC are methods for distributing power. However, as the power requirements of the load changes, both attempt to maintain their voltage values. For example, in your house, the AC voltage coming in is always at a (relatively) fixed 230V AC. Over time, the power requirements of your house change as electrical appliances turn on or off. This means that the resulting alternating current generated by this voltage may be higher or lower at any given point in time. If no electricity is being used, then no current flows into your house. But the 230V AC is still present.

AC Waveforms

While a DC voltage doesn't change, AC voltage is always alternating back and forth in a fixed way. The voltage is constantly changing from some positive value, down to 0, then down further into a negative value. It then comes back up and repeats the cycle over and over again. Since the graph of voltage over time looks somewhat like a wave of water, it's called a *waveform*.

Figure 3.10, on the following page is an example graph of an AC waveform. It shows how the voltage is constantly changing. The height of the voltage, from midpoint to peak, is called the *amplitude*. The time between two waves is known as the *period*.

Electricity in the U.S. and many other countries is generated at 60 *Hertz*, or waveform cycles per second. The period of these waves is 1/60th of a second. Thus, frequency and period are reciprocal values.

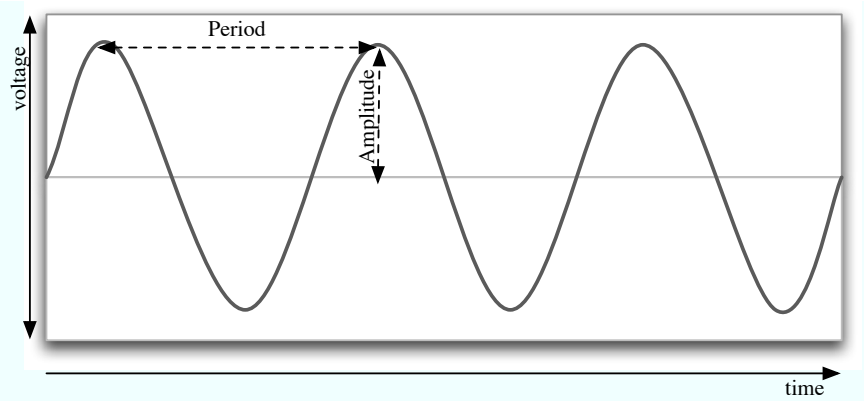


Figure 3.10: Waveform Terminology

$$\text{Period} = \frac{1}{\text{Frequency}} \qquad \text{Frequency} = \frac{1}{\text{Period}}$$

Figure 3.11: Waveform Equations

Another interesting concept in AC power is the *wavelength*, or the physical distance between waves. While the period is the time between successive peaks in the wave, the wavelength is the distance separating two such peaks. In order to figure this out, we have to know how fast the wave is moving.

We can use the equation in Figure 3.12, on the next page to figure out the wavelength of a 60 Hertz AC power wave. If we assume that waves in copper travel at the speed of light (note, this is an approximation), then the wavelength comes out to be 5,000,000 meters, or about 3100 miles. What this means is that if you had a piece of wire 3100 miles long and used it for 60 Hertz electrical power distribution, every time the waveform you were sending out was at its peak, at the other end of

$$\text{Wavelength} = \frac{\text{Velocity}}{\text{Frequency}}$$

wavelength of light in a vacuum

$$\text{Wavelength} = \frac{300{,}000{,}000 \text{ meters / sec}}{\text{Frequency}}$$

$$\text{Wavelength} = \frac{333 \text{ meters / sec}}{\text{Frequency}}$$

wavelength of sound in air

Figure 3.12: Wavelength Equation

the wire the previous peak would just be arriving. 3100 miles is a very long wavelength.

Why 60 Hertz?

There is no fundamental physical law that states that AC power must be generated at 60 Hertz. In fact, many parts of the world use 50 Hertz as the AC generation frequency. It's possible to generate any desired frequency of AC waveform with a properly designed generator.

Today's use of 60 Hertz for AC power generation stems from an initial design decision by Tesla. The general consensus on Tesla's decision is that it is the lowest frequency that would not cause a light to flicker visibly. Since AC power is rapidly switching directions, there are brief instants where no current flows into the light. This causes the light to flicker. Tesla noted that at 60 Hertz and above, the human eye could not discern the flicking effect anymore.

3.5 AC Power Distribution

Industrialization in the early 1900s drove the rapid expansion of electrical transmission lines to connect power generation plants with end

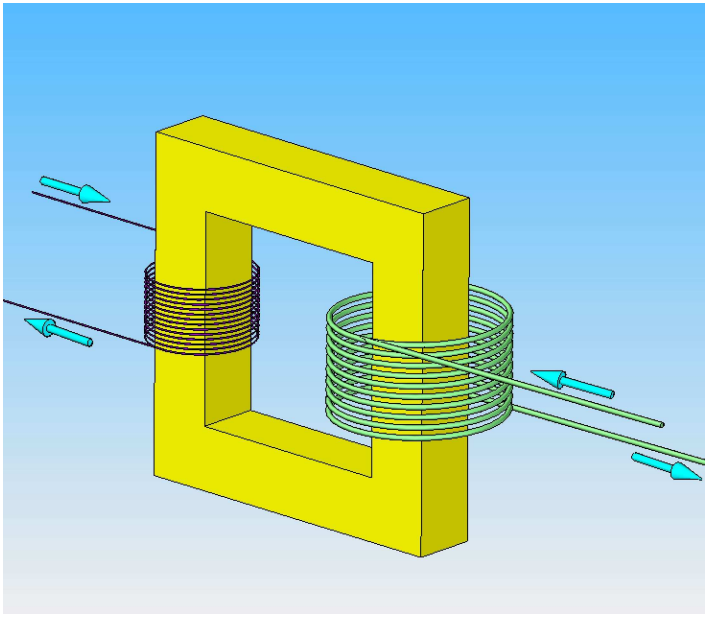


Figure 3.13: 3D model of a transformer

stations. In order to achieve this, generation plants would create electrical power at very high voltages to transfer it more efficiently over the lines. These high voltages, however, were not suitable for end use and thus had to be lowered before reaching their destinations. This was accomplished by using *transformers*.

More than meets the eye

An electrical transformer is a device that transforms an AC waveform from one amplitude to another. Transformers that create larger voltages from smaller ones are known as *step-up* transformers. Their counterparts that work in the opposite direction are *step-down* transformers.

At its heart, a transformer is a relatively simple device. As seen in Figure 3.13, it's simply a ring shaped piece of metal, usually iron, with two loops of wire wound around each side. As AC is applied to the *primary* side of the transformer, the electrical energy transforms into magnetic energy in the exact same way as that of an inductor (see Section 2.5, *Inductors*, on page 30 to recall how the conversion process works). The inducted magnetic energy travels through the iron core of the trans-

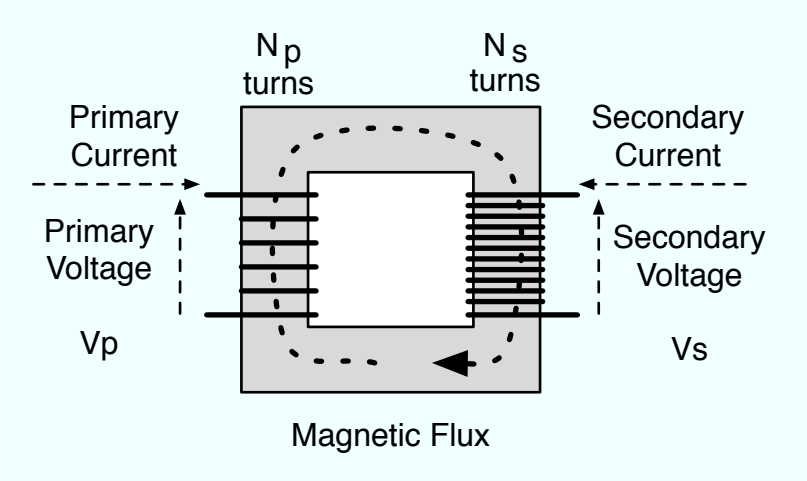


Figure 3.14: Descriptive Model of a Transformer

$$Vs = \frac{Ns}{Np} Vp$$

Figure 3.15: Transformer Equation

former and through the other loop of wire, known as the *secondary*. This magnetic energy converts back into electrical energy creating an electrical current through the secondary wire.

The resulting output current on the secondary side is directly related to the number of turns of wire that are wound across the transformer as described by the transformer equation, Figure 3.15. The equation shows that the output voltage amplitude is a ratio of the number of turns on the secondary side to the number of turns on the primary side.

What this means is that if the transformer has the same number of coils of wire on both the primary and secondary sides, then the output voltage would equal the input voltage. However, if the transformer has twice as many coils on the secondary side, then the output voltage would be twice the input voltage.

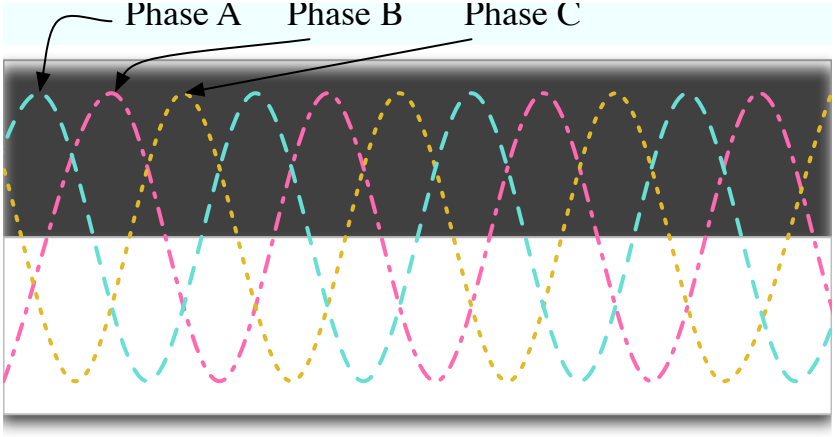


Figure 3.16: Three Phase Power

As you can imagine, the possibilities are endless. This very simple device provides an extremely easy way to step voltages down from the very high ones created by the power generation facility to the much lower ones needed by the consumer. But note that the conversion process from electrical, to magnetic, and back to electrical only works for alternating current. If the current instead was direct, it would not induce the magnetic field needed to transform the energy to the secondary side of the transformer.

Something for Nothing

While transformers can be used to step AC voltages up, they don't magically create new power. While there is a little bit of loss of power in the transformer itself, for the most part the power going into the primary side of the transformer is wholly delivered to the secondary side. This means that if a transformer steps up a voltage on the secondary side, the resulting amount of current the secondary side creates will be reduced. The electrical power on both sides is roughly equal.

Power in the United States

Electrical power generation begins at a power plant. The plant has to have some way of generating the electrical power which is generally accomplished by a rotating electrical AC generator. In order to create that rotation, power plants may make use of a waterfall, wind power,

diesel engines, or gas turbines. Most commonly, AC power is generated via a steam turbine. The steam can be created by burning coal, oil or natural gas. It can also be created by a nuclear reactor.

We've already seen what AC power looks like. However, power generation in the US is a *three-phase* system. What this means is that instead of a single waveform, the power company generates three different waveforms. Each waveform looks the same, but is slightly offset in time from the other two, like in Figure 3.16, on the previous page. The only thing special about three-phase power is that we're sending out three different AC waveforms (on three different wires), each slightly offset from each other in time.

In the home, generally only one of these phases is needed to supply the basic power needs. However, all three phases become important when dealing with industrial equipment.

The importance of 3-phase power

With both one and two phase power, there is a period of time at which the AC waveform is passing through zero (see Figure 3.17, on the following page as an example). If a piece of machinery is using single phase power, for example, as the voltage crosses through zero the instantaneous power consumption of the machine is also zero. As previously mentioned, this scenario is acceptable when working with light bulbs, as the light flicker is not noticeable. However, some industrial equipment is not as forgiving. In these pieces of equipment, certain electrical parts may exhibit bad effects when presented with electrical power that goes "off"periodically. The power delivered to these machines as a result of one phase AC is not smooth and steady, and it may cause a reduction in their expected lifetimes.

With three phases of AC, there is no time period in which the instantaneous power consumption of the connected machine is at zero. Any time one of the waveforms is passing through zero another one is reaching its peak. A machine that can utilize three phase power has a more even distribution of power, because one of the phases will always be at its peak when another one is at its zero crossing.

One very common use of 3-phase power is specially designed electric motors. Since the power being supplied to the motor can be drawn from three different sources that provide more even flow, the cost and complexity of these motors can be reduced.

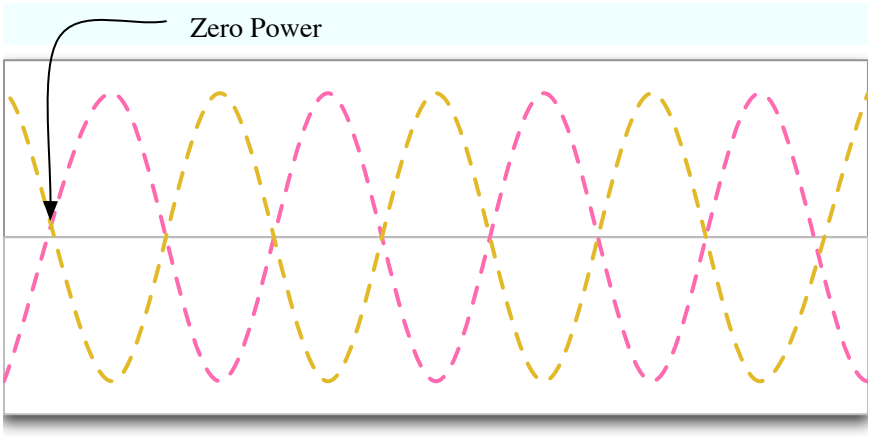


Figure 3.17: Two phase power waveform

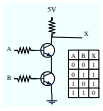
So why three phases and not four or more phases? It comes down to cost. The cost of adding yet another wire for the power company to run is significant. The overall cost was higher than any other efficiency gains it might bring to 4-phase motor design. Three phases became a natural settling point.

From the station to your house

The three phases of power leaving the electrical company's generators are in the thousands of volts. In order to be transmitted long distances efficiently, the power company uses step-up transformers to convert these voltages to hundreds of thousands of volts. These high voltage wires can be seen hanging from large towers.

The high voltage lines eventually reach sub-stations, where they are stepped back down for further distribution—typically to around 10,000 volts. From there, the voltage is distributed to its destinations along utility poles or buried cable. It may also be distributed to other transformers for further distribution, such as at the entrance to a subdivision.

Each commercial residence and business that uses electrical power has a transformer dedicated to its electrical service. This transformer takes the distribution voltage, around 7200 volts, and converts it to the volt-



The Buzz...

Why are high voltage lines run through the air?

You may be wondering why these lines are run through the air and not buried deep underground. This is because at these high voltages, the amount of insulation required to cover the wires to keep them from touching the ground would be financially impractical. Also, the wires would have to be buried very deep in order to keep someone from digging into them. And finally, if there is an electrical issue, such as a broken wire, it's much easier to fix a wire that is exposed instead of one you cannot easily see where the break might be.

That's not to say that power lines cannot be run underground. In many newer residential subdivisions they in fact are. However, these are lower voltage lines and not the tens of thousands of volts found on long distance transmission lines.

age utilized by the consumer. Most homes use 240VAC (that's 240 Volts, AC) and many businesses, particularly heavy industry, use 480VAC.

A home service transformer is actually designed to deliver both 240VAC and 120VAC through a special transformer *tap* arrangement as shown in Figure 3.18, on the next page. In this figure, the secondary side of the transformer has an extra wire that is connected halfway into the secondary windings. To utilize 240VAC, the two ends of the secondary side of the transformer are used as normal. To get 120VAC, the center wire is used along with one from either end of the transformer.

Most consumer electronics operate off 120VAC, though heavier power consumption machines like refrigerators and clothes dryers may need 240VAC to operate. These higher voltage appliances typically are connected to differently shaped receptacles so there's no confusion as to which voltage is present at which receptacle.

3.6 What is Ground?

The word ground gets thrown around a lot in electronics, and it seems to cause quite a bit of confusion. Historically, ground represents the physical earth. As electrical systems were being developed, it was found

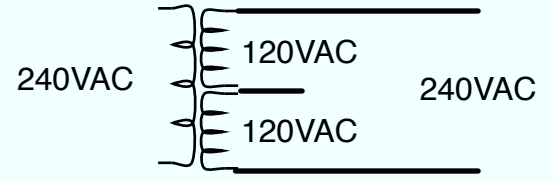


Figure 3.18: A home transformer with a center tap

that the earth is an electrically neutral body. Since voltage is measured as the potential difference between two points, having a common frame of reference is important. The earth provides a nice reference point when discussing various electrical circuits.

Thus, a ground represents the idea of using the earth as a common point for voltage measurements. The terminology has mutated a little bit over the years, though. Now, a ground can be a more generic term for a common frame of reference for voltage measurements within a circuit, even though it may not be related to the earth itself.

Signal Grounds

When we talk about voltage, which is the potential difference created by an electrical field, this difference has to be between two things. If this voltage is created by a battery, for example, the battery has to have two terminals. The potential difference between these two terminals is what creates the electricity. When wire is connected to these terminals and electrical components are activated at the ends of the wire, current flows out of one terminal and through the components back into the other terminal. The loop has to be complete in order for the current to flow.

If we had multiple batteries, we would need multiple sets of wires for each. Or would we? It turns out, we can commonize one of the wires, like in Figure 3.19, on the following page. It's possible for multiple energy sources, like batteries, to share the same return path. Using a common return path for multiple sources saves on parts cost and makes the circuit easier to understand.

Consider a case where there are two sources with no common return path, like in Figure 3.20, on the next page. What is the voltage difference between points A and B in this figure? 4V? 10V? The answer is:

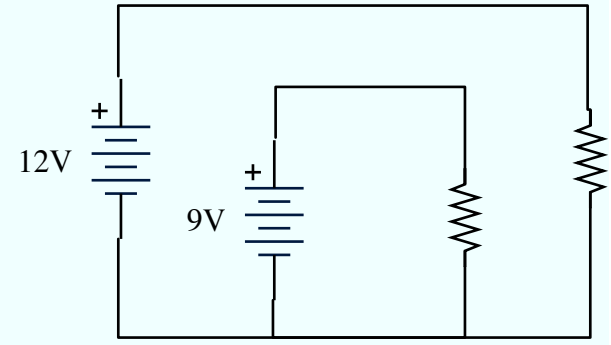


Figure 3.19: Two batteries sharing a common wire

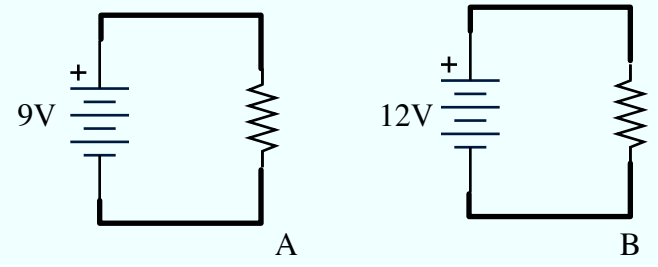


Figure 3.20: Two batteries and loads with no common wire

we don't know. There is nothing connected between these two circuits, so there's no relationship. It's possible that there is 1000V difference between these two circuits. The simple point is that we don't know. We have no common frame of reference.

So, when we refer to a signal ground, what we are usually referring to is the return path for small voltage sources, like small batteries or other small electrical signals. In many cases, it makes sense to connect one side of all of our voltage sources together to the common signal ground. This helps to commonize all of our voltage measurements to one single reference point.

The practicality of combining voltage source returns into a single common signal ground can be shown with a simple example.

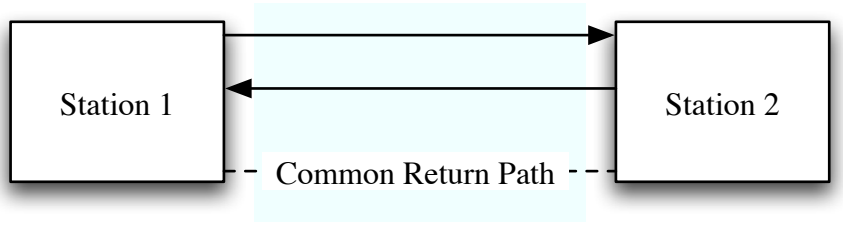


Figure 3.21: Two telegraph stations with one common return path

Figure 3.21 shows two telegraph stations connected via copper wire. The stations make use of two signal wires, so they can both transmit and receive messages at the same time. In order to generate the voltages that both can understand, a common return wire is also used (as noted by the dashed line). They could have used two common return wires—one for each signal. This extra expense is unnecessary. Instead, they are able to commonize their voltages against just one wire.

Earth Ground

An earth ground is simply an electrical connection to the earth (which is a good conductor) to provide a return path for current. Antennas and metallic towers, such as those used for cellular phones, are directly connected to the earth ground. This means that if they were to ever be hit by lightning, the resulting current would have a place to go. As well, lightning rods on houses and other buildings are generally earth grounded to copper rods driven many feet into the earth.

Your feet right now may be earth grounded. This means that if your body was to come into contact with electricity, it would have a place to go.

Electrical Power Ground

Since as humans we are frequently in contact with the earth and the human body can conduct electricity, there is some danger when we come into contact with other conductors. A good example of a conductor in your house is a copper water pipe. Imagine there was a voltage present on this pipe, for any reason (perhaps an electrical wire came loose somewhere and was touching it). When you touch the pipe, your body completes a path for the current to travel, since you also are probably earth grounded. As a result, you get a serious jolt of electricity.

Luckily that shouldn't happen. Electrical code regulations specify that copper pipes in your house must be connected to earth ground at some central point. This means that the current in any stray electrical wire that touches the pipe would have a place to go: through the pipe and into the ground (most likely tripping a breaker or blowing a fuse).

Because the copper pipes are connected to earth ground, they are at the same electrical potential as the ground. If you have one hand on a pipe and one foot on the ground, no current can flow into your body because the electrical potentials between your hand and foot are the same.

Ground Safety in the Home

All buildings (at least, those built within the past few decades) have a connection to earth's ground built in nearby, generally by an 8 foot piece of copper pipe driven into the earth. The ground plug at each of the electrical outlets in the home connects back to this earth ground. In turn, all appliances with exposed metallic parts connect to this ground plug. This ensures that there are no exposed metallic elements in the home that could potentially be a source of electrocution should a stray current carrying power line come into contact with them.

3.7 AC Power Safety

While the ground plug system provides safety against some potential electrical hazards, other hazards still exist. An errant short circuit, either inside of an appliance or by electrical connection to the human body, poses significant hazard. Under certain conditions, the home's 120VAC can be lethal, and for those of us who have experienced a 120VAC jolt, the experience is far from pleasant. Even at the appliance level, an electrical fault can result in permanent damage and sometimes cause a fire. Because of these hazards, a number of safety features are employed both along the power distribution system and in the home to help minimize accidents.

The first line of defense in the home is the *breaker box*, or the *fuse box* in older homes. All of the receptacles in the house tie back to a breaker (though multiple receptacles may share a breaker). If the amount of current going through the breaker (or fuse) exceeds the rating, the device stops the flow of current. When too much current passes through a fuse, a small piece of metal inside becomes hot and physi-

cally burns up, causing an open circuit in which no more current can flow. In the breaker's case, a switch on the front trips. The switch simply must be reset before it can be used again.

The main advantage to a breaker over a fuse is that it can be reused after a fault, whereas a fuse must be completely replaced with a new fuse.

A standard home breaker is rated for 10 Amps. This is generally enough capacity for a small room in a house for lighting, and some small appliances such as a television and radio. Larger requirements may make use of a 20 Amp breaker, or even more.

Ground Fault Circuit Interrupters

Circuit breakers and fuses are great as a first line of defense from too much current flowing into an electrical circuit. However, current as small as 1 amp can be lethal to a human even for a very brief period of time. If you were to touch an electrical outlet and an additional 1 amp of current started to run through your body, this may not be enough current to trip the circuit breaker or blow the fuse. An example of this is shown in Figure 3.22, on the next page.

Because of the inherent dangers of providing electricity in places where a higher likelihood exists for possible electrocution, such as near water sources, electrical codes commonly mandate use of a *Ground Fault Circuit Interrupter*, or GFCI.

The GFCI works by sensing an imbalance between the amount of current that is flowing into one side of the receptacle and the amount of current flowing out the other side. In the ideal situation, this amount of current is equal. If an imbalance exists, then the difference in current must be flowing somewhere else. A GFCI outlet is sensitive to the milliamp level and is designed to trip very fast—within milliseconds. GFCIs presume that an imbalanced electrical current could be flowing through a person's body and into the ground. Their fast trip reaction keeps the fault current from triggering a fibrillation of the heart.

3.8 Taking Measurements

One very useful tool we have in our exploration of electronics is a *multimeter*. This device allows us to measure different aspects of electrical circuits. A small multimeter is illustrated in Figure 3.23, on page 62.

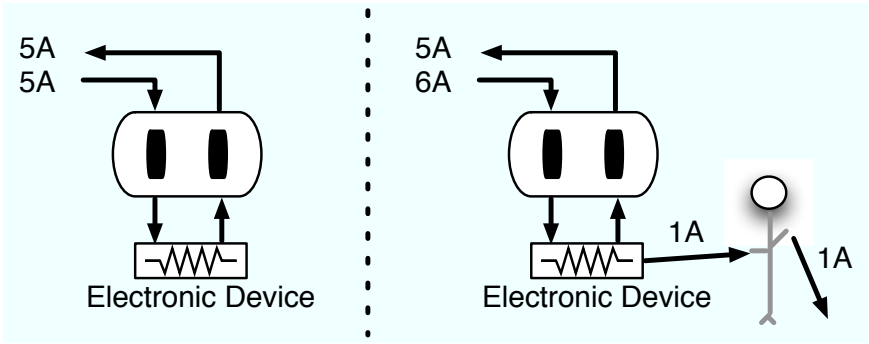


Figure 3.22: Normal Current Flow and Current Flow with a Human Fault

Two terminals allow us to connect probes so that we can take measurements. The red colored probe designates the positive side of the meter.

Modern multimeters can measure a myriad of electrical properties: current, voltage, resistance, capacitance, and more. This is accomplished by changing a switch on the front of the meter to tell it which property you want to measure. In the case of the meter in Figure 3.23, on the next page, we are measuring voltage.

Voltmeter

A multimeter in the voltage setting becomes a *voltmeter*. The meter has two terminals, positive and negative, which can be connected to the device of interest. The readout is then displayed on the screen, like in Figure 3.24, on the following page. In earlier times, a voltmeters were referred to as *galvanometers* since they measured the response to galvanic reactions.

Ammeter

Most multimeters are also capable of measuring current, in which case it comes an *ammeter*. In order for the ammeter to work, we must introduce it into the circuit so that the current we are interested in measuring must go *THROUGH* the meter. This is illustrated in Figure 3.25, on page 63.

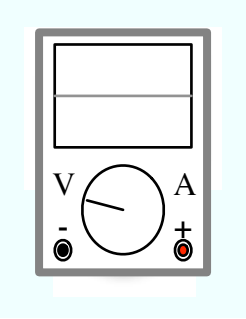


Figure 3.23: A multimeter

One has an Analog display. The other has a digital display. To take the measurement, we simply touch the probes to the location we're interested in measuring.

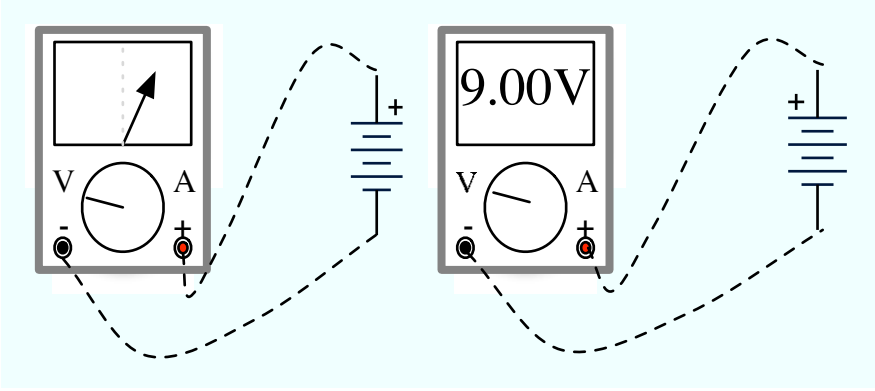


Figure 3.24: Two voltmeters.

Notice how the ammeter is used differently than the voltmeter. Here, we must make a complete loop and put the ammeter in that loop.

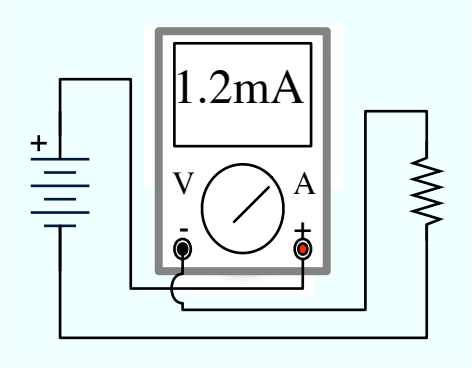


Figure 3.25: Example of an Ammeter.

It's important to note that when using an ammeter, the circuit actually must be broken and then the ammeter must be inserted directly into the middle of the break.

Other Meters

Many multimeters offer even more measurement capabilities:

- An *ohmmeter* measures the resistance of a component in a circuit.
- A *continuity* tester will sound a bell if the two leads of the meter are touching two continuous parts of a circuit. This is useful for checking for breaks in a circuit, for example.
- Many meters also offer settings to check diodes, transistors, and capacitors.

Clip-on Ammeters

The meters we've looked at so far made their measurements through direct contact with the circuits of interest. However, another meter exists that is able to take current measurements without any direct contact to the circuit. This device is known as a clip-on ammeter. It consists of a round clip that can be put around a wire that carries current. Because of the relationship between current flow and electric/magnetic fields, this device is able to measure current flow in a wire without any direct contact to the wire.

Figure 3.26: Picture of a clip-on ammeter

A picture of a clip-on ammeter is shown in Figure 3.26, on the following page.

The Oscilloscope

An *oscilloscope* is another valuable measurement instrument. With previous meters, we were able to look at what amounted to discrete values, such as the DC voltage in a circuit. The oscilloscope goes a bit further; it provides the ability to graph the measurement information over time.

In general, an oscilloscope is a voltage measurement device. However, special probes are available that can be used to measure current.

The abilities of an oscilloscope over a standard voltmeter are many:

- Observation of the voltage over time to see if it is changing or not.
- Calculation of the frequency of an alternating waveform.
- Visualization of electrical noise that may be coming into the circuit.

The downside of an oscilloscope is that it is generally larger, bulkier, and more expensive than a normal voltmeter. However, it is an invaluable tool for circuit and signal analysis.

Figure 3.27: Picture of an Oscilloscope

Chapter 4

Making Waves

We've already discussed a bit about electrical waves and some of their properties in Section 3.4, *AC Power Fundamentals*, on page 47. Since the concept of electrical waves is so important, we need to take a more in depth look.

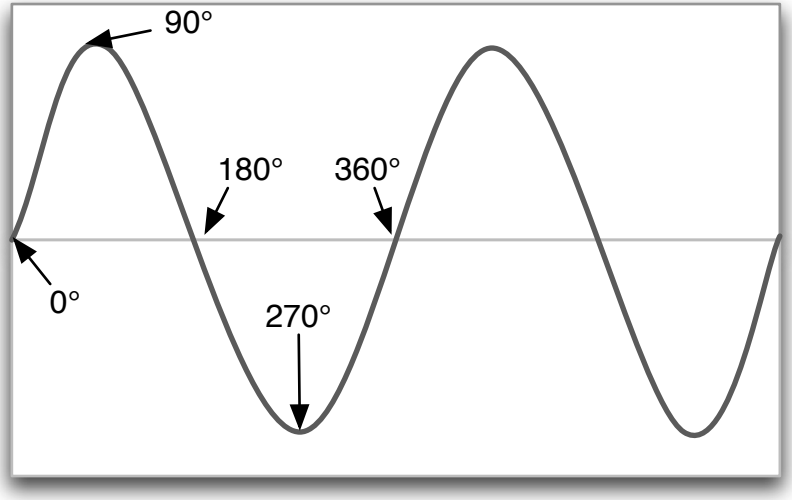
4.1 Electrical Waves

The voltage waveforms we've looked at so far, like the one in Figure 3.5, on page 40 are called *sine waves* (or sinusoidal waves) because the amount of induced voltage is proportional to the sine of the angle of rotation to the circular motion that is producing the voltage.

The sine is a trigonometric function shown in Figure 4.1, on the next page; it is the ratio of the opposite side to the hypotenuse of a right triangle. The sine function takes an input in angular measurement, either degrees or radians, and outputs a value between -1 and 1. The function repeats every 360 degrees (which is equivalent to 2π, or about 6.28, radians).

Engineers and mathematicians tend to prefer the use of radians over degrees when discussing trigonometric values. One radian is the angle created by a circular arc whose length is one radius. One degree, on the other hand, is just 1/360th of a circle. There's a more natural mathematical basis for the radian than the degree, and when solving some trigonometric Calculus equations, the calculations are easier when using radians instead of degrees. The downside is that thinking in terms of radians is slightly more complex because there aren't a whole number of them in a circle.

Angle θ		
Degrees	Radians	sin θ
0°	0	0
30°	π/6	0.5
45°	π/4	0.707
60°	π/3	0.866
90°	π/2	1.0
180°	π	0
270°	3π/2	-1.0
360°	2π	0



sine function

Figure 4.1: The sine function with table

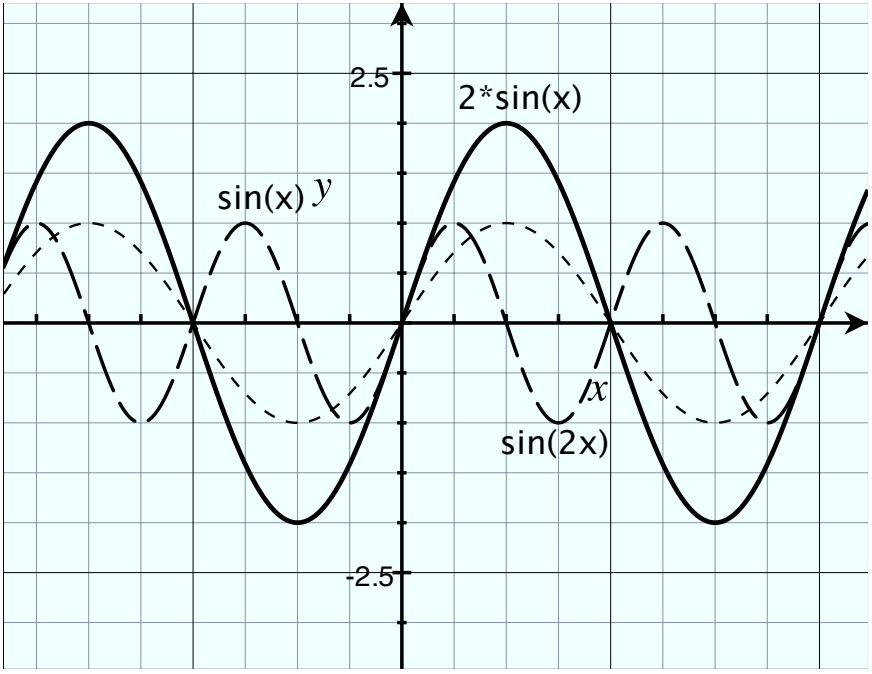


Figure 4.2: Three sine waves

Wave Characteristics

When we talk about sine waves, there are a few characteristics that come into play that help describe them: the repetition of the oscillation and the height of the oscillation. For example, consider the sine waves in Figure 4.2. On this plot are three different sine waves with three different functional representations. There are two different frequencies and two different heights.

The parameters of sine waves are shown in Figure 4.3, on the next page. The characterization of the items in this figure are:

- Amplitude - The height of the wave.
- Frequency - The repetitiveness of the wave.
- Time - The time, in seconds
- Phase Shift - The offset of the wave from zero. The wave in Figure 4.4, on the following page shows a phase shifted sine wave.

$$A * \sin(2\pi f * t + \phi)$$

Basic sine wave function

A - Amplitude
f - Frequency
t- Time
ϕ - Phase Shift

Figure 4.3: Basic Wave Equation

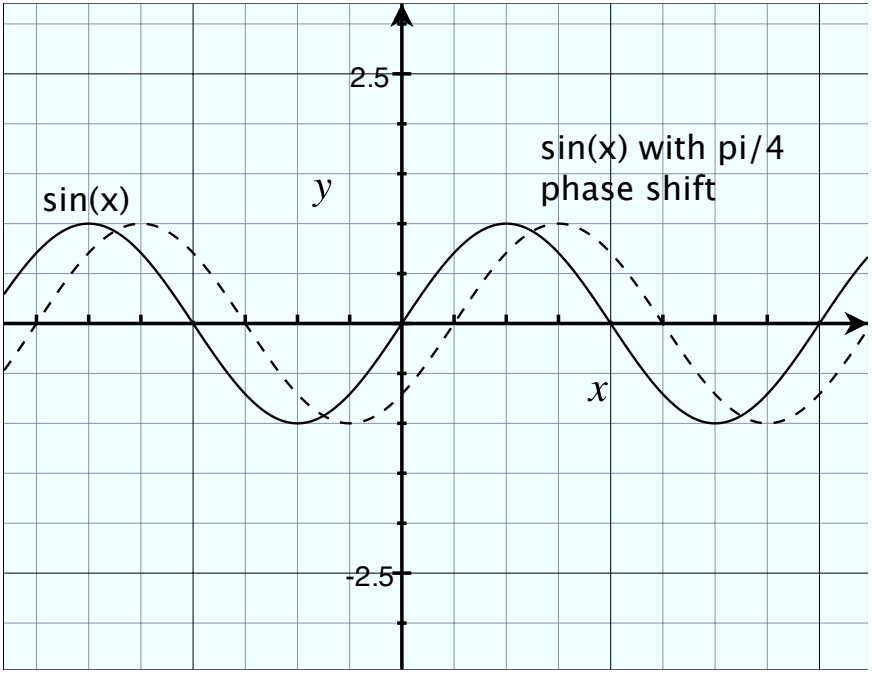


Figure 4.4: The sine function with a 45 degree ($\pi/4$ radians) phase shift

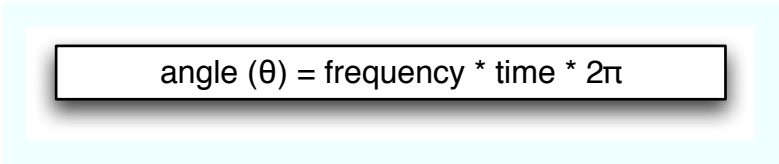


Figure 4.5: The sine angle to frequency relationship

Wave Frequency

As the generator turns, we can measure the number of cycles it is turning per second. This value is the *frequency* of the sine wave. If the generator completes 60 revolutions in 1 second, then the frequency of the generated voltage is 60 cycles per second, or 60 Hertz.

If we know the frequency, something interesting happens: we can predict the angle of the generator at any point in time. The frequency, in cycles per second, multiplied by the time, in seconds, gives us the total number of cycles that the generator has turned (as noted in Figure 4.5). And since each cycle is 2π radians, we know the total number of radians that the generator has turned. For example, if the generator is turning at 60 Hertz then in 1 second it will have turned exactly 60 cycles. 60 cycles is 120π radians.

Non-sine waves

There are infinitely many different waveform shapes. Two other common waveform types are square waves and triangle waves as shown in Figure 4.6, on the following page. There's nothing strange about these types of waves, they're just different types. For example, the square wave could be created by the action of someone turning a switch on and off, or by a clock signal. A triangle wave may be created by an electrical circuit charging and discharging.

Additive Waves

Commonly, more than one wave may be transmitted at a time. For example, when you press a key on the piano (middle A, for example) it creates a SOUND wave with a frequency of 440Hz, known as the *fundamental frequency*. But it also creates many more waves at higher frequencies that accompany the 440Hz wave. These extra frequencies are known as *harmonic frequencies*.

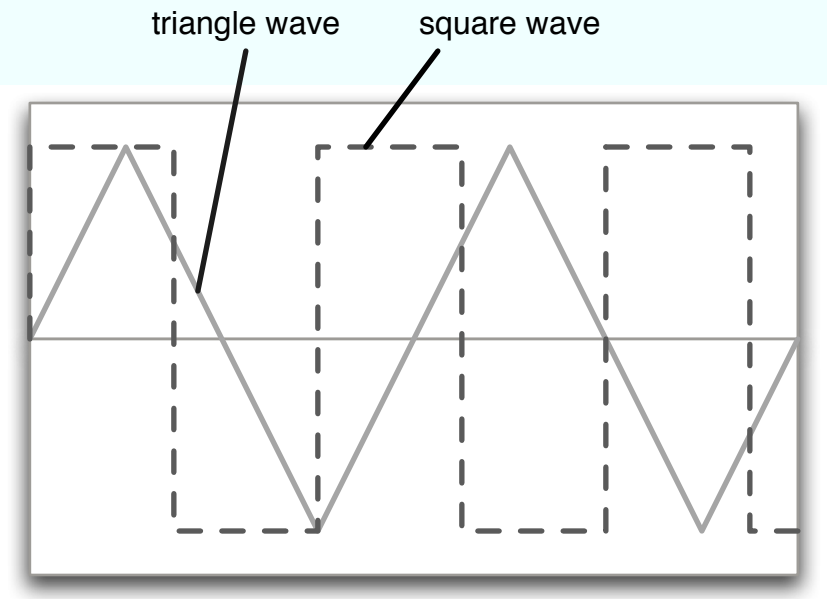


Figure 4.6: Square and triangle waves

It's these harmonic frequencies that allow you to distinguish the difference between a piano playing a middle A and a guitar playing the same note. The fundamental pitch is the same, but the sounds are different. The difference is in the harmonic waves.

The same concept is true for the voice. Your ears use this information to distinguish between different speakers.

When more than one frequency is present in a signal, the resulting waveform is just the additive result of the individual sine functions. For example, Figure 4.7, on the next page shows a waveform that is the additive result of the three sine waves shown in Figure 4.2, on page 68. Note that it doesn't look like a sine wave that we're used to. It repeats, it has a frequency, but it isn't sinusoidal (that is, it doesn't look like a sine wave). It's still a wave however; it's just composed of multiple individual sine waves added together.

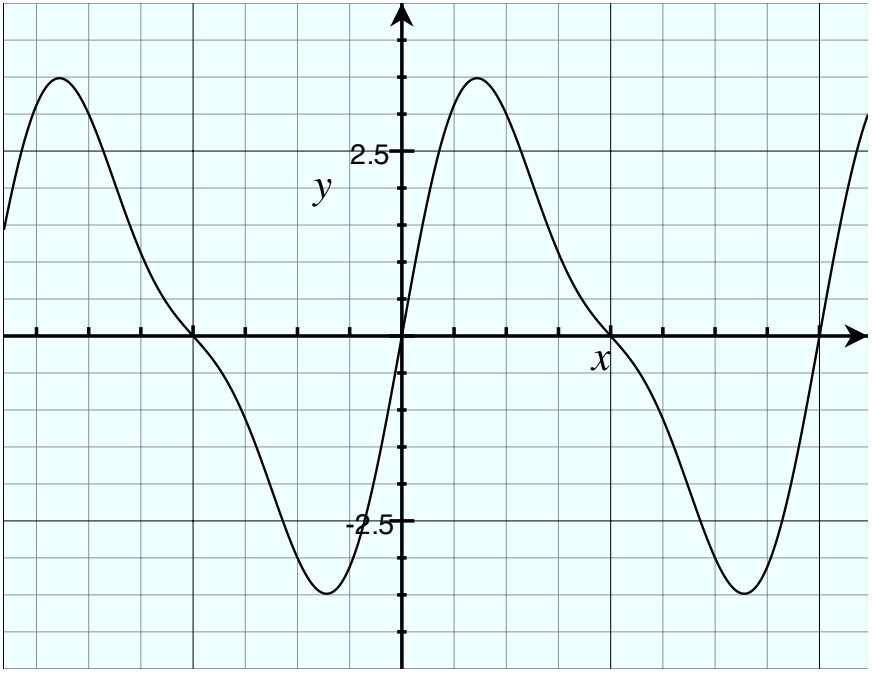


Figure 4.7: Additive Sine Waves

Fourier Series

In 1822, Joseph Fourier discovered an interesting property of waves. Any arbitrary waveform, of any shape, can be represented as an additive group of sine waves. His concept of the *Fourier Series* meant that any repeating pattern could be simplified into the simple addition of sine waves.

Let's look at this in practice. A classic example of this in action is a square wave. Broken down into its Fourier series, a square wave is just the addition of odd multiples of sine waves, mathematically shown in Figure 4.8, on the next page. This series is infinite—that is, there are an infinite number of sine waves that make up the square wave. But as the frequency of each additional sine wave gets larger, the amplitude becomes smaller.

We can demonstrate the creation of a square wave by adding these sine waves together. In Figure 4.9, on page 74, we attempt to recreate the square wave by adding up the first few sine waves of the Fourier

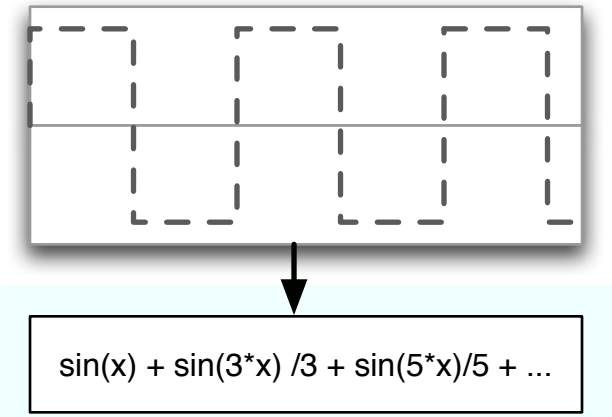


Figure 4.8: A square wave as a fourier series

series. Note that it doesn't make a perfect square wave, though as we add more sine waves corresponding to higher and higher harmonics the shape becomes more square.

Fourier's discovery shows that all arbitrary wave shapes can be created by the simple addition of one or more sine waves of different frequencies. This makes the sine wave the basic wave type of wave that all others can be composed of.

Frequency Response

One of the consequences of utilizing certain electronic components is that sometimes it's not possible to transmit all of those frequencies across a medium.

In some circuits, the addition of capacitors and inductors cause filtering of some high or low frequencies. This can be shown by the *frequency response* of a circuit. A frequency response is a correlation between an input signal and an output signal of a given electronic circuit. An example frequency response is shown on Figure 4.10, on page 75. The frequency response chart shows a graph over a wide range of frequencies. A line is drawn to show that at certain frequencies what happens to the output of the circuit. In some cases, the output becomes larger than the input (gain), and in other cases it becomes smaller (attenuation).

The graphs should overlap each other, but are shown offset so they are easier to view. The lower curve shows a Fourier series using just five frequencies and the top curve uses the first thirty frequencies.

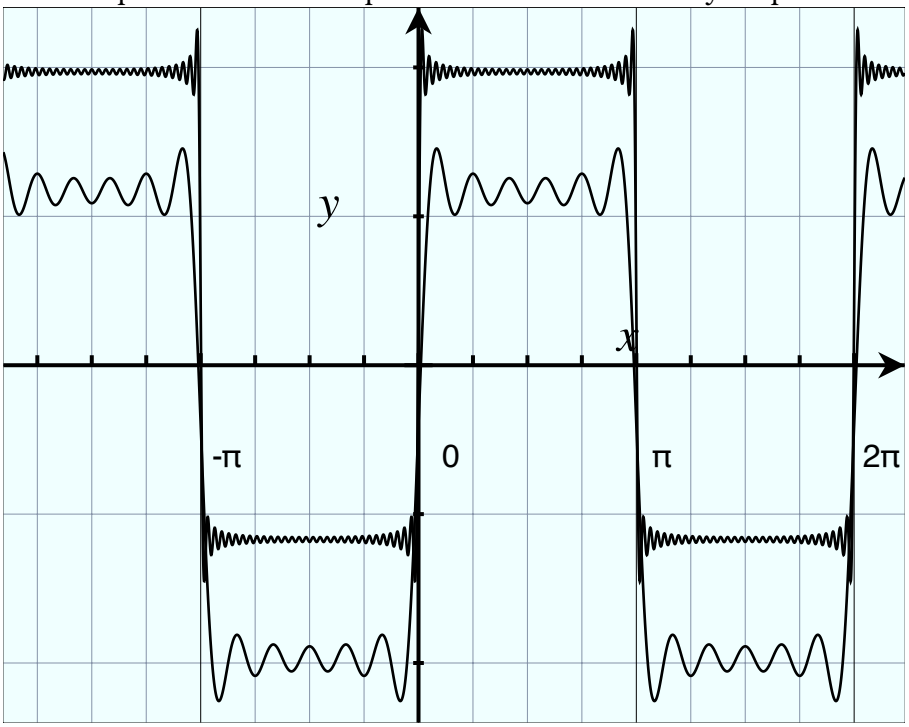


Figure 4.9: Building a square wave out of sine waves

Complex circuits have more interesting frequency responses. For example, a simple resistor/capacitor pair generates the frequency response shown in Figure 4.11, on page 76. The diagram shows that for frequencies less than 10kHz the output voltage looks the same as the input voltage. The horizontal line indicates that as the frequency increases beyond 10kHz the output becomes smaller. The output will eventually become so small that it is basically undetectable.

This is a very interesting concept. What it means is the output of electrical circuits is dependent on the frequency of the input signal. At some frequencies, the output looks just like the input. If you were to connect an oscilloscope to both the input and the output of these circuits, the displayed waveforms would be identical.

The lower graph shows a device whose output becomes smaller as the frequency increases.

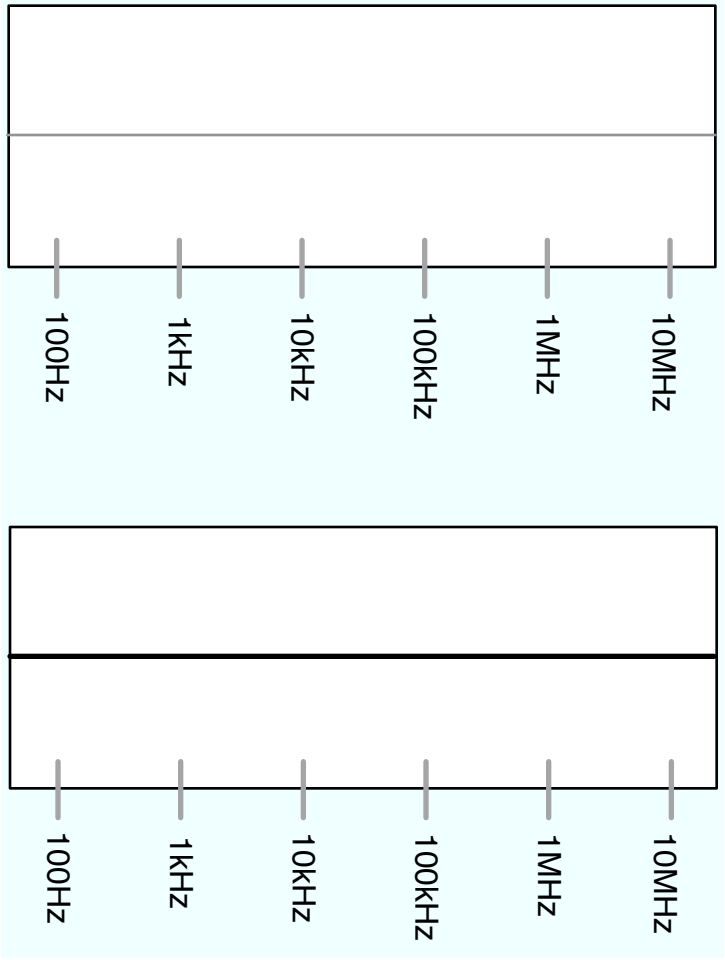


Figure 4.10: An example frequency response. Anything above the center line represents amplification. Anything below the center line represents attenuation.

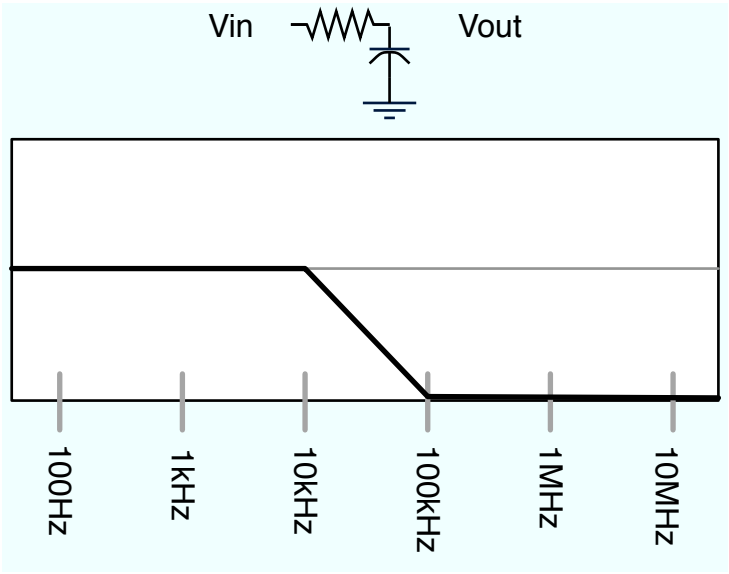


Figure 4.11: A frequency response of a resistor and capacitor pair.

But at other frequencies, the output may be smaller or even nonexistent. An oscilloscope connected to these circuits would show that the output does not look the same as the input. The circuit is changing the frequency information of input.

If we put in a signal that had more than one frequency, the output result may look very different. Some frequencies that were present in the input signal may not be present in the output signal.

What would happen if we tried to send a 1kHz square wave through the circuit in Figure 4.11? Remember, a square wave is simply made up of a group of sine waves of various frequencies added together. Some of those sine waves would be passed through the circuit unchanged. Others may not make it through the circuit. The result is that the output wave will not look like a square wave anymore.

Filtering

Circuits in which the output at certain frequencies may look different than the input are known as electronic *filters*, so named because they filter some frequencies from being seen at the output. Filters are grouped into a few different styles:

- Lowpass - Removes all frequencies above a certain level.
- Highpass - Removes all frequencies below a certain level.
- Bandpass - Removes out all frequencies except for a certain range.
- Notch - Removes out all frequencies within a certain range.

Furthermore, electronic filters are grouped into two categories: active and passive. Passive filters are built using discrete components like resistors, capacitors, and inductors. In this case, the output amplitude will never be larger than the input. Active filters are built using discrete components, but also include the use of amplifiers within the circuit. Because of the amplifier use, an active filter's output amplitude may actually be larger than the input.

Cutoff Frequencies

All filters have a range of frequencies over which they perform their work. For example, a lowpass filter diminishes frequencies above a certain value and lets through frequencies below a certain value. The frequency value where a filter starts to work is known as the *cutoff frequency*.

In the example Figure 4.11, on the preceding page the filter circuit is a passive lowpass filter and the cutoff frequency is 10kHz. As the input frequency increases beyond the cutoff frequency, the output signal gets smaller and smaller. Theoretically, the output signal always continues to become smaller and smaller as the input frequency increases, but at some point it becomes unmeasurable and as such we can deem it to be nonexistent.

However, note the section of the figure between 10kHz and 100kHz in which the output is getting smaller. During this range, the output signal does still exist though it is somewhat attenuated. This distinction is important: just because the input frequency is above the cutoff frequency does not mean the output signal will be nonexistant. It merely means it will be somewhat reduced. As we move further from the cutoff frequency, more attenuation takes place.

For the example circuit, the output frequency is attenuated by a factor of 10 as the frequency goes up by a factor of 10. That is, if our cutoff frequency is 10kHz then the output will be 1/10th its original amplitude at 100kHz and 1/100th the original amplitude at 1MHz.

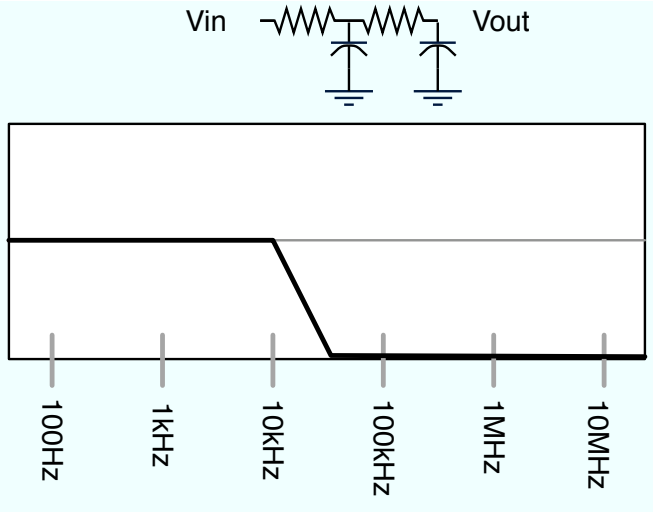


Figure 4.12: Frequency response of a second order filter

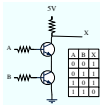
The amount of attenuation is designated by the *order* of the filter. In this case, our filter is a first order—a generalization that means its frequency attenuation is performed in one stage. Higher order filters can be made by adding more stages. For example, a second order filter can be created by simply putting another resistor and capacitor inline like in Figure 4.12. Notice that with this second order filter, the attenuation *roll-off*, or how fast the attenuation takes place, above 10kHz is at a greater angle than its first order counterpart.

Third, fourth, or higher order filters are also easily built by adding more stages.

4.2 Analog and Digital

In the electronics world, the two words that tend to turn up most often are analog and digital. Both represent concepts of signal transmission. Understanding what they really mean is very important.

We live in an analog world. In this sense, the word analog means *continuous*. For example, sounds that we hear are transmitted via analog sound waves. Radio signals that are transmitted through the air are also analog waveforms.



The Buzz...

Understanding Waves

A wave is a *disturbance* that travels through a medium between two locations. There are many familiar types of waves, from sound and light, to ocean and stadium (you know, the kind when a bunch of people at a sporting event stand up and throw their hands in the air).

In a wave, a disturbance from one side of a medium propagates through the medium to the other side. Think of a slinky, for example. At rest the slinky doesn't move. However, if you were to suddenly move one end of the slinky up and down, this motion would carry down the slinky and to the other side. The energy from the disturbance you created moves through the slinky to the other side.

Let's think about the world of sound for a minute. Sounds that we hear are based on the vibration of air modules that our ears are capable of picking up.

But how do we distinguish one sound from another? What makes individual sounds unique? It's the information contained in the waves.

For example, if I was to whistle into a microphone and capture a sample of what the sound looked like, it may look something like Figure 4.13, on the following page. The sample graph that is shown is a representation of *displacement* versus time. That is, the vertical portion of the graph represents the deflection of the microphone's diaphragm as a result of sound waves causing pressure against it.

All sounds look something like this example waveform, though they may have different characteristics. But the key idea is that the graph of the displacement of the microphone is a continuous, almost fluid, wave.

The Scoop on Digital

In contrast to the concept of analog information, the idea of something digital means it is *discrete*. In general, many people think of binary as being a good example of digital information. The idea of two, constrast-

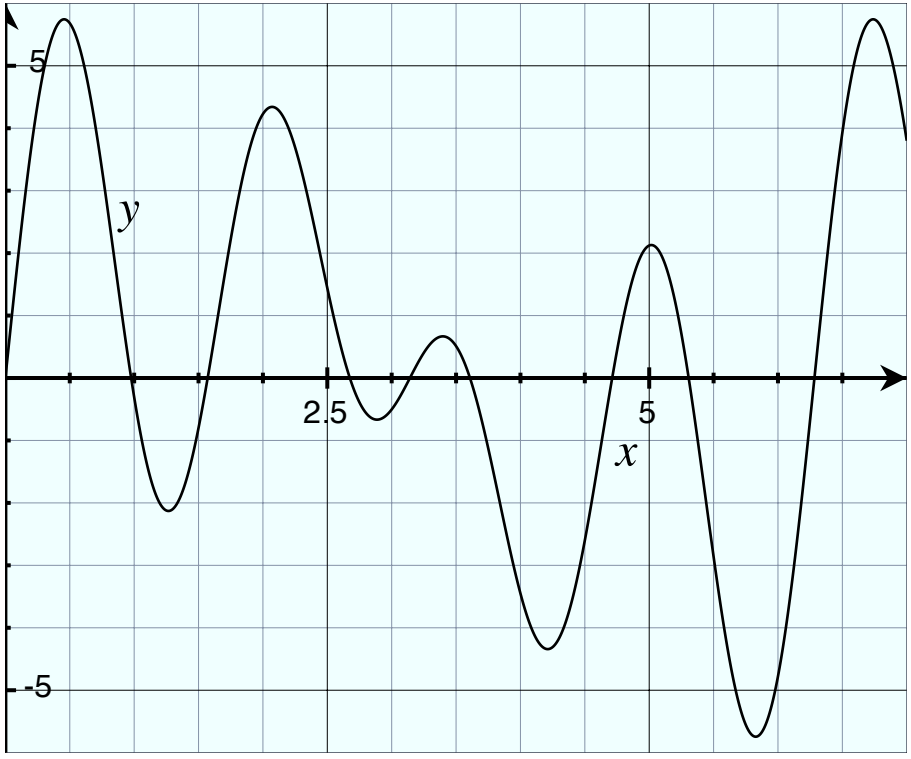


Figure 4.13: A sample waveform

ing states such as LOW/HIGH, OFF/ON, 0/1, RED/BLUE or any other two opposite states makes for a good representation of binary.

But digital can mean more than just binary. Digital means discrete. That is, as the waveform changes over time, a digital waveform is able to "jump" between two values without going through the values in between. This isn't possible in the real world with our sound example, because the diaphragm cannot just discretely jump between two arbitrary positions.

In Figure 4.14, on the next page, the sound waveform has been digitized. That is, all of the displacement values of the waveform are now discrete values (in this case, they are all integers).

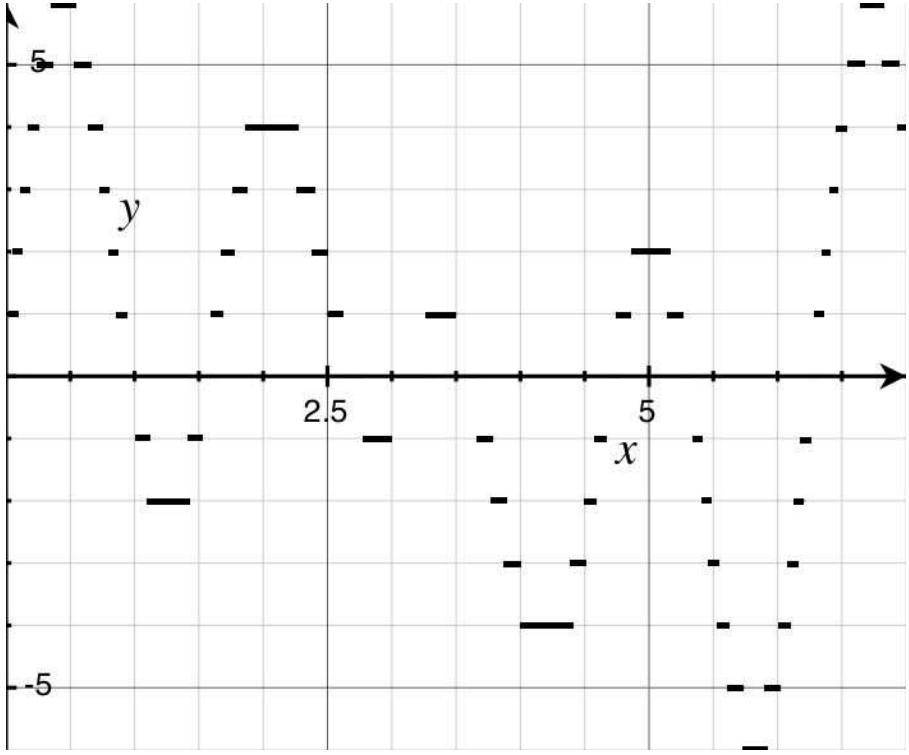


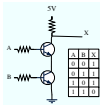
Figure 4.14: A sample waveform, digitized

Digital Conversion

The conversion between analog and digital happens around us all of the time. A good example of this is a portable music player. The music that is saved onto the device originated in the analog world—the singers' voices and the music instruments all started as analog waveforms as recorded by a microphone. However, once the data finds its way into the computer, it becomes digitized.

The computer that records the original sound does it via a process known as *sampling*. It continuously monitors the microphone and takes samples of the waveform from the microphone. The microphone converts the diaphragm displacement into a voltage which the computer can measure.

When the computer takes a voltage measurement from the microphone, it's not able to measure the number exactly. It has to make an approx-



The Buzz...

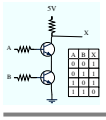
Audio Sampling

Music stored in Compact Disc format is sampled at 44.1KHz (that's 44100 samples per second). It's sampled with 16-bit accuracy, which means there are 65536 possible discrete sampling states.

44.1Khz was chosen as a sampling frequency because the upper end of human hearing is around 20Khz, and this value is just below the Nyquist frequency (see The Buzz, on the following page). 16-bit data values were chosen because they provided good range of discrete values for the sampled sound.

At 44.1KHz, 16 bits, and 2 stereo channels, this creates 176,400 bytes of data every second. The physical capacity of Compact Discs limits them to 74 minutes of music at this rate. Note that any changes to the sampling rate or the size of the sampled value has an impact on data storage. For example, if we wanted 96 Khz sampling with 24 bit data, this would create 576,000 bytes of data every second, which reduces the CD's capacity to just 23 minutes of music.

imation, because it has a fixed, limited amount of memory to store its data. The computer makes an approximation of the sampled data and saves the value as close to the original value as possible.



The Buzz...

The Nyquist Frequency

The Nyquist frequency represents the bandwidth of a sampled signal. It is equal to 1/2 of the sampling frequency. For digital audio sampled at 44.1KHz, the Nyquist frequency is 22.05 Khz.

The purpose of the Nyquist frequency is to designate the upper frequency that can be present in the signal. The idea is that if a frequency exists in the signal above the Nyquist frequency, then its frequency information cannot be reconstructed after the signal has been sampled. For example, since digital audio sampled at 44.1KHz has a Nyquist frequency of 22.05Khz, then no frequency above 22.05 KHz can be present in the sampled audio. If frequencies above 22.05KHz were present, they would not be sampled properly, and would result in *aliasing*, which causes a distortion.

The implication of this is that the sampling frequency of a system must be at least twice the largest frequency present in that system. Otherwise, the frequency information of the system cannot be reconstructed properly after sampling.

Harry Nyquist, for whom the frequency is named, was an engineer at Bell Labs. He was heavily involved in theoretical work in determining the bandwidth requirements for information transmission, most notably for the telegraph.

Chapter 5

The Power Supply

All of the electronics inside of your computer (or most other consumer electronics) depend on the power supply.

This little beast is responsible for converting, managing, and maintaining the power requirements of the machine.

Most consumer electronics make use of DC to provide power to their internal components. They need DC power because they rely on a constant flow of power in order to maintain the state they are operating in. Since AC power is constantly switching on and off it's not suitable for powering many devices.

In the last chapter we discussed that electrical power is distributed to end locations via AC. This presents a slight problem—we must create DC from AC.

5.1 Rectification

To create DC from AC, it must go through a conversion process known as *rectification* that converts AC into DC. In order to rectify, we need a new component: a *diode*.

Diodes

A diode is an electrical component that allows current to flow in only one direction, as demonstrated in Figure 5.1, on the next page. A diode is an electrical equivalent of a check valve. In plumbing, a check valve allows water to flow in one direction only. Any water trying to flow in the opposite direction is stopped by a closed valve. A diode works the same way with current.

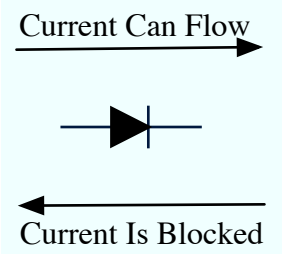


Figure 5.1: Diagram of Diode Current Flow

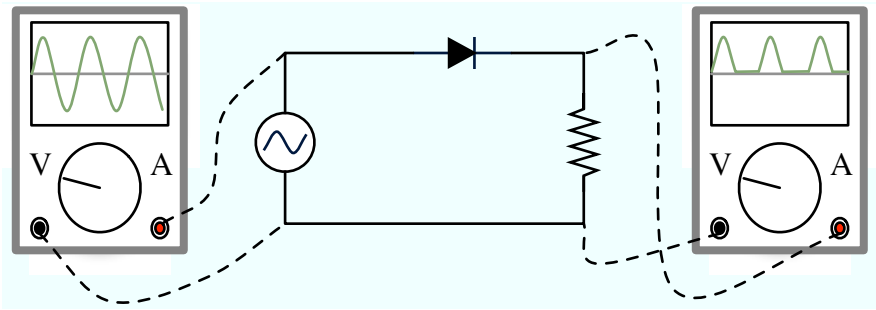


Figure 5.2: A Half Wave Rectifier Circuit

How a semiconductor diode works is explained in detail in Chapter 6, *Semiconductors*, on page 99. For now, though, note that current can only flow in one direction through the diode.

Half Wave Rectifier

The most basic diode rectifier we can have is called a *half-wave rectifier*, and is shown in Figure 5.2. In a half wave rectifier, current can only flow in one direction, so we can only produce a voltage in one direction. This is illustrated in Figure 5.3, on the following page. When the source voltage is positive, current can flow in the circuit and create a positive voltage on our resistor. When the source voltage goes negative, no current flows (the diode acts like an open circuit) and thus there is no voltage across our resistor.

The interesting thing about this simple half-wave rectifier is that if you average the voltage over time, you get a positive value. This isn't true

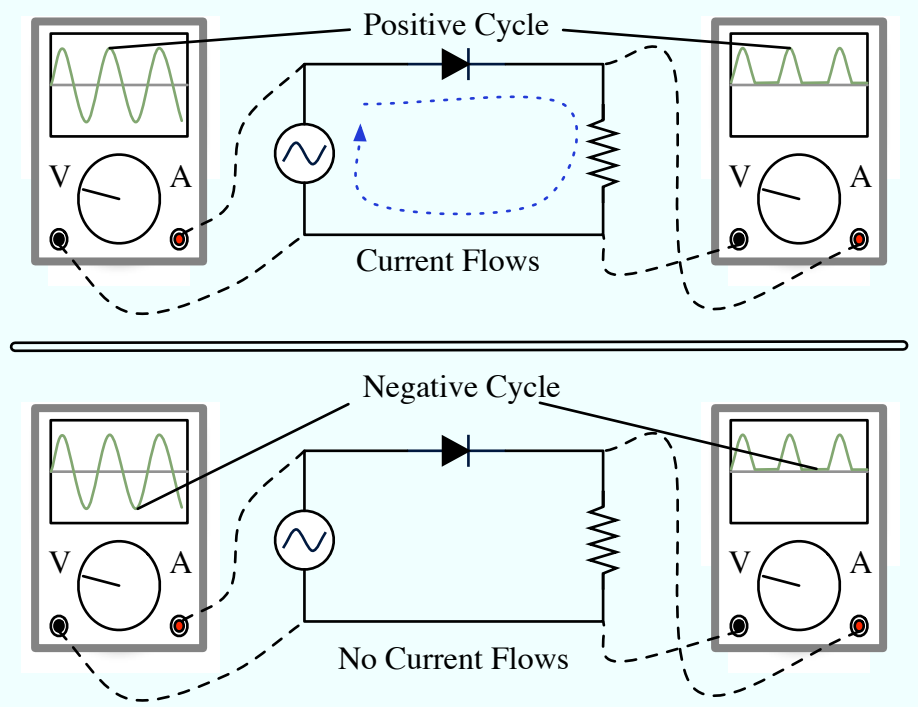


Figure 5.3: A Half Wave Rectifier Circuit

with regular AC, as over time it always averages out to zero. But over time the rectified half-wave does indeed average out to a positive value.

Since the average value is not zero, a half-wave rectified AC voltage can also be viewed as a DC voltage (the average value) with a little bit of fluctuation (the AC part).

Full Wave Rectifier

If we take the idea of half wave rectifier another step, we can create a *full-wave rectifier*. This circuit is made up of four diodes in a *bridge* configuration, as seen in Figure 5.4, on the next page. This configuration allows current to flow in alternating pairs of the diodes, shown in Figure 5.5, on the following page.

The current flow for each cycle in the full wave rectifier is shown in Figure 5.6, on page 88. Note that no matter which part of the source AC voltage phase we are in, the current through the resistor is always

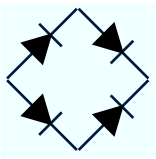


Figure 5.4: A Diode Bridge

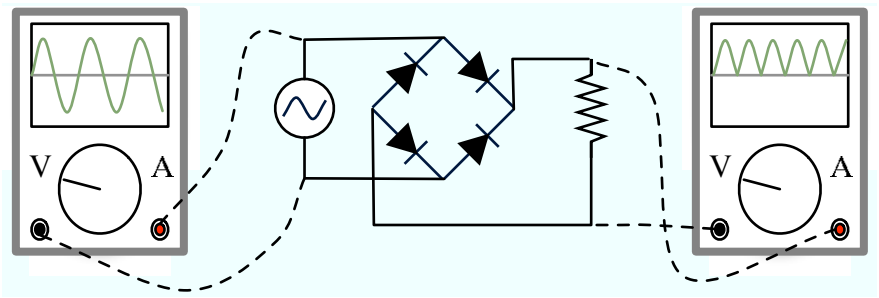


Figure 5.5: A Full Wave Rectifier

flowing in the same direction. This means that our output voltage is always going to be positive. The voltmeter on the right reflects that.

Be aware of the differences between full wave and half wave rectification. With full wave rectification, we are able to transmit the whole portion of the wave to the other side of the bridge. This new rectified voltage, much like its half wave counterpart, also has a positive average value. This means that a full wave rectified voltage can also be thought of as a DC voltage that fluctuates a little bit.

In the half wave circuit there was a period of time between each wave peak where the wave value was 0, while a full wave circuit fills in those gaps with another wave peak. Because of this difference, the averaged voltage of a full wave rectified circuit is twice as much as its half wave counterpart.

With both rectification methods we have created some DC voltage out of our AC voltage. However, we haven't created a crisp clean DC voltage; instead, by removing the negative part of the AC voltage we've created a new voltage that, over time, has a positive average value. In effect,

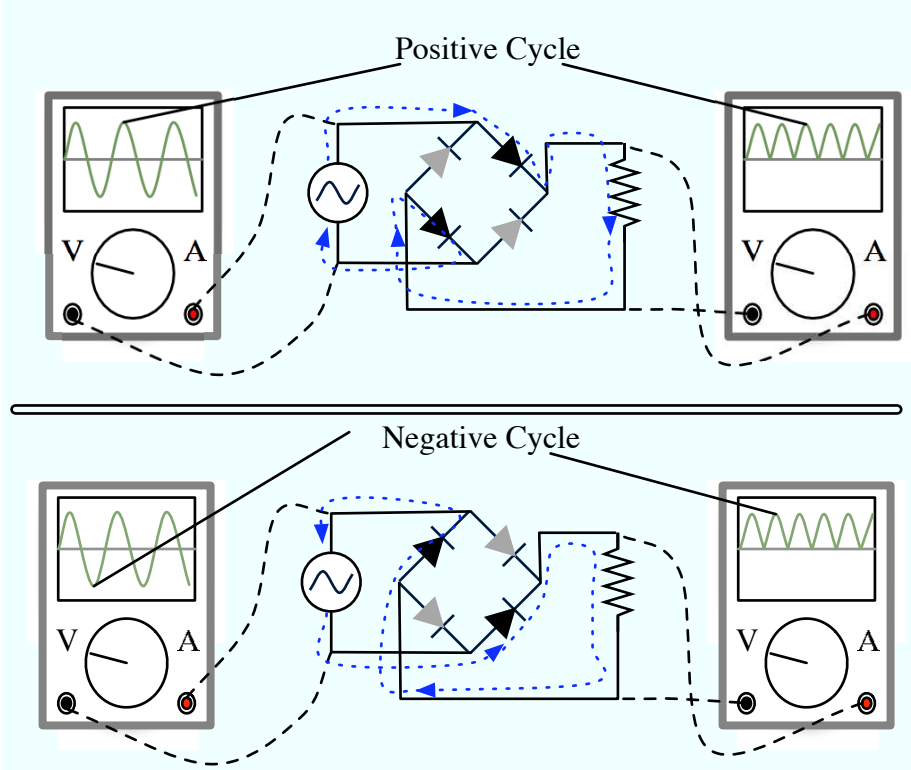


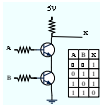
Figure 5.6: A Full Wave Rectifier with Current Flow

we've created a DC voltage that fluctuates. The amount of fluctuation is known as the *ripple* voltage.

Depending on the device we are trying to provide power to, one of these two rectifiers might be good enough to serve as a power supply. However, there's one small trick we can do to make our full wave rectifier even better—add a capacitor.

This is shown in Figure 5.7, on the following page

The capacitor attempts to maintain a constant voltage. It does this by storing charge inside of itself, and when the voltage starts to change it uses this excess charge to try and make up for the difference—keeping the voltage constant. You can think of it like a little battery that doesn't have a very long life. However, before the capacitor loses all of its stored charge, the source voltage waveform returns and the



The Buzz...

Capacitors Can Kill

You're probably aware that opening up a computer power supply while it's still plugged in the wall is a very dangerous idea. However, just unplugging the power supply doesn't make it safe. This is because the capacitors that help smooth the DC voltages retain charge even after the power has been disconnected. Capacitors do slowly discharge and depending on their capacitance this discharging can take anywhere from milliseconds to minutes.

Power supplies are not easy to open for this very reason. Because of the danger, we don't advise trying to open one.

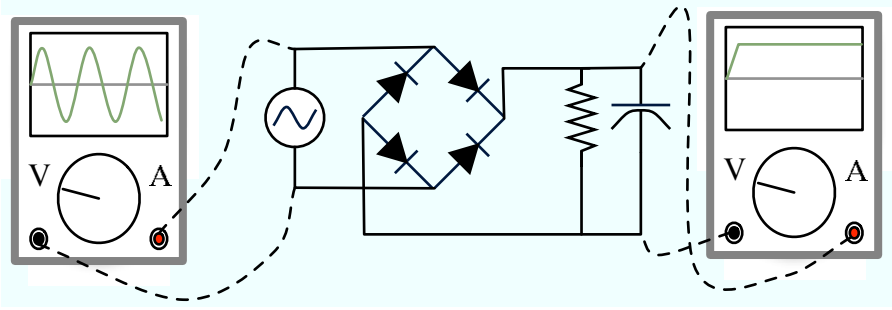


Figure 5.7: A Full Wave Rectifier with Smoothing Capacitor at the Output

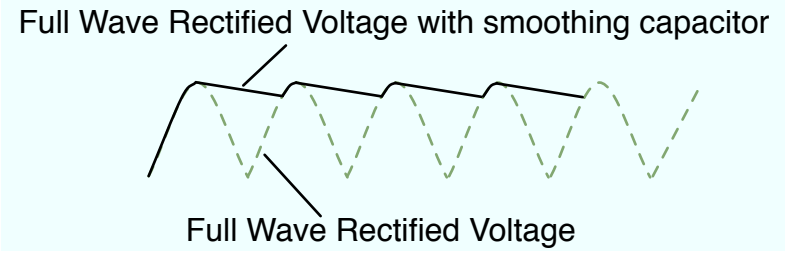


Figure 5.8: Capacitor Smoothing of a Full Wave Rectifier

capacitor is recharged again.

In Figure 5.8, the outputs of a full wave rectified circuit and one with the added smoothing capacitor is shown. Compare this output to Figure 5.7, on the preceding page. As shown, there is still some small ripple in the output voltage. However, it is significantly less than without the smoothing capacitor.

Putting it all together

We've seen the components that are essential to changing AC to DC and help us build our power supply. The use of a diode bridge and capacitors are really all that we need to rectify AC into DC.

5.2 Switching Power Supply

One small issue we haven't addressed yet is how to get the voltage value we want out of our rectifier. Many computer components, for example, are designed to use 5 Volts DC. How do we achieve this?

The easiest and most conventional way is to use a transformer on the AC source side to change our incoming AC waveform into a smaller voltage waveform so that the rectifier output gives us the desired voltage. For example, we may use a transformer that converts 120VAC into 12VAC before it goes into our rectifier. By the time it gets out of the rectifier, the output may be close to what we are looking for (5VDC). Then we can use some resistors to trim this DC voltage slightly in order to achieve our final desired voltage.

This is how rectification was done, historically. But there are some issues:

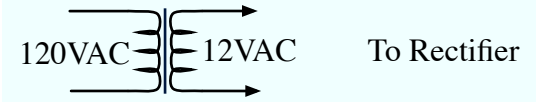


Figure 5.9: A transformer pre-rectifier

- AC power is delivered at 60 Hertz. AC transformers used to convert 60 Hertz power from one voltage to another are bulky and heavy.
- Using resistors within the rectifier causes some power loss, making the overall power supply less efficient.
- The capacitors used to keep the DC output relatively smooth are large and heavy, as well.

To counteract these nuances, modern day power supplies make use of a technique known as *high frequency switching*.

High Frequency Switching

A high frequency switching power supply eliminates the bulky transformer that normally would take the 120VAC and change it to somewhere around 12VAC before it was rectified. Instead, the 120VAC is directly rectified to DC without the transformer.

At this point, we have rectified high voltage DC. The next stage of the power supply feeds this DC voltage to high frequency switching transistors that cycle on and off very fast, somewhere around 10 kilohertz (that's 10,000 times per second). This fast switching essentially turns the DC back into a very high frequency AC signal.

The high frequency AC signal is then delivered to a transformer that converts the voltage down to a level that is desired at the output of the supply. This final AC voltage is rectified and filtered a second time to create a final DC output voltage.

The added step of using a higher frequency may seem like a burden, but it is actually a smart choice. Higher frequency AC signals require smaller and lighter transformers, as well as smaller and lighter capacitors for the rectifiers. Special circuitry within the supply also can tightly control the output voltages to very precise levels.

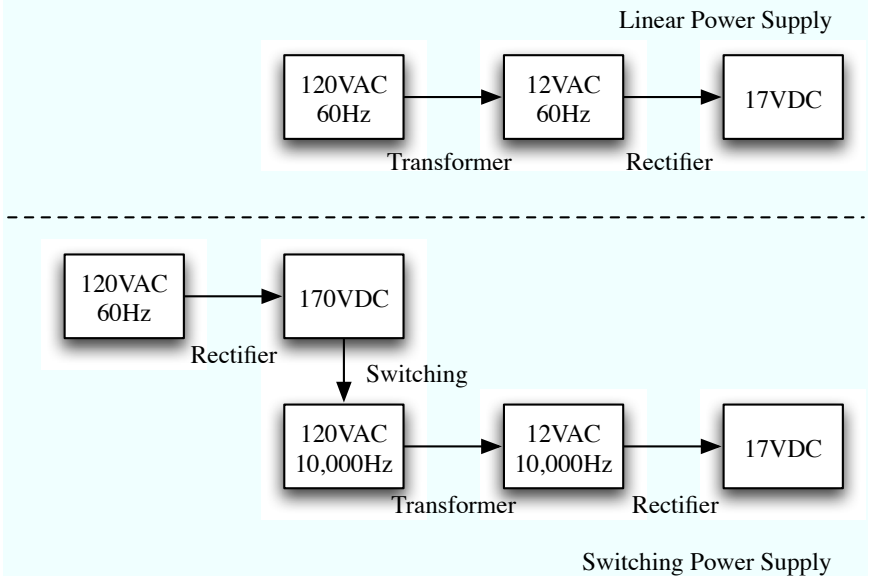
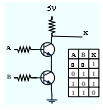


Figure 5.10: An overview of linear and switching power supplies



The Buzz...

DC in the server room

AC is efficient for long distance distribution, but once it reaches the server room it has to be converted to DC to be useful. In a server room, where there may be hundreds of computers making this conversion, the process may generate a lot of wasted heat which translates into wasted dollars.

Some administrators are finding that they can save on cooling expenses by taking out the AC-to-DC conversion process from within the power supply and centralizing it by distributing DC power directly to the servers. It's not a trivial task, but for corporations with hundreds or thousands of machines it may end up being much more economical.

5.3 Bus Voltages

In order to operate, each device inside the computer needs a source of electrical power. A device's manufacturer will specify how and where to connect power to each device. For example, a computer hard drive manufacturer will specify that a certain voltages need to be applied to two pins on the device for it to function. Since there are many devices in the computer that need power, a common distribution line is the most cost effective way to provide it.

Most computer device manufacturers have standardized their equipment to utilize commonly available power within the computer. A typical computer power supply delivers multiple voltages, including +12, +5, +3.3, -5 and -12.

The supply creates these voltages on three separate lines, or busses. Equipment can then attach to ones of these lines and use this voltage as it sees fit, like in Figure 5.11, on the next page.

The following table shows an overview of some of the equipment utilizing these voltages.

- *+12V* - Primarily used to power motors in disk drives. Also used by some cooling fans.
- *+5V* - On older computers, this voltage powered most of the chips on the motherboard including the CPU. Some circuitry on disk drives and plug-in cards may use +5V as well.
- *+3.3V* - Most modern computers power the mother board components, including the CPU and RAM, with +3.3V. It's also used by many plug-in peripheral cards like video cards.
- *-5V* - Older computers provided this for ISA cards which needed it. Many modern power supplies still provide it, though it is rarely used today.
- *-12V* - Serial (RS232) ports do signaling at +12V and -12V, so this voltage is supplied to the serial port. Some newer systems without serial ports may not it.

Note that while some supplies provide -12V and -5V power, there typically isn't a supply of -3.3V power available. This is because the use of negative voltages from the power supply is slowly being phased out, since in many cases the relatively small number of devices that still use negative voltages can generate them internally without the general need

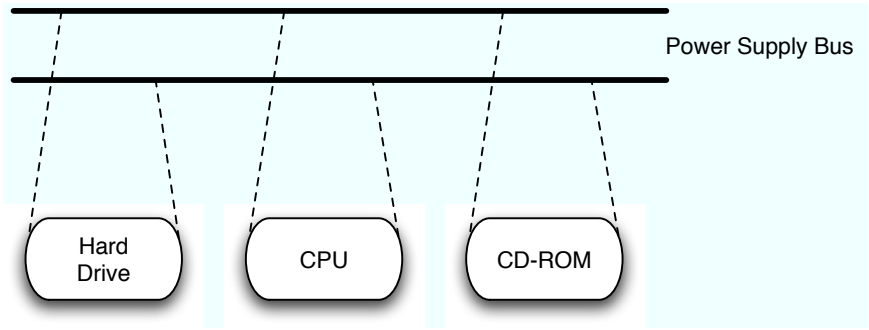


Figure 5.11: The Power Supply Bus

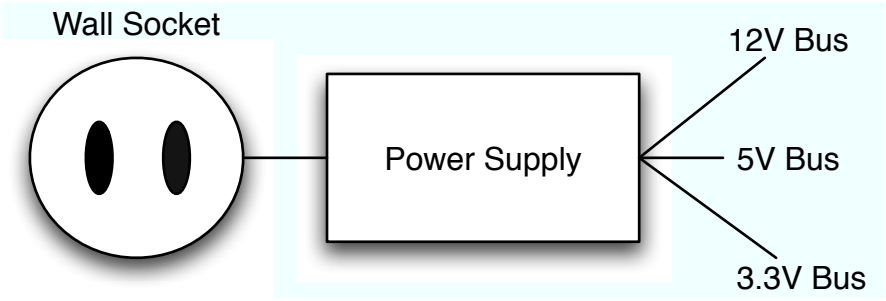


Figure 5.12: A Power Supply Diagram

for a common bus. When +3.3V became a standard supply voltage, it was decided to not provide a -3.3V counterpart that would eventually just be phased out anyway.

The power supply converts the voltage coming from the wall socket into the voltages that are usable by the computer devices, like in Figure 5.12.

Power Delivery

Power from the supply is delivered to the various computer components via cables and connectors. Each peripheral inside of the computer has some connection to get its power from the central power supply.

One standard electrical connector, known as a Molex connector (after the name of the company who makes it) is shown in Figure 5.13, on the

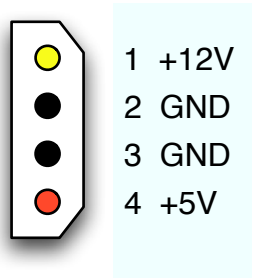


Figure 5.13: A hard disk/CD-ROM drive power connector

next page. This connector delivers both +12 and +5 volts to the hard disk drives and CD drives inside of the computer.

5.4 Power Consumption

Each component inside of the computer needs electrical power in order to function. The power consumption required by any individual component may be small, but when added up all of the components inside of the computer may need a significant amount of power.

When a power supply is rated, it's rated by its *output power*. This is typically listed in Watts. For example, a power supply may be rated for 350 Watts. This is the total amount of power the supply is able to provide to the components within the computer. Based on all of the things in the computer needing power, this may or may not be enough.

Furthermore, it's not the whole picture. The 350 Watt rating is a characterization, but it doesn't specify the amount of power available at the various voltage levels. This is why the supply will specify, for each voltage it provides, power specifications for that particular level. A sample nameplate from a computer power supply is shown in Figure 5.14, on the following page.

Instantanous Power

A 350 Watt power supply isn't always producing 350 Watts. This is just a peak value that it is capable of producing, also known as *instantaneous* power. Most of the time it will be creating less power than that.

Figure 5.14: Nameplate from a power supply

Each component manufacturer will specify a power rating for its device. This is generally how much power the device uses during normal operating conditions. Some devices may need to draw more power occasionally. For example, a hard drive motor may draw up to twice its normal operating current when it is speeding up from rest.

Most power supply manufacturers take this into consideration and add extra capacity into the power supply to account for short bursts of extra power. This is generally specified as a *peak* power output.

5.5 Power Management

Advanced Power Management

To help reduce total power consumption, Intel and Microsoft created a specification known as *APM* to allow the computer to perform power management by turning off various components such as hard drives and monitors after periods of inactivity. APM put the control of the power management in the hands of the BIOS.

Advanced Configuration and Power Interface

An industry standard known as the Advanced Configuration and Power Interface has recently been introduced. ACPI put the control of the power management in the hands of the computer operating system.

It defines seven operational states (note that four of these are sub-states of a main state):

- *G0 Working* - The normal on state of the computer when it is active. Within this state it's possible to put the CPU and other peripheral devices into their own power saving states.
- *G1 Sleeping* - In sleep mode, the power supply cuts off power to more peripherals. There are four substates, known as S1,S2,S3, and S4. Each is progressively a deeper state of sleep.
- *G2 Soft Off* - This "off" mode occurs when the computer shuts itself down.
- *G3 Mechanical Off* - This is the classical "off" state, which occurs when after loss of power.

Part II

Microprocessor Technology

An expert is a man who has made all the mistakes which can be made in a very narrow field.

► Niels Bohr

Chapter 6

Semiconductors

6.1 Electrons through a Vacuum

In the first section of the book, we've introduced to electronics, electrical power, and how it relates to the computer system. In order to continue our study, we need to start focusing on some of the specifics of what makes a computer operate. This section of the book is centered around the processor—the brain that makes the computer tick.

Today's processors are made of millions of tiny semiconductor transistors. So before we can get too far into our study of processors we need to take a look at the building blocks of those transistors.

The Edison Effect

When Thomas Edison was working on his incandescent bulb design in the 1880s, he ended up choosing a filament made of burnt bamboo. However, after a few hours of time, carbon from the filament built up on the inside walls of the bulb causing it to turn black.

Edison wanted to understand why this was happening. The carbon appeared to be coming from filament toward the power supply and was moving through the vacuum to the walls of the bulb. He surmised that the carbon must be able to carry electrical current even through the vacuum. Edison knew that the particles leaving the filament were negatively charged. To help, he added a second electrode to the bulb, between the filament and the bulb, like in Figure 6.1, on the next page. He reasoned that if he was able to place some positive charge onto this electrode it would attract the carbon and keep it from sticking to the wall.

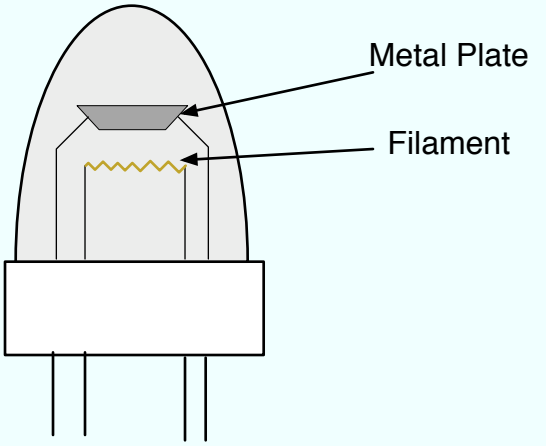


Figure 6.1: Edison's Bulb with Added Electrode and Plate

He found something strange: when the polarity of the electrode was positive with respect to the filament, current would flow into the electrode. However, when the polarities were reversed no current would flow.

Edison was not able to explain the reason why (and the electron would not be identified until some years later). His added electrode did not change the blackening problem caused by the carbon either. So he simply filed it away as an interesting concept and moved on to other projects. He did, however, file a patent on his device in case it turned out to have some special commercial application.

The Electron (Vacuum) Tube

Edison had showed his invention to many people, including a British professor named Ambrose Fleming. Fleming had experimented greatly with the device. He found that it rectified AC current into DC. But there was still a lack of understanding as to why the device worked. Then, in 1889 Joseph Thompson discovered subatomic particles. Fleming quickly realized that the electron was being emitted from the filament and it gave reason as to why a positively charge electrode would attract them.

Based on his knowledge, Fleming created what he called an "Oscillation Valve", the first formal diode. He had been working a lot with wireless

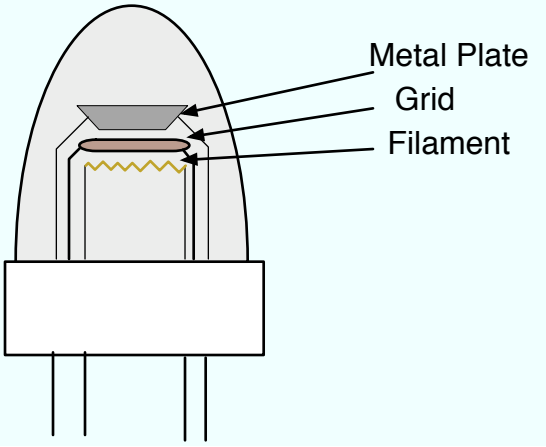
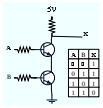


Figure 6.2: DeForest's Audion



The Buzz...

What's a Crystal Detector?

Some natural minerals are able to detect radio signals. Using one of these materials along with a very thin wire known as a cat whisker, a capacitor, and an inductor, a circuit can be created that is able to receive radio signals.

communications and this invention was very helpful in the detection of the wireless signals.

Despite its apparent usefulness, the oscillation valve was not widely used. It was expensive to make and used a large amount of power. Competing devices resulted including a crystal detector (see The Buzz).

But investigations into vacuum tube technology continued. In 1907, Lee DeForest added a third electrode made of a grid mesh as shown in Figure 6.2. He found that by varying the amount of negative charge on the grid, he could control the amount of current that flowed through the two other electrodes. He called his invention the Audion.

DeForest had no idea how profound his invention was. Vacuum tubes

were ideal for amplification. One interested customer, A.T.&.T. became interested in being able to broadcast voice signals further—even all the way across the country. They bought the patent from DeForest.

Over time, new applications for vacuum electron tubes were found. New devices with more electrodes were invented that allowed other types of control. But nothing was able to overcome how large and bulky vacuum tubes were. The heat generated by the electrodes made large scale use impractical.

Luckily, some scientists were searching for a replacement to the vacuum tube. We discuss this replacement, the solid state *transistor*, in more detail in Chapter 7, *Transistors*, on page 109. However, to understand how this replacement works we have to look into the physics of semiconductors.

6.2 Semiconductors

From the name, you can probably infer that semiconductors are just average conductors. The main features of a natural semiconductor are:

- A higher resistance than metal conductors, but a lower resistance than insulators.
- A valence number of +4. (refer to Appendix A, on page 217 for a refresher on what this means)

The two most commonly used semiconductor elements are Silicon and Germanium. Their +4 valence numbers mean that they have a very stable covalent bond structure, as seen in Figure 6.4, on page 104. In its natural form like this, semiconductor silicon is known as an *intrinsic* semiconductor.

One way that semiconductors differ from conductors, such as metals, is in their how their resistances change with temperature. In metals, a rise in temperature causes the atoms to exhibit more vibration which creates collisions in the structure, impeding the flow of electrons. In a semiconductor, however, added heat actually causes the resistance to decrease. This is because the added energy goes into the valence electrons and makes it easier for them to jump into the conduction band and become charge carriers.

+3	+4	+5
Boron	Carbon	Nitrogen
Aluminum	Silicon	Phosphorus
Gallium	Germanium	Arsenic
Indium	Tin	Antimony
Thallium	Lead	Bismuth

Figure 6.3: The periodic table of +3,+4, and +5 valence elements

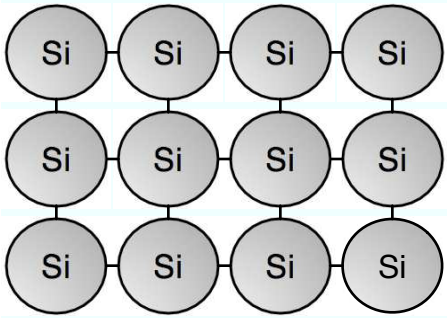


Figure 6.4: silicon's covalent bond structure

6.3 Doping

The usefulness of pure silicon as a semiconductor is limited. However, if we add small amounts of non-silicon material into the structure things become much more interesting.

This process of adding *impurities* to a semiconductor is known as *doping*. Doping results in *extrinsic* semiconductors, meaning they are not in their natural form.

Since a normal semiconductor has a valence of +4, a small amount of impurity will cause a charge imbalance. Take for instance the adding of a phosphorus atom to the structure as in Figure 6.5, on the next page. This phosphorus atom in the structure bonds with the silicon atoms around it. However, with its valence of +5, it has an extra electron available for bonding that is unused in the structure.

Doping a pure semiconductor with a small amount of material with a valence number of +5 (which inclues Phosphorus, Arsenic, and Anitmony) creates an *n-type* semiconductor. It is referred to this because of the excess of free electrons in the material.

Similiarly, you create a a *p-type* semiconductor by doping a pure semiconductor with a small amount of material with valence number of +3 (Boron, Aluminum, Gallium, and Indium). This results because of a hole that is left by the absence of an electron in the covalent bond structure.

Note that doping a semiconductor does not add or remove any charge. The resulting product is still electrically neutral. Doping simply redis-

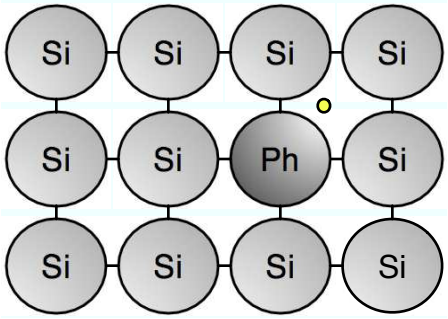


Figure 6.5: A silicon structure with added phosphorus impurity and free electron

tributes valence electrons so more or less free charges are available for conduction.

Understanding Holes

The idea of holes is a bit intriguing. A hole really is a place where an electron could be, or moreso *wants* to be. If there is an electron nearby, it will jump into a hole and fill it up.

Electron holes aren't really holes at all. It's just a convenient description for visualizing energy interactions between electrons and nuclei. In order for there to be an electron hole at all, some energy has to be used to free an electron from the grasp of the nucleus. The removal of the electron tips the nucleus slightly out of balance. It then begins using this energy to attract another nearby electrons to join back up.

In Section 2.4, *Current Conventions*, on page 23, we talked about the difference between hole and electron current. The same idea exists in semiconductors. Within the n-type semiconductor we think of electrons being the major current carrier. In the p-type semiconductor, holes are the major current carrier.

One interesting thing to remember about holes is that the movement of the hole through the p-type material is due to the movement of the *bound* electrons in the structure. That is, the crystal re-bonds from atom to another atom and the hole "moves" in the opposite direction.

Just remember that a hole is nothing more than an empty place where an electron could be.

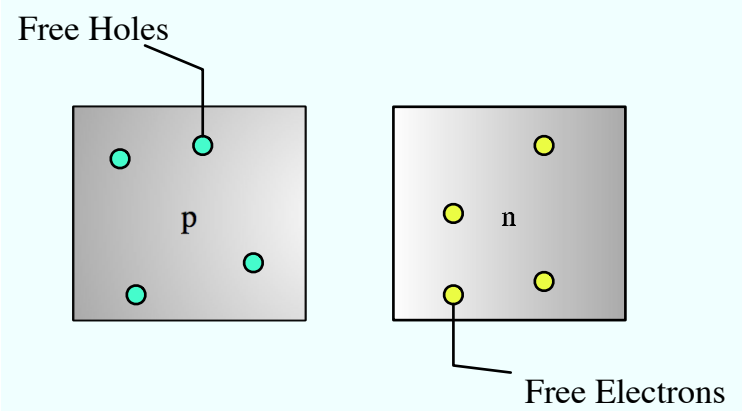


Figure 6.6: P-type and N-type materials

6.4 The PN Junction

The first interesting thing we can do with our doped semiconductors is to put a small piece of n-type semiconductor next to a small piece of p-type semiconductor. The extra electrons in the n-type conductor attempt to move over into available holes in the p-type conductor. At the same time, some of the holes in the p-type conductor end up moving over to the n-type to meet up with electrons. When this happens, we end up with an excess of electrons on the p-type side and extra electrons on the n-type side, creating an electrical imbalance.

This electrical imbalance is known as the *barrier potential* and is shown in Figure 6.7, on the next page. In silicon, this barrier potential is about 0.7 Volts.

The p-n junction becomes interesting when we apply an external voltage to it. But in which direction shall we apply the voltage? There are two possibilities which we call the forward and reverse bias.

6.5 P-N Bias

Forward Bias

If we apply a positive voltage to the p-type material and a negative voltage to the n-type material, we are applying a *forward bias* to the semiconductor. First, the negative voltage at the n-type material is going

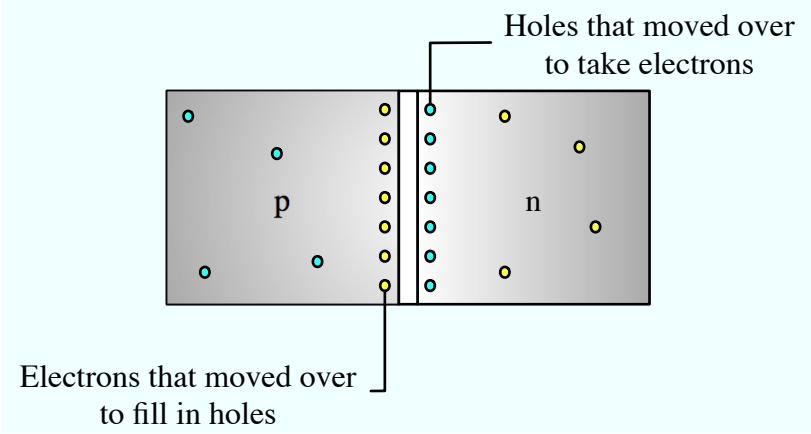


Figure 6.7: The barrier between p-type and n-type materials

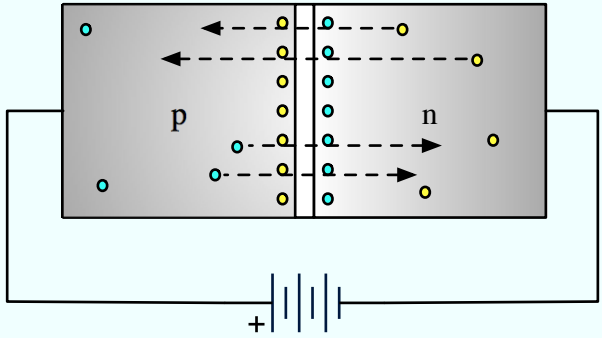


Figure 6.8: A forward biased p-n junction

to attempt to push electrons towards the junction in the middle. The positive voltage at the p-type material will push the holes towards the barrier as well. This reduces the barrier potential.

If the barrier potential is reduced enough, the charge carriers can move through the barrier and out the other side. This means that current flows.

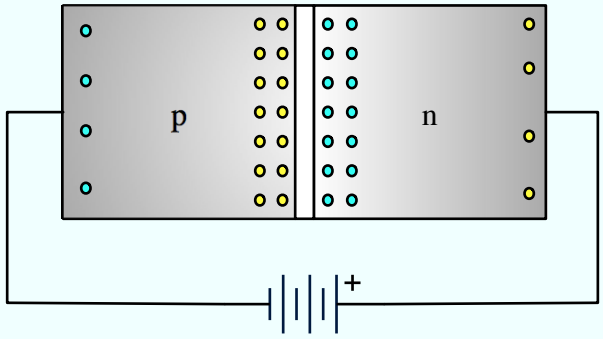


Figure 6.9: A reverse biased p-n junction

Reverse Bias

Applying a reverse voltage to our semiconductor material is known as *reverse bias*. In this condition, the electrons are pulled away from the barrier on the n-type side and the holes are pulled away from the barrier on the p-type side. This results in a larger barrier, which creates a much greater resistance for charges to flow through. The net result is that no current flows through the barrier.

If you want to succeed, double your failure rate.
► Thomas Watson, Inventor of the Transistor

Chapter 7

Transistors

7.1 The History

After World War II, A.T.&T.'s Bell Laboratories started putting more research into semiconductor technologies. A team of William Shockley, Walter Brattain, and John Bardeen was assembled to work on a semiconductor replacement to the vacuum tube.

In the spring of 1945, Shockley had designed the first semiconductor amplifier. His group refined the concepts, even competing with one another to show up each other with a better design.

By 1948, a refined enough product was available and Bell Labs introduced it to the public. Sales were slow, so Shockley quit Bell Labs to start Shockley Semiconductor Laboratories to focus on a more desirable product.

But Shockley's abrasive personality eventually drove away some of his top people; they started their own company: Fairchild Semiconductor. Soon, other companies such as Intel and Texas Instruments started working on their own transistor designs.

It wasn't long before the viability of the semiconductor transistor as a replacement to the vacuum tube caught on.

7.2 The use of transistors

Transistors are three terminal semiconductor devices that are primarily used for two purposes: amplification and switching. One terminal of the transistor is typically used as a control terminal, which a voltage or current applied to the terminal causes the transistor's characteristics

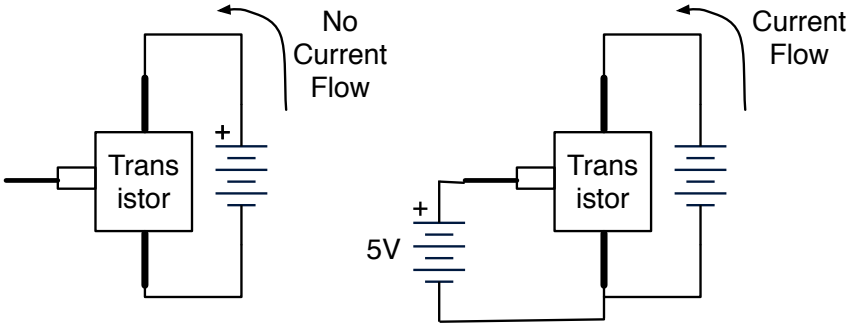


Figure 7.1: A simple transistor switch

to change. This input change results in a change in current or voltage between the other two terminals.

Switching

In their natural state, transistors (depending on the type) tend to either act like open circuits or short circuits. That is, they either easily allow or readily block current flowing through their two main terminals. By applying a voltage or current to the input terminal, it's possible to reverse the transistor into the opposite mode it started in. In this fashion, the transistor acts like a switch.

A simple model of this can be seen in Figure 7.1. In this instance, the transistor does not allow current to flow when no voltage is applied to its input. With 5V applied to the input, current can now flow through the transistor.

Amplification

Transistor amplification happens when a voltage or current is applied to the input that is in between the on and off states of the switching application. Following our example from the previous section, we would find that the amount of current that flows through the transistor depends on the voltage applied to the input. For example, a 1V at the input may allow 20mA of current to flow. 2V may allow 40mA. 3V may allow 60mA. 5V and above may only allow 80mA of current to flow.

This region of the transistor, between its on and off states, is known as the *active* region. Depending on the type of transistor, we can use this

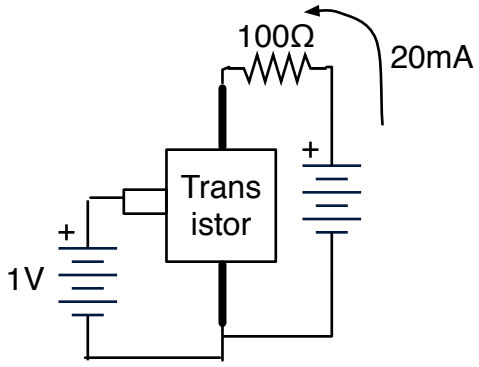


Figure 7.2: A simple transistor switch

region to get a large output from a small input. For example, the 20mA output from the 1V input can be passed on to a 100 Ohm resistor, creating a 20 V output (20 mA * 100 Ohm), as shown in Figure 7.2. 1V at the input of the transistor is *amplified* to create 20V at the output.

7.3 Bipolar Junction Transistor

After our diode design from the previous chapter, the next logical step is to put three doped semiconductors next to each other, like in Figure 7.3, on the following page. This configuration creates what is known as a Bipolar Junction Transistor (BJT). There are two types, the n-p-n and the p-n-p. Each material can be connected to a small piece of wire so we can do interesting things with them. For example, an n-p-n transistor and the three terminals, the *collector*, the *base*, and the *emitter* are shown in Figure 7.4, on the next page.

The schematic symbol for a BJT is shown in Figure 7.5, on page 113. Since we are putting three pieces of semiconductor material together, we can also think of a BJT as similar to two diodes put together. For example, in Figure 7.6, on page 113, we see that an n-p-n transistor is equivalent to two diodes back to back (i.e. the p-type materials are connected).

A common way of using a BJT is to forward bias the base to the emitter. Since this connection and junction is effectively a diode, some electron current will flow from the emitter to the base.

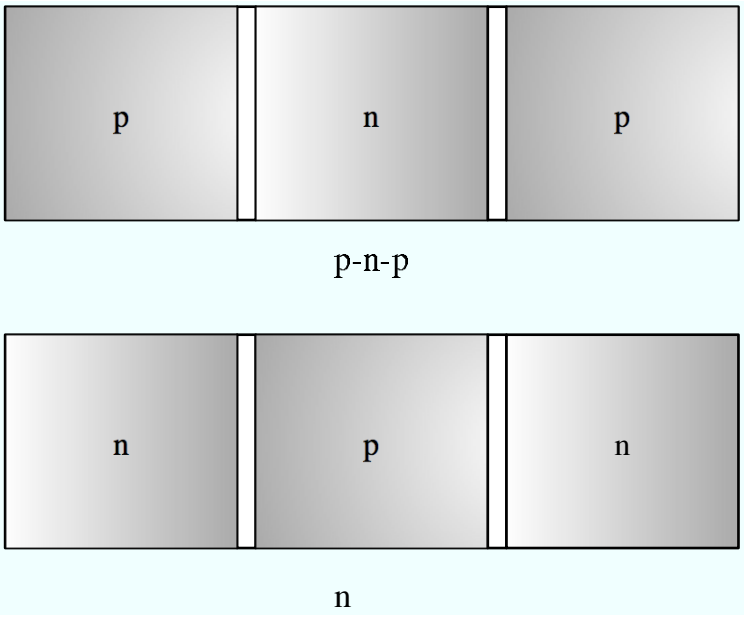


Figure 7.3: An n-p-n material

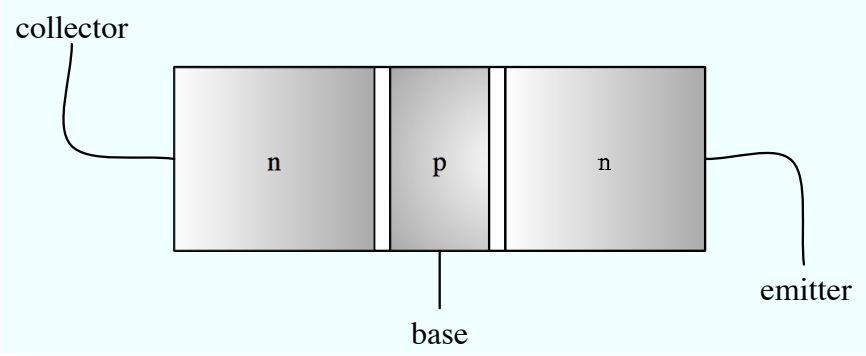


Figure 7.4: Collector, base, and emitter

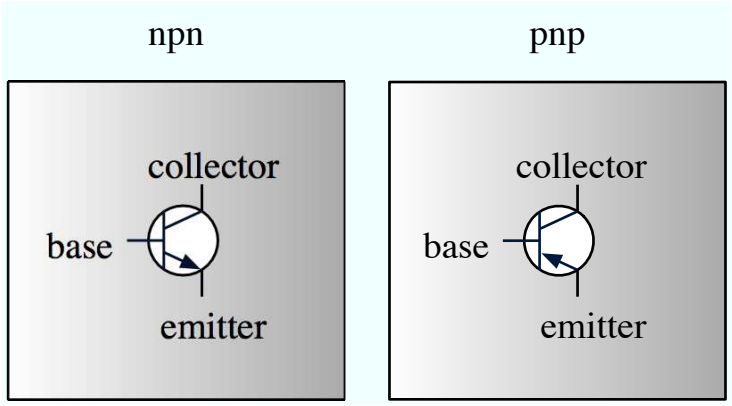


Figure 7.5: BJT Schematic Symbols

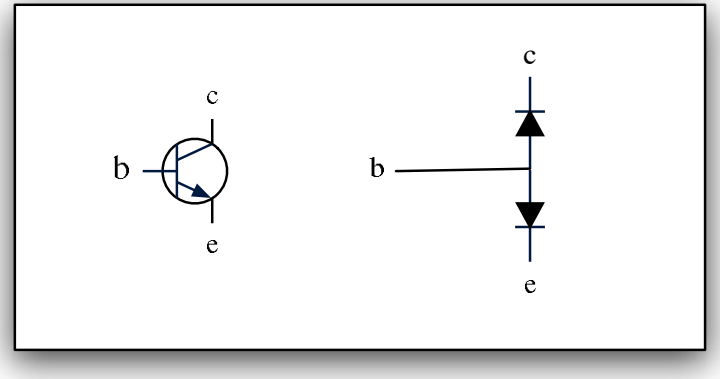


Figure 7.6: An n-p-n BJT equivalent circuit

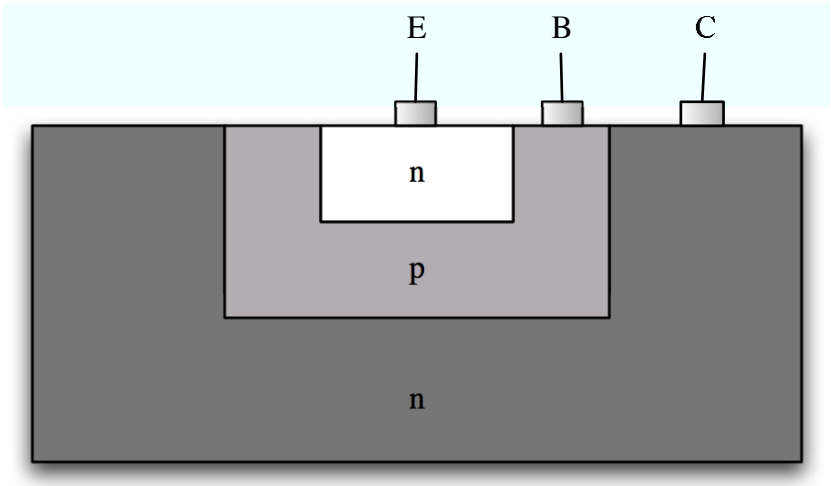


Figure 7.7: Side view of an n-p-n transistor

Similarly, the collector is forward biased to the emitter. This means that the collector to base connection is *reverse biased*. Normally, no current would flow from the base to the collector due to this reverse bias. But the reason is not the same as that of a diode.

No current would flow because there is no source of current carriers available to cross the depletion region. However, since we have electrons flowing from the emitter to the base, they are available to move from the base to the collector—and they do.

Thinking about current flow in a transistor is both tricky and technical. In summary, though, a BJT is a current *amplifier*. For a small amount of (conventional) current sent into the base, a much larger amount of current can be drawn through the collector. If you vary the amount of current in the base, the amount of current going through the collector will change as well, albeit much more drastically.

7.4 Field Effect Transistor

Another semiconductor transistor is the Field Effect Transistor, or FET. Much like the BJT, a FET has three terminals known as the Gate, Drain, and Source. The FET is constructed a little bit differently than a BJT, and is shown in Figure 7.9, on page 116. The two most common

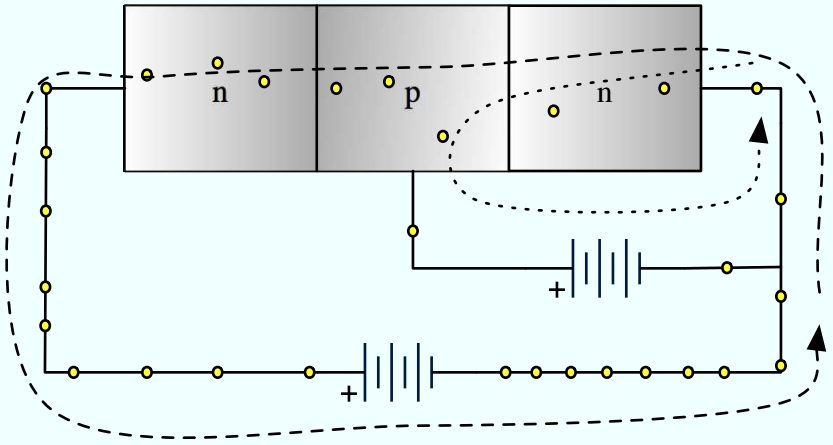


Figure 7.8: Current Flow in a BJT

types of FETs, the Metal Oxide Semiconductor FET (MOSFET) and the Junction FET (JFET) are shown.

Like the BJT, we can use the FET to control and amplify current.

A FET is operated by controlling *channel*. The examples shown in Figure 7.9, on the next page have n-type channels; p-type channel FETs are also commonly used.

A voltage is applied between the drain and the source terminals which are at opposite sides of the channel. The gate is then used to control the current that flows between the two.

The MOSFET

A MOSFET has a gate that is insulated from the channel.

A voltage that is applied to the gate will attract charge from the channel toward the gate. This charge cannot move through the gate because of the insulation. But, this "moved" charge does change the conductivity of the channel.

The most common MOSFET is the *enhancement mode* MOSFET. In an n-channel enhancement mode MOSFET the resistance from the source to drain is relatively high; therefore, very little current can flow. However, a positive voltage at the gate causes the channel to induce negative charges which allows more electrons to flow from the source to the

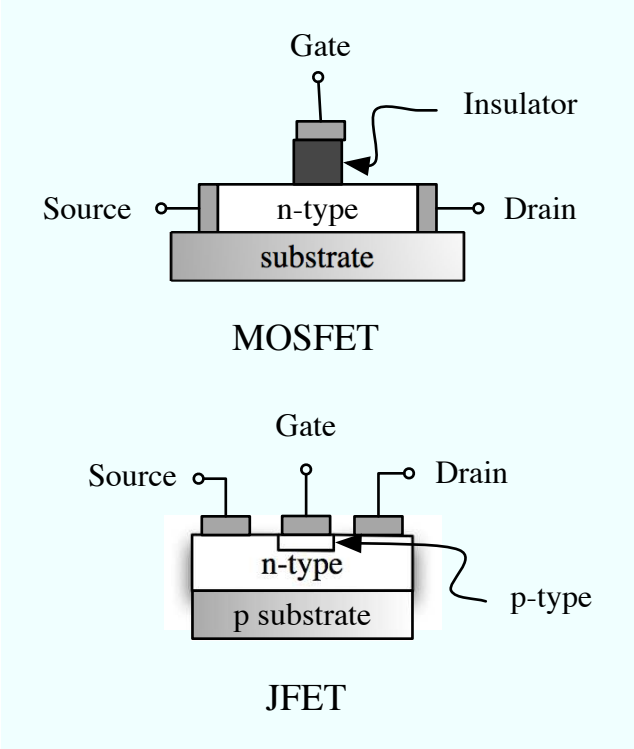


Figure 7.9: Construction of FETs

drain. This means that the resistance of the channel is easily controlled by the gate.

The JFET

In the JFET, a reverse biased junction is used at the gate instead of an insulating material like in a MOSFET.

7.5 The Use of Transistor

BJTs and FETs have many uses. Since both can be used to control a different current or voltage, both are highly suitable as amplifiers. Circuit designers have some preferences in using one over the other

- BJTs have a higher *transconductance* than FETs. This means that their relative ability to amplify a signal is considerably higher.

Because of this, BJTs are more commonly used as signal amplifiers.

- In their "off" state, FETs have higher resistances than BJTs. This means that the current they conduct when turned off is very small. Because of this, FETs are commonly used in switching applications where power consumption is tightly controlled.

From the standpoint of a microprocessor, we are really only concerned with the switching abilities of a transistor and not the amplifying abilities. In the current processor world, the MOSFETs are much more prevalent than BJTs.

7.6 Transistor Logic

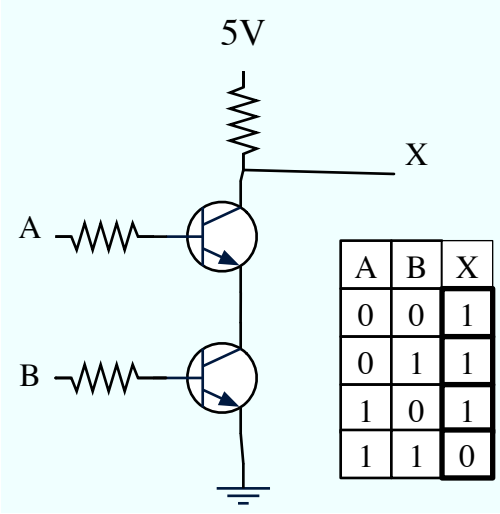
In 1962, Texas Instruments released a series of integrated circuit chips known as the 7400 series, which provided a wide variety of digital logic functions such as AND, OR, and NOT gates. The technology was known as *TTL*, which stands for Transistor to Transistor Logic. They were implemented using BJTs and resistors, then were packaged into chips which could easily be integrated into circuits without the need for having to worry about how to implement the logic.

As an example of the implementation of some of these circuits, the NAND and NOR gates are presented below:

A basic logic circuit built with BJTs is shown in Figure 7.10, on the following page. In this case, we have implemented a NAND logic circuit. If we consider the A and B terminals as inputs, and the X as an output, we can see the results in the table in the figure. The output X will always be 5 volts, because no current can flow through the resistor and down through the transistors. When BOTH transistor bases active, meaning that both transistors will conduct current, then the current can flow down through the transistors. In this condition, the output is now at 0 volts (ground).

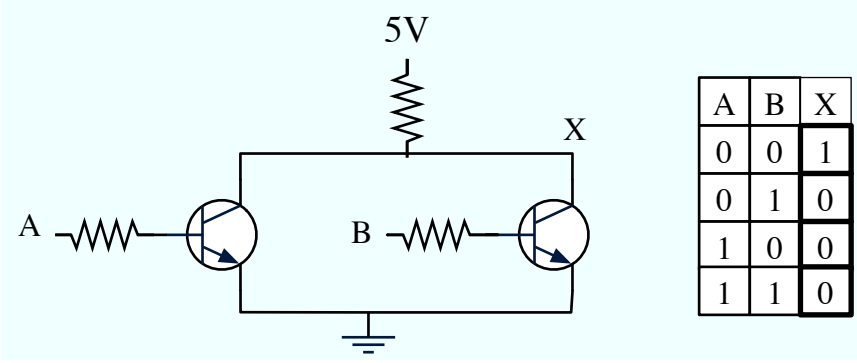
Similar to a NAND gate, a BJT NOR gate can be built like in Figure 7.11, on the next page.

The 7400 TTL series of integrated circuits also provided logic circuits like flip-flops and shift registers, which we will look more closely at in Section 7.8, *The Flip Flop*, on page 120.



A	B	X
0	0	1
0	1	1
1	0	1
1	1	0

Figure 7.10: BJT NAND Gate



A	B	X
0	0	1
0	1	0
1	0	0
1	1	0

Figure 7.11: BJT NOR Gate

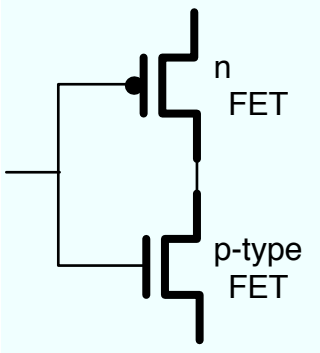


Figure 7.12: CMOS Transistors

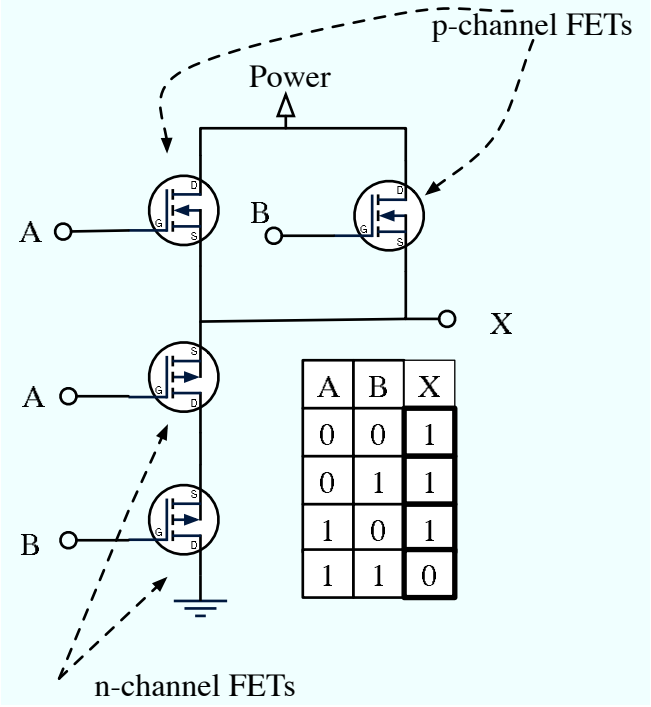
7.7 CMOS

A replacement for TTL is CMOS, or Complementary Metal Oxide Semiconductor. CMOS is a circuit design technique that utilizes MOSFETs in pairs to handle functions similar to the TTL circuits. In 1968, RCA released the 4000 series of chips that provided many of the same characteristics as the 7400 series chips. The 4000 series counterparts were slower, meaning they could not be switched on and off as fast as the 7400 series implementations. On the other hand, the 4000 series chips consumed much less power and could operate over a much wider range of voltages.

CMOS uses a pair of MOSFETs, one p-type and one n-type as shown in Figure 7.12. The advantage to this setup is in power consumption. If an n-channel and a p-channel FETs are connected up in the same way with the same controlling gate voltage, one will be “on” when the other is “off”.

A CMOS based implementation of a NAN D gate is shown in Figure 7.13, on the following page.

The main advantage to CMOS design is in reduced power consumption. With other designs, like TTL, some power is required to maintain one of the switched states. For example, in the NAND gate example in Figure 7.10, on the previous page, the inputs A and B will draw some electrical power when high. This is because the base input of the BJTs draws some current.



A	B	X
0	0	1
0	1	1
1	0	1
1	1	0

Figure 7.13: CMOS NAND Gate

In a CMOS design, no power is consumed when the FET pair is switched either on or off. Instead, the only power consumption occurs during the switching between states.

7.8 Transistor circuits

Transistors are used as building blocks to create integrated circuits. This section takes a look at some of the common base circuits we can create with transistors that will be essential in creating multi-purpose integrated circuits such as microprocessors.

The Flip Flop

A *flip-flop* is a circuit that is able to remain in one of two states. They are an important part of digital logic circuits because they have memory that normal logic gates, like the NAND and NOR gates, lack.

The RS Flip-flop

The basic type of flip-flop can be created by connecting two NAND gates like in Figure 7.14, on the following page. This circuit forms an *RS*, or reset-set flip-flop. This circuit has two inputs, S and R, and two outputs, Q and Qbar. Q and Qbar are opposites. When one is on, the other is off.

The RS flip-flop works like this:

1. *Hold.*- In this condition, nothing happens. This happens when the S and R inputs are both HIGH. The outputs stay as they were.
2. *Set.*- Setting causes the output Q to go HIGH. This happens when the S input goes LOW.
3. *Reset.*- Reset causes the output Q to go LOW. This happens when the R input goes LOW.

The flip-flop is a very simple device, but it is also very profound. It is able to maintain a state. When we perform a SET operation by making the S input go LOW and the Q output goes HIGH, later when the S input goes back to HIGH the Q output stays the same. The flip-flop retains its state. It has memory. The only way to clear this operation is to later make the the R input go LOW.

This behavior is commonly referred to as *latching*.

The RS flip-flop is very easy to implement, as noted in the figure. However, a small problem arises if both inputs were to go LOW, which is an undefined/prohibited operation. In this condition, both Q and the Qbar outputs would be the same (HIGH).

The D Flip-flop

The D flip-flop has only one data input, D, and a clock input, CLK. The D (for data) input is used to delay the output from the input. The output, Q, will match the input, D, whenever the clock input CLK transitions from low to high.

The representation of the D flip-flop is shown in Figure 7.15, on the next page.

The JK Flip-flop

Another common type of flip-flop is the JK flip-flop, and is shown in Figure 7.16, on page 123. This flip-flop is considered the universal flip-flop. The JK flip-flop has two data inputs, J and K. It also has a clock

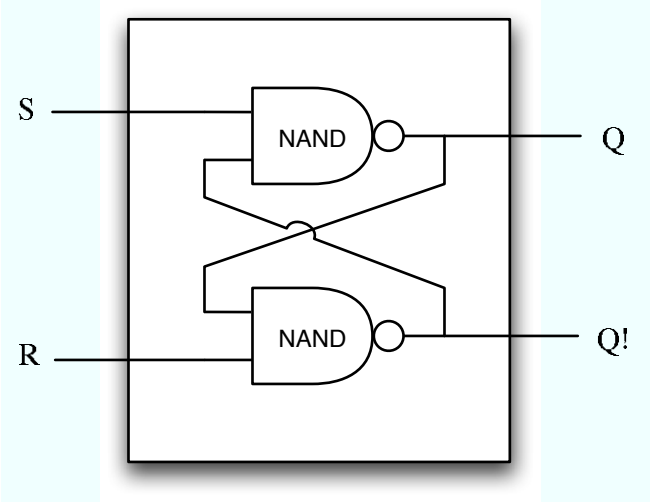


Figure 7.14: RS Flip-Flop

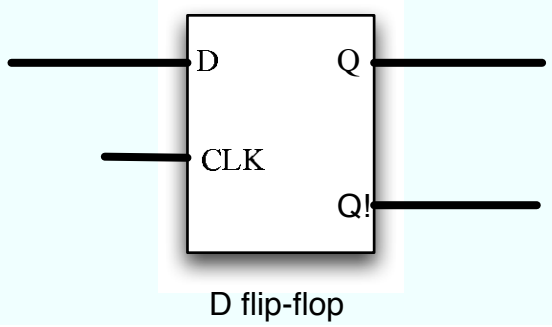


Figure 7.15: D Flip-Flop

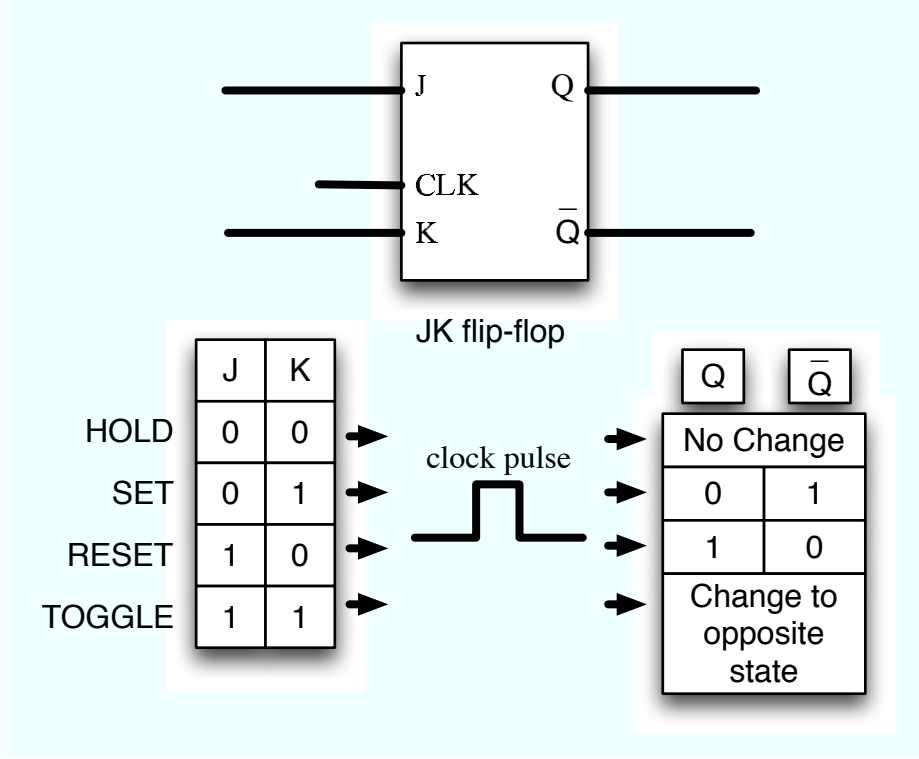


Figure 7.16: JK Flip-Flop

input, CLK. It has the same Q and Qbar outputs as the RS flip-flop does.

The JK flip-flop is a *clocked* flip-flop, which means that its outputs will only change when the clock (CLK) is triggered. This generally happens during the transition between a LOW state and a HIGH state, known as a clock pulse.

The operation of the JK flip-flop is as follows:

1. *Hold* - Whenever the clock pulses, the output remains the same.
2. *Set* - The output Q goes HIGH when the clock pulses.
3. *Reset* - The output Q goes LOW when the clock pulses.
4. *Toggle* - The output Q changes from HIGH to LOW or LOW to HIGH when the clock pulses.

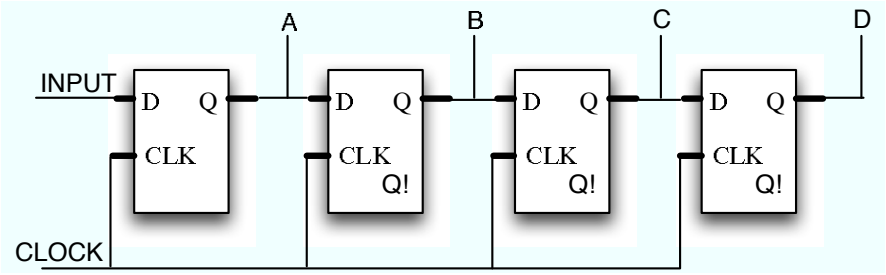


Figure 7.17: A 4-bit shift register

Flip Flop Circuits

Utilizing flip-flops, we can build various types of functional circuits.

The Shift Register

One such circuit, the *shift register*, is shown in Figure 7.17. This shift register is made up of four chained D flip-flops. The clock inputs of each flip-flop is tied together, and the data outputsput of each flip-flop is connected to the data input of the next. The shift register works as follows:

1. Place a bit value (LOW or HIGH) at the INPUT terminal.
2. Pulse the CLOCK input. The value is now latched to the first flip-flop.
3. Repeat the process three more times. On the final CLOCK pulse, all four values have now been latched into each of the four flip-flops.

Ripple Counters

A chain of J-K flip-flops can create a *ripple counter*. A ripple counter counts the number of input pulses to the CLOCK input and outputs that value as a binary representation of the number of counts. The circuit for a 3-bit ripple counter is shown in Figure 7.18, on the following page.

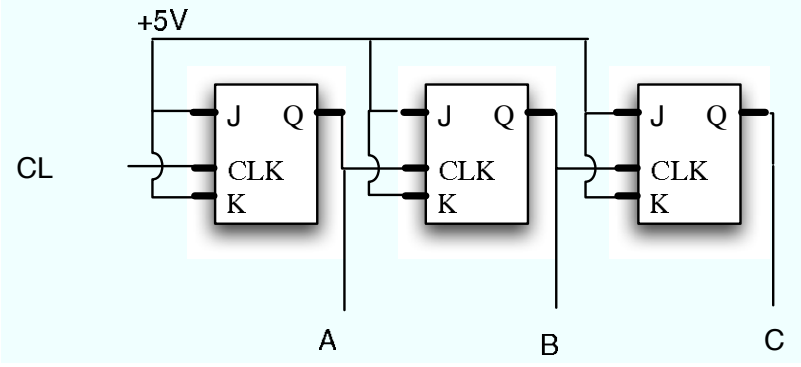


Figure 7.18: A 3-bit ripple counter

There is no reason anyone would want a computer in their home.

► Ken Olson, Digital Equipment Corp.

Chapter 8

The Processor

8.1 The history of the processor

The first integrated circuit, with multiple discrete components in a single package, was invented by Texas Instruments engineer Jack Kilby in 1958. For the next 10 years, the process of packaging transistors into a small integrated package was refined.

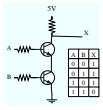
In 1968, Robert Noyce, Gordon Moore, and Andy Grove left Fairchild Semiconductor, a dominant producer of integrated circuits. They created a new company called Intel where they could use their talents towards future integrated circuit design.

Building a Processor

The first commercially available microprocessor was the 4004 created by Intel released in November 1971. A 16 pin, 4-bit processor, it was originally designed for use in calculators for the Japanese company Busicom.

One of the chip designers, Federico Faggin, believed that there were markets for the chip beyond just Busicom's calculators. To prove his point, Faggin used a 4004 chip to create a control circuit for a production tester of 4004 chips. Faggin started a movement with Intel to consider the commercial release of the 4004 as a multipurpose processor. This required re-negotiation of contracts with Busicom, but ultimately proved to be very successful.

Around the same time, Texas Instruments (TI) developed and released its own 4-bit microprocessor, the TMS 1000. TI's processor was a little bit different in that it didn't require any external chips to store data like



The Buzz...

Beyond Semiconductors

Robert Noyce, Gordon Moore, and Andy Grove have made contributions far beyond their initial semiconductor designs.

Robert Noyce was known as the "Mayor of Silicon Valley". Beyond his technical contributions, he is largely credited with beginning the trend of allowing a casual work atmosphere. He believed in giving freedom to bright young employees to prosper. These trends are still followed today in many Silicon Valley companies.

Gordon Moore is most famously known for Moore's Law, in which he stated that the number of transistors on a computer chip would double every 18-24 months. He was Chairman and CEO from 1979 to 1987, after which he became Chairman of the Board.

Andy Grove became president of Intel in 1979 and CEO in 1987. From 1997 to 2005 he served as Chairman of the Board and currently serves as a management advisor.

the 4004 did. This chip was significantly bigger than the 4004 and had more pins (28), but was designed to be a multipurpose processor from the start.

Intel continued moving forward with new chip designs. Its next major release, the 4040, was the successor to the 4004. Shortly thereafter, the 8008 was released. It was designed for the Computer Terminal Corporation for use within one of their programmable terminal products. The prototype had problems in its memory circuits which required a redesign. It was delivered late to CTC; too late, in fact, to be used in their product.

Intel marketed the 8008 to other companies with some success. However, many customers were requesting things that weren't in the 8008. Federico Faggin, as its lead designer, took note of these requests.

A Revolution

In 1974, Intel released the 8080 chip. It is widely considered to be the first general purpose microprocessor. Its use in many early microcomputers, like the Altair 8800, gave it widespread popularity.

Soon the 8080 had competitors. Motorola released the 6800 and Zilog, a company founded by Faggin after departing Intel, released the Z80. Soon, many companies were competing for market share in the processor world.

However, it was a bit of marketing on Intel's part that sealed the deal for their dominance in the processor market. In 1978 Intel released the 8086 and 8088 (the latter of which was the same as the 8086, but had a reduced data bus size to work more easily with other cheaper chips). The release of the 8086/8 was done simultaneously with a marketing program and sales campaign known as "Operation Crush". Operation Crush trained Intel employees to focus on customer support, drive sales, and show long term commitments to supporting their products.

Operation Crush worked. In 1981, IBM chose Intel's 8088 for use in their new personal computer line. While the choice of Intel may not have been what led to the ultimate success of the PC, it certainly helped Intel become a dominant player in the microprocessor market.

Over time, new processors were released, like the follow-on 80286, 80386, and 80486. Eventually Intel released the Pentium line of processors which continue to remain popular choices even today.

8.2 Processor Fundamentals

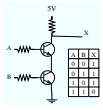
Processors provide the following components:

- Data storage in the form of registers.
- Data manipulation in the form of a logic library.
- Data paths in the form of the pipelines.
- Clock and synchronization circuitry

Data Storage

Registers

Registers are areas of memory internal to the processor. They are used directly by the processor for calculation and data access. For example,

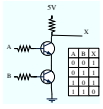


The Buzz...

What about Babbage?

Long before the invention of the transistor, the vacuum tube, or even electricity, Charles Babbage had created the first analytic computer known as the Difference Engine. Fed up with the problems of using mathematical tables for calculations, Babbage built a machine that could calculate the lookups of numeric tables mechanically and be correct every time.

Babbage also designed, though never built, the Analytical Engine. This machine, in theory, was programmable via the use of cards. The concept of being able to write programs, on cards, and then reuse them to get calculations is the foundation for modern day computer processors.



The Buzz...

What is Magnetic Core Memory?

Magnetic core memory was used for data storage on early computers. It is composed of small iron rings in a grid structure with wires running through them. By manipulating the various wires, a magnetic field could be induced in the ring and could then later be read as a 0 or 1.

if the processor needed to add two numbers together, it would most likely look for both of the numbers in two different registers, perform the addition, and store the result back in one of those registers.

Registers are measured by the number of bits they hold. On an Intel Pentium 4 chip, for example, most of the registers are 32 bit registers.

Registers can be implemented by flip-flops, magnetic core memory, or more commonly, as a register file.

Register Files

A register file is a common way of creating and accessing registers within a processor. It contains the physical registers of the CPU. The implementation of the register file is the same as Static RAM (see Section 9.3, *Static RAM*, on page 142). In other words, the CPU registers are nothing more than special purpose implementations of RAM.

The ALU

The Arithmetic Logic Unit (ALU) is part of the processor that handles the core mathematical abilities of the processor. In general, it performs the following types of operations:

- Arithemetic such as addition, subtraction, and sometimes multiplication.
- Logic operations, such as AND, OR, and NOT.
- Bit manipulations, such as shifting or rotating bits within a byte.

Clocks

CPUs as well as most logic circuits are *synchronous*, meaning that operations are performed sequentially. The clock signal handles this synchronization. The clock signal is usually in the form of a square wave and the frequency of the wave depends on the abilities of the processor and other electronics and the circuit designer.

The clock frequency has to be set to ensure that all parts of a data operation are in place before performing that operation. However, some parts of the processor may work faster than others meaning that some data may be sitting idle waiting for the next clock signal while other data parts are still be processed. So, an important part of CPU design involves making sure that the various parts of the processor are good at handling parallel tasks.

8.3 Processor Packaging

Dual Inline Packaging

The first commercially available processors were packaged as DIPs, or Dual Inline Packages (sometimes also called DILs). It is a rectangular package with two rows of pins protruding downwards. DIPs come in a wide variety of pin counts, from 8 to 64.

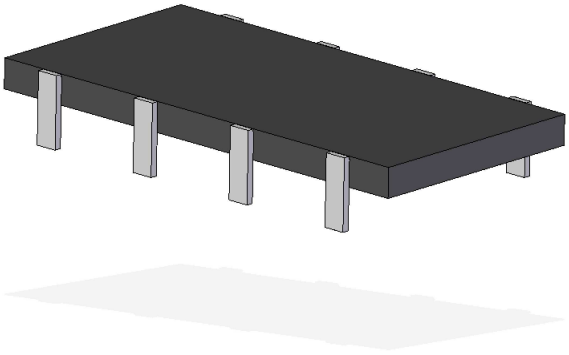


Figure 8.1: A DIP style chip

For processors, DIPs were very popular in the early days of computers thorough the early 1980s. They continue to remain popular for other types of integrated circuits, particularly for hobbyist tools like programmable memory chips, because they are easy to handle.

By the mid 1980s, however, the number of pins on the processor started to get very large and it was becoming clear that DIPs were not able to handle the count.

Pin Grid Array

The next step in processor packaging was PGA, or Pin Grid Array. With PGA, the pins were moved to the bottom of the processor in a large grid pattern instead of out the side like in the DIP. The advantage to this was the number of pins available was a function of the surface area of the processor. Pin counts were able to go into the hundreds.

Other Packaging Types

Integrated circuits are found in many other package types. Some of the more common types are:

- PLCC - Plastic Leadless Chip Carrier. This is a four sided package with the electrical connectors along each of the sides.

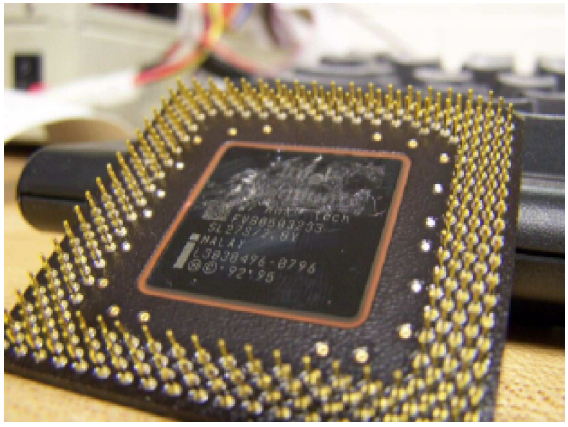


Figure 8.2: A processor with PGA

- SOIC - Small Outline Integrated Circuit. This package is similar to DIP, but shorter and more narrow.
- BGA - Ball Grid Array. Similar to PGA, this package type uses small balls of solder at the connection method instead of physical pins coming out the bottom of the package. BGA has the advantage of mounting directly to the surface of the board instead of having pins that must go completely through the board.

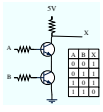
8.4 Processor Cooling

With all of the electronics inside of it, the processor generates heat while it is working. How much heat the processor generates is dependent on:

- The efficiency of the electronic design.
- The clock speed.
- The operational voltage.

Processors cannot be allowed to get too hot, or they risk damaging themselves. Modern processors have thermal sensors within the housing of the processor which can be monitored. If the temperature exceeds a critical value the sensor can shutdown the processor.

To help with generated heat, it is dissipated by the use of a heat sink and a fan. A fan blowing air past the CPU aids in heat dissipation



The Buzz...

What is the Peltier effect?

In 1821 a physicist named Thomas Seebeck discovered that if two pieces of wire, made from different metals, were attached together at one end and heated, a voltage would result between them. The relationship between the amount of heat (temperature) and the resulting voltage could be measured, interpolated, and put into a table. Seebeck's discovery, known as the thermocouple, is widely used today as a method for measuring temperatures.

The Peltier effect, discovered in 1834 by Jean Peltier, is the opposite of the Seebeck effect. When current is passed through two different metals at a junction, heat transfers from one junction to the other. One junction gets colder while the other one gets warmer.

Simply, a thermoelectric Peltier cooler transfers heat from one side of the device to the other. This can be used to cool things such as processors by keeping the cold side of the cooler next to the processor.

by transferring heat away from the CPU. The heat sink helps as well by creating more surface area over which the heat can be dissipated. Because the heat sink contact with the processor may not be ideal, a thermal paste is sometimes used between the two. The thermal paste helps conduct heat into the heat sink and away from the processor.

More recently, cooling techniques such as water cooling and thermoelectric cooling using the Peltier effect are becoming commonplace.

Part of the inhumanity of the computer is that, once it is competently programmed and working smoothly, it is completely honest.

► Isaac Asimov

Chapter 9

The Motherboard

The processor may be the heart and soul of the computer, but it's highly dependent on all sorts of other devices to get things done. We'll take a look at these peripheral devices in the next section.

The progression of design of today's motherboards stems back to the early 1980s and IBM's release of the first PC. When it was released, IBM also released the design of their motherboard and specifications on how to talk to their BIOS as open standards. The idea behind it was to keep a proprietary lock on their BIOS design, which meant they would always be one step ahead of the game in the industry. The plan failed miserably, however, when competitors quickly reverse engineered the BIOS.

This opened up the market to IBM compatible clones made by other companies. Many of the original designs used in IBM's PC motherboard are responsible for the designs in motherboards today.

In this chapter we'll look at what makes up a motherboard, some of the types of things found on typical motherboards, and some of the issues facing motherboard design.

9.1 Circuit Connections

Ancillary to the processor, there are many other circuits in the computer that exist in order for the processor to work its magic. The components include many of the items we've already learned about: resistors, capacitors, diodes, and more.

So far, our discussion of connecting electrical components together has been done via wires. This is a great way of making interconnections

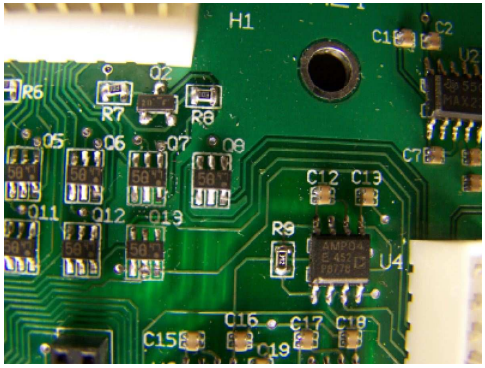


Figure 9.1: A printed circuit board

in the lab, but it's not very practical once the number of components becomes large. Furthermore, the mobility of computers today is far too high to make using copper wire to connect individual pieces together practical. Also, how would the resistors, capacitors, and diodes be fastened down inside of the computer.

The Printed Circuit Board

The printed circuit board (PCB) is the answer. Most likely you are already familiar with printed circuit boards, as almost every consumer electronics item has at least one in them. They are used to both hold components in place and provide the electrical connections between the components. A picture of one is shown in Figure 9.1.

The PCB is made up of a number of parts that make everything work together.

- Components - Various electronic components are attached to the board. In Figure 9.1, the components are surrounded by white borders. These borders are made from an ink that is silk screened on. Each component is also labeled with a designator, like R9. The letter signifies the component type (R for resistor) and the number signifies it is the 9th resistor.

  Components are attached to the board using solder. In the example figure, these components are attached to the PCB directly to the surface by small *pads*. Parts that are soldered directly to the surface are known as *surface mount* components.

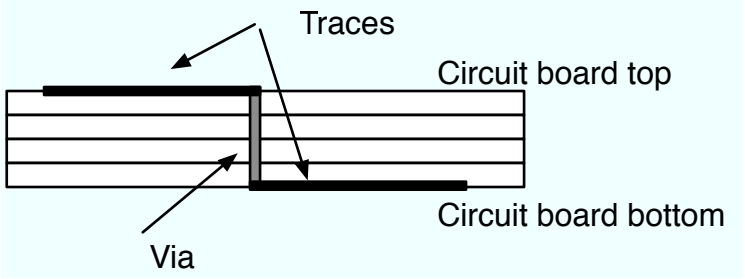


Figure 9.2: A PCB side view showing traces connection with a via

Some components which have pins may be soldered to the board through small holes. These components are known as *through hole* components.

- Traces - The lines that connect the components together are called *traces*. Traces are small copper "wires" that are directly attached to the board and connect components together. A large number of traces are seen in the figure.
- Vias - One problem with traces on a printed circuit board is that they are purely two dimension. There is no way to jump a trace over another one. This presents a problem when there get to be a large number of traces—how do you connect up components where the traces need to jump over each other? The answer is with *vias*. Vias are small holes that have been drilled in the board. A small copper ring lines the hole, and runs through the board all the way to the other side. Vias allow traces to become three dimensional. A side view of a PCB showing traces and a via is shown in Figure 9.2.

Printed circuit boards can have multiple layers. Some boards have just one or two layers, which consist of the front and back of the board itself. However, it is possible to sandwich more layers in between. Vias then can be used to run traces in between multiple layers. It is not uncommon to see up to 16 different layers in a printed circuit board, though designing and building these boards is significantly more expensive than a one or two layer board.

Making a Printed Circuit Board

The first step in making a circuit board, such as a mother board, involves design. Typically design software is used where a designer can put down chips, resistors, capacitors, and other devices. The designer will specify how connections need to be made. For example, she may indicate to the software that she wants pin 2 from chip U7 to connect to pin 4 on chip U12.

For small circuit boards, the designer may choose to make the connection herself by specifying the geometry of the board, where each component will be placed, and manually drawing the traces and vias. However, for larger projects such as a motherboard, the whole layout process is usually automated by software.

Creating the Board

Once the design and layout of the circuit board is complete, the process of creating the board begins.

Boards are created using the following steps:

1. A blank board is ordered and cut to the size required by the designer. If the board is to have multiple inner layers, each layer will be created the same way and bonded together in the end.
2. The blank board has a solid sheet of copper. A negative mask is applied to the board, after which a chemical is used to remove the copper from the places on the board where no copper is needed. The mask is then removed from the board.
3. Vias are drilled into the board using computer controlled milling machines, or in some cases lasers.
4. The drilled via walls are plated with copper.
5. Pads which will have components attached are plated with solder.
6. Artwork and lettering are silk screened onto the board.
7. Components are placed onto the board by hand or by machine. They are soldered into place.
8. In some cases, electrical testing is done to ensure the board has been populated correctly.

9.2 Bus Types

A *bus* is a system that transfers data between two components in a computer.

Today's PC motherboards typically use a bus architecture known as Northbridge and Southbridge. The Northbridge is a set of chips that handle communication between the CPU, memory, and some expansion slots. The Southbridge handles some of the slower components such as the serial ports, USB, and the hard drive controllers. The Southbridge connects to the Northbridge as shown on Figure 9.3, on the next page.

Front Side Bus

The front side bus connects the CPU to other components on the motherboard, like memory, the BIOS, and expansion slots. A number of different electrical implementations exist based on the manufacturer. However, the FSB does contribute to some very important aspects of the computer.

CPU Frequency

The CPU clock is directly controlled by the FSB. The FSB operates at a lower frequency than the CPU. The clock signal that controls the FSB is run through a frequency multiplier (see Section 9.4, *Frequency Multipliers*, on page 146) before being fed to the CPU.

Memory Frequency

The clock that controls system memory also derives from the FSB clock.

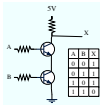
Expansion Busses

IBM's original PC had a motherboard with *expansion slots* designed to connect peripheral cards to the motherboard.

ISA

ISA, the Industry Standard Architecture, is an expansion bus type originating with IBM's PC in 1981. It was originally designed as an 8-bit system, meaning that 8 bits of data could be transmitted at a time. The bus data rate was 4.77MHz.

In 1984, with the release of Intel's 80286 processor, the ISA bus was enhanced to a 16 bit bus with a clock speed of 8 Mhz.



The Buzz...

Front side bus logic

Intel processor based FSBs use an implemetation known as Gunning Transceiver Logic (GTL). GTL is a form of electric signaling that relies on very small voltage changes (less than a volt) that allows for high speed data transfers. AMD Athlon based processors, on the other hand, use a FSB implemenation known as EV6.

As computers today are starting to see applications with multiple processors, the scope of the FSB (which was designed for one processor machines) is changing. Newer replacement technologies like AMD's HyperTransport or Intel's IHA are replacing traditional front side busses.

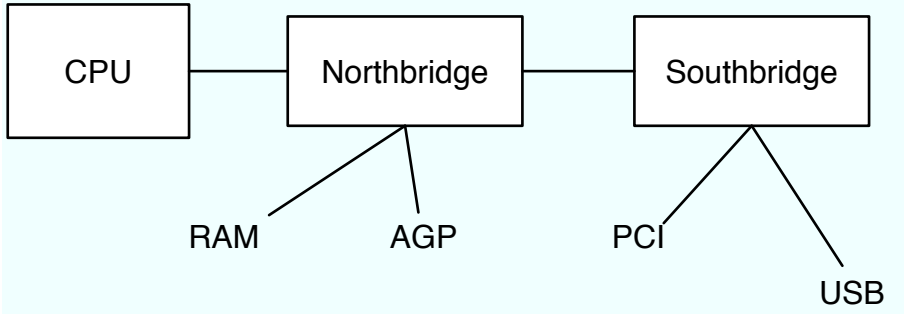


Figure 9.3: A CPU with Northbridge and Southbridge Controllers

One downside to the ISA bus and the expansion cards dealt with configuration. Almost every card had jumpers which had to be individually configured to set things like the IRQ line number and IO ports. This led to possible conflicts with other cards already installed in the system.

Over time, more problems with the ISA bus became apparent. The fixed bus speed started lagging behind the increasing clock rates that new processors were starting to run at. The manual configuration via jumper switches only allowed a limited number of settings, which had to be fixed in the hardware. These weren't problems at first; they only grew out of the increasing number of manufacturers of PCs and peripherals.

MCA

IBM took note of the problems starting to plague the ISA bus type. They also took notice of how their market share had eroded in the PC market due to the open design of their motherboards and the ISA. To combat this, they created a new bus type known as the Micro Channel Architecture. They would retain the rights to this bus style via patents and license the rights to create peripherals using it.

MCA was designed as a 32 bit bus, but allowed for a 16 bit mode as well. The clock rate was set to 10MHz, but also now used an individual bus controller chip to handle bus communications instead of relying directly on the processor. Cards were also allowed to *bus-master*, meaning they could talk with other cards on the bus without the need of the processor.

Amongst other improvements, the MCA also added support for something known as POS, or *Portable Option Select* which allowed for settings to be configured in software instead of in hardware.

Save for IBMs own machines, MCA never became widely adopted.

EISA

The industry responded with the *Enhanced* Industry Standard Architecture. This expanded bus was a 32-bit design, and supported bus-mastering as well. It also had the same shape as previous ISA cards, meaning individual slots could support older style ISA cards as well.

Though the technical aspects weren't as great as the MCA, the EISA was more widely adopted by PC manufacturers.

PCI

In 1990, Intel started work on the next generation bus style, the Peripheral Component Interconnect. It took some time to gain steam, but by the mid 1990s PCI became the industry standard alongside EISA. As of 2006, it is still the industry standard.

PCI supported *plug-and-play*, meaning that software was used to handle the configuration of resources. It supported both 32-bit and 64-bit peripherals. The clock rate was 33.33MHz, and allowed for the use of 3.3V or 5V signals.

Some variants to PCI exist, including PCI-X which has higher data transfer rates.

AGP

The PCI expansion bus was well suited for general expansion cards. But as displayed screen graphics became more utilized with the invention of the graphical user interface and multitasking operating systems, the needs for display cards rose. PCI was not able to cope with these speed requirements.

In the late 1990s, Intel released the Accelerated Graphics Port. AGP provided a direct link to the processor for a graphic card, making it a superior choice over PCI. The AGP bus is a 32-bit bus with a 66 MHz clock, providing data transfer of 254 MB/sec. But newer versions with increased data transfer rates can achieve up to 1018MB/sec.

PCI Express

The latest entry into the expansion bus market, released in late 2004, is PCI Express, sometimes known as PCX or PCI-e. PCI Express uses serial data transfer instead of parallel data transfer like all previously discussed bus types. PCI Express devices don't rely on a hardware bus, but instead are connected via a star like network, similiar to ethernet.

PCI Express is designed around a bi-directional "lane" system, where serial data travels down 1 wire. This can be contrasted to the 32-bit PCI bus system where all devices shared 32 wires. All PCI Express devices must support at least 1 lane, though they can support more for greater data transfer rates.

9.3 RAM

RAM, or Random Access Memory, is the one of the most important parts of the whole computer system. So named because data can be accessed at any location at any time (unlike earlier sequential data types, such as magnetic tapes, in which data had to be accessed in order). RAM provides the processor and other components with storage for data. This storage is designed to be short-term, most notably because after the system powers off the data is not saved.

RAM is normally in the form of integrated circuits, built into small cards or *sticks*. These sticks can be placed on the motherboard in slots, similiar to how expansion cards are placed on the motherboard.

There are two main types of RAM used today, static and dynamic.

Static RAM

Static RAM, or SRAM, is a form of semiconductor memory. This type of RAM retains its value as long as power is applied.

Every memory bit in static RAM is stored within four transistors, forming two *cross coupled inverters*. Two additional transistors are used to control reading and writing of the value. So, one bit of SRAM requires 6 transitors to implement as shown in Figure 9.4, on the next page.

SRAM operates in three different modes: standby, read, and write.

Standby

If the data stored in the RAM cell is currently not being used, the RAM is in standby mode. The inner four transistors, being cross coupled inverters, store the value indefinitely.

Read

To read the value stored in the RAM cell we simply utilizes the two access transistors on the outer part of the circuit. Enabling these transistors causes the value stored in the RAM cell to appear on each of the bit lines. Note that the bit lines are complementary: a HIGH value on one bit line means a LOW value on the other bit line.

Write

To write a value to the RAM cell, two access transistors are used to tie the bit lines to the inner RAM cell. In this case there are two options. We can write the value that is already being stored in the RAM cell, in

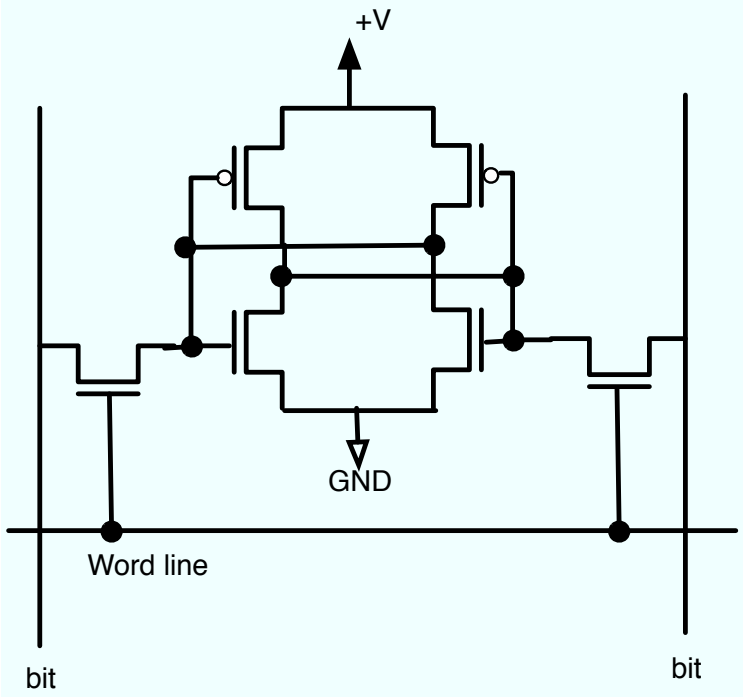


Figure 9.4: Static Ram

which case nothing changes. Or, we can write the opposite value and change the state of the cross-coupled inverter.

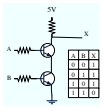
Dynamic RAM

Dynamic RAM, or DRAM, stores each bit of data as charge in a capacitor. However, since capacitors tend to slowly discharge, this value must be constantly refreshed. Because of this refresh requirement, the memory is considered dynamic.

DRAM is easier to implement than SRAM, because it requires only one transistor (as a switch) and one capacitor per bit.

9.4 System Clock

In digital electronics, the clock is used to synchronize the actions of circuits. In this regard, the clock is important because it allows the ability to consistently perform operations at a set time. For example, if



The Buzz...

The scoop on quartz

Quartz is the most common mineral found on Earth. It's found in almost every rock type and is frequently the most common component within a rock. It is composed of a crystal structure of silica (Silicon Dioxide, or SO_2). Major varieties of quartz include amethyst, onyx, and jasper.

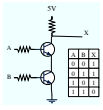
Because of its commonality, quartz is relatively inexpensive. Since it oscillates stably and predictably, it makes a great source for a time base for electronic circuits.

Figure 9.5: A crystal oscillator on a circuit board

some data is required at the input to a circuit, the clock signal can be used to trigger the circuit as to when the data is available.

Clock Generation

Most motherboards generate clock signals by using a *crystal oscillator*. In most cases this is built using a small piece of quartz crystal. When packaged properly, a small piece of quartz exhibits a property known as *piezoelectrity*. A piezoelectric material bends slightly when exposed to an electric field. In addition, a piezoelectric material will produce a voltage (and thus an electric field) when bended slightly.



The Buzz...

Understanding Resonance

Picture a child on a swing set going back and forth. The child is moving at a certain frequency. If we were pushing the child to keep them going, we have to push them at a certain part of the swing in order to do the best amount of work. That is, we tend to push them just as they start moving downward again. If we changed how we pushed the child — for example, if we pushed them every couple of seconds regardless of where they were located—our pushes do not provide the same amount of energy transfer as before.

The idea of resonance is that mechanical and electrical systems tend to absorb the most energy if the frequency of the input matches the natural frequency of the system. All mechanical and electrical systems have natural frequencies that they tend to want to oscillate at, based on their size and composition (or in the case of electrical systems, their components). The swinging child's natural frequency is based on the length of the ropes of the swing, the weight of the child, and environmental factors like wind speed.

The crystal, shown in its metallic packaging in Figure 9.5, on the preceding page, is used in a small circuit known as a *crystal oscillator*. Initially, the crystal is given some random noise voltage, which causes some vibration in the crystal. The crystal is designed to vibrate optimally at one certain frequency, known as the *resonance* frequency. The crystal's vibration causes a small voltage to be created. This voltage is then fed back into the circuitry that caused the initial vibration.

Because the crystal tends to want to vibrate at its resonance frequency, it will generate the most output signal power at that frequency. This gets fed back into the oscillator circuitry, causing an even stronger vibration at that frequency. Eventually, all of the other noise and frequencies die out and the output of the crystal is some waveform oscillating at a known frequency.

The choice of quartz for the clock oscillator signal is important. For example, it's completely possible to build an oscillator using a resis-

tor, a capacitor, and an inductor. Picking a proper value for each will give a circuit that oscillates at a desired resonant frequency. However, these electrical components are not immune to temperature drift, where the resonance frequency may change slightly based on temperature. Quartz is less susceptible to temperature drift, though not immune. Some quartz oscillator applications may measure the temperature around the oscillator and use that information to compensate for any drift in the resonance frequency that may occur.

Quartz oscillators can be tuned to vibrate at any frequency, though a handful of standard frequencies are the most common. In digital clock applications, a oscillator tuned at 32,768 Hz may be used. This allows for some digital logic to get down to a base time of one second.

Changing Frequencies

Frequency Dividers

Sometimes its desired to divide down a clock frequency into a lower frequency. One easy way of accomplishing this is with the use of a ripple counter as discussed in Section 7.8, *Ripple Counters*, on page 124. Each progressive stage of the ripple counter effectively divides the output frequency by 2.

Frequency Multipliers

The clock frequencies that are generated in today's computers are simply too fast for use with crystal oscillators, as there is an upper limit to where crystal oscillators vibrate reliably. In these cases, we may want to generate a clock signal that is higher than the output of our oscillator. This is where the frequency multiplier comes into effect. The logic for a simple frequency multiplier is shown in Figure 9.6, on the following page.

A frequency multiplier utilizes an electronic circuit known as a VCO, or Voltage Controller Oscillator. This circuit is able to vary its output frequency in relationship to a voltage that is applied to the input.

To make a frequency multipler, the output of a VCO is fed into a frequency divider. The resulting divided frequency is compared to a reference frequency created by a crystal oscillator. If the frequencies don't match, the difference between then is fed back into the VCO. This feedback causes the output of the frequency divider to *lock on* to the frequency of the crystal oscillator, making an exact copy. Thus, the actual

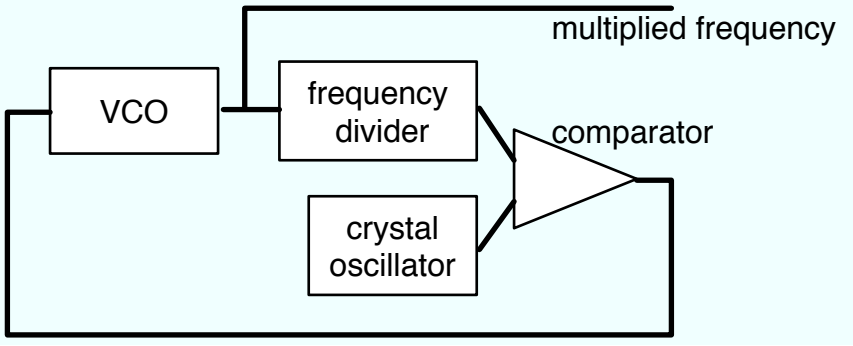


Figure 9.6: Frequency Multiplier

output of the VCO, which is on the input side of the frequency divider, is a multiplied version of the crystal oscillator.

Clock Issues

As the speeds of the clock have increased over the years, so have the problems associated with the faster clock signals.

Wavetype

Clock signals are usually associated with square waves. This is because clock circuits are designed to trigger on the transition between the digital state changes, so a crisp, fast state change like that in a square wave is desirable.

For many years, the use of a square wave as a clock signal was feasible because the clock speeds were low, relatively speaking. But as we have seen from our discussion of the Fourier Series, a square wave is simply the sum of sine wave harmonics. This means that a given square wave has many higher frequency components. Unfortunately, these higher frequency components can generate electromagnetic interference that causes problems with other parts of the circuits.

Transistor Switching

As the clock cycles back and forth, transistors within the integrated circuit transition between states. Transistors like FETs dissipate power during these transistions. As clock frequencies get higher and the transistors have to switch on and off more rapidly, the amount of power

that dissipates increases dramatically. Heat issues with higher clock speeds can be a real issue.

Data Transmission

Internal mechanisms that use clock signals typically do so to operate on some form of data. Before the clock triggers we need some assurance that the data is already at the inputs. As the clock frequency gets faster, this makes it more difficult to ensure that the data is indeed already available.

Furthermore, electrical signals travel at a finite speed. A transitioning clock signal at one end of a wire takes some time to reach the other end of the wire. The timing issues involved with this become much more apparent at high clock speeds.

9.5 BIOS

BIOS stands for Basic Input/Output System. The BIOS is a small piece of software that is executed when the computer is first powered on. Its main function is to make the system ready for the operating system to take control.

Historically, the BIOS was stored on a ROM chip connected to the motherboard. However, since ROM chips are only programmable once, this meant that the BIOS software could never be changed without physically removing and changing the chip containing the program. As personal computers became more complex the need for an evolving BIOS became apparent. By the early 1990s, the system BIOS was distributed on EEPROM devices, which have the capability to be reprogrammed.

The BIOS connects to nonvolatile memory that contains internal settings for the computer. Historically these settings were contained within CMOS, which sometimes led to the terms BIOS and CMOS to be used interchangeably. Storing these settings in CMOS, however, had one major drawback. It required a small, but always active power in order to retain the settings. This was generally accomplished by the placing of a small battery on the motherboard that provided power to maintain the BIOS settings in the CMOS even when the system was off. Today, due to the use of EEPROM and Flash technology which can store data even when power is removed, the internal battery (which is shown in Figure 9.7, on the next page) is only used to maintain the internal real-time clock of the system.

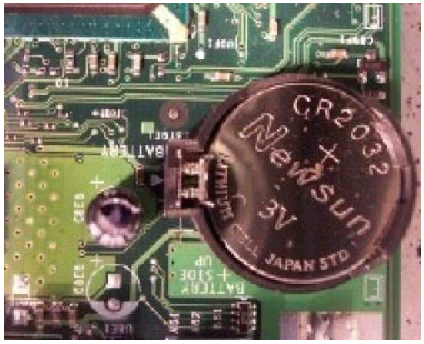


Figure 9.7: The battery on a motherboard

9.6 Other Devices

IDE

IDE, or Integrated Drive Electronics, is a connection type found on motherboards that is used to connect to internal peripheral devices like hard disk drives, CD and DVD ROM drives, and other storage devices. There are other terms used in connection with IDE including EIDE, an enchanced version that was released later, as well as ATA, the Advanced Technology Attachment. As more devices became available for the ATA/IDE interface, an extension known as ATAPI (Advanced Technology Attachment Packet Interface) also became known. Today, all of these terms mean roughly the same thing.

Cabling

ATA, or Parallel ATA (PATA) as it is sometimes now called, connects to devices via 40 pin ribbon cables. Each cable connects from the mother board and runs to one or two devices known as a master/slave configuration. There are 16 data pins available on the 40-pin cable, so data is transferred 16 bits at a time.

The 40 pin ribbon cable was suitable for many years, but recently has cause a bit of an issue for designers.

- Flat cables such as ribbon cables are much more susceptible to electrical noise from their surroundings.
- The cables have short maximum length specifications (up to 18 inches) and make it difficult to connect devices that aren't very

close to the motherboard.

- The large flat cables can get in the way of circulating air within the computer chassis and create issues when it comes to cooling.
- Most cables have connectors to allow two peripherals to be connected. The signaling aspects of having one or two devices connected change as the termination of the cable changes, causing electromagnetic reflection of signals. This can affect data transfer rates.

To overcome some of these issues, manufacturers have released an 80 pin ribbon cable version that utilizes ground pins in between signal pins to help with noise issues. They've also made some changes to the layouts of where the various connectors lie on the cable and how the end devices become connected. This has allowed higher data transfer rates for newer style devices.

SATA

A more recent development, SATA or Serial ATA, was introduced in 2003. It utilizes a different style of connector than PATA opting instead of a 7 pin connector. SATA devices transit data serially using differential signalling at fast clock speeds.

Part III

Peripheral Technology

Who is General Failure and why is he reading my hard disk?

► Author Unknown

Chapter 10

Data Storage

In the early days of computing, computers did not retain programs between power cycles. Each time the computer was powered up, the program had to be re-entered. Obviously, this repetition caused people to start looking for ways of storing programs so that they could be reused much more easily.

The first such storage medium was paper, made famous by the use of punch cards. A light based sensor like a photodiode was used to search for the absence or presence of light on the punch card. These punch cards were a breakthrough in data storage, but were still non-ideal to use as they didn't allow much margin for error. Any mistake on the card required the re-creation of an entirely new card.

Magnetic tape was the next large breakthrough in data storage. Data bits could be stored in a magnetic field on a small piece of tape—and better yet, the tape was reuseable. Early offerings in the magnetic tape world were large tape reels used in mainframe applications. Eventually, personal computers like the TRS80 and Apple II began offering cassette tape storage on the same medium as audio cassette tapes of the era. A major downside of cassette tapes was that the data was not available in a random access formation but instead had to be accessed sequentially by moving through the tape to get to where the data was stored. For large data sets, this typically resulted in a lot of jumping between locations on the tape while the end user simply sat idle, waiting for their data to be read.

Soon, manufacturers began offering the *floppy* disk, a removable magnetic storage medium like a cassette tape. The difference is that the data was stored on a large flat magnetic surface and the read/write

heads could move across the disk to where the data was stored. This allowed for random data access and was a large improvement over cassette based systems. Original floppy disks were 8 inches square but a smaller sized 5.25 inch version became much more widely used. Eventually a more durable 3.5 inch version became the preferred choice in floppy disks.

Floppy disks were widely popular because they allowed computer programs to be portable. As manufacturing costs of floppy disks declined, they became commodity items allowing many computer users to own hundreds or even thousands of them. Floppy disks regularly failed, but their inexpensive cost meant they were simply "throw away" items.

As time progressed, the data storage needs of computer users grew beyond the physical capabilities of the floppy. Today, few PC manufacturers even offer floppy disk drives as a standard item on their computers. However, another magnetic storage medium that evolved along side of the floppy disk continues to be used today.

10.1 Hard Disk Drives

The original hard disk drive was offered in the 1950s by IBM as a storage solution for large collections of data (5 megabytes at the time). In contrast, today's hard drives boast storage spaces of 300GB or more, within in a considerably smaller package size.

Data Storage

Inside of a hard drive are round metal disks called *platters*. Most drives have at least two platters, though some have as many as 8. These platters are generally made of aluminum or, more recently ceramic, and are coated with with a magnetized compound. This magnetic compound is where the data gets stored. Each platter can store data on both sides.

The magnetized compound sprayed on the platters is a *ferromagnetic* compound. Because of its structure, the magnetic dipoles in a ferromagnetic compound are easily influenced by an external magnetic field—an effect known as paramagnetism. The magnetic dipoles of the compound will align themselves with the applied magnetic field. The dipoles tend to stay in this aligned state, even after the external magnetic field is removed.

The basic unit of data storage on a hard drive is the *sector*.

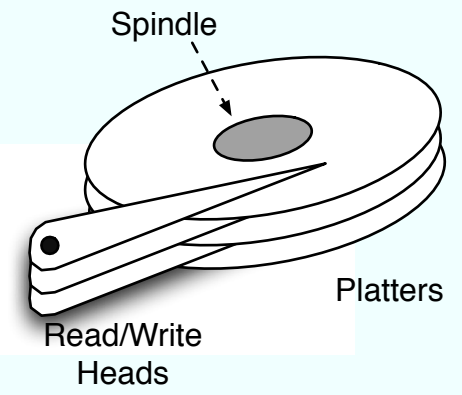


Figure 10.1: CD Velocity

The platter is formed of multiple sectors that make up the geometry of the platter. Within each of these sectors are very small regions of magnetized material which make up the individual bits stored on the platter. The magnetized surface on the platter is made of small pieces of magnetizable material known as grains, and these grains respond to external magnetic fields that are imposed on them.

Data is stored by magnetizing the grains to "point" in the same direction within each small region of the platter. The grains generate a small magnetic field around themselves and this field is sensed by a read-head. Actually, for robustness, the data is encoded by a change in direction between the magnetic fields between small regions.

The Spindle

The platters are stacked vertically and separated by a small gap. In the middle, each platter attaches to a *spindle* which is attached to a spindle motor.

Read and Write Heads

Data gets on and off of the platter through a *head*, which is a small arm that can swipe across the surface of the platter, much like a needle on a record player. There is one head for each side of each platter. When the platters are spinning, the motion creates air pressure which lifts the heads off of the platter, so the heads do not physically touch the

platters during spinning. When not spinning, the heads rest in a parked position.

Data is written to the platter by the creation of a magnetic field in the write head. This magnetic field causes the ferromagnetic material on the platter to magnetize. During reads, the head moves past the ferromagnetic material on the platter and the magnetism stored in the material on the platter generates a small current in the head.

In older drives, the read and write functions were combined onto the same head by a small U shaped piece of magnetizable material wrapped in a coil of wire. More recently, a separate read head as been has been created which uses a small piece of material that is *magnetorestrictive*, meaning that its electrical resistance characteristic changes in the presence of a magnetic field. This advance has paved the way for the large hard drive densities we have today.

Servo Motor

The heads attach to an actuator, which moves them across the radius of the platter. The movement of the heads is controlled by a *servo motor*. A servo is a special purpose DC motor that is designed for precise control, usually for either position or speed. Servo motors are generally used where the application requires very rapid acceleration and deceleration. The motor design allows for intermittent current many times higher than the continuous current rating, meaning that high amounts of torque can be generated for very short periods of time. Other design factors, such as a lightweight rotor and a short, fat shape mean that most servo motors must be small in size, not much bigger than a few inches.

10.2 Optical Disk Drives

Compact Discs

A CD is a round piece of clear polycarbonate plastic 1.20 millimeters thick and 120 millimeters in diameter. A 15mm hole in the center of the CD is generally used by a clamp on the spinning motor to attach to the CD.

CDs are made through a process known as injection molding. The mold is rigid, and forms *pits* into the polycarbonate layer. Afterwards, a thin layer of aluminum is sprayed onto the disk over the formed pits. The aluminum is then coated in a thin layer of acrylic lacquer for protection.

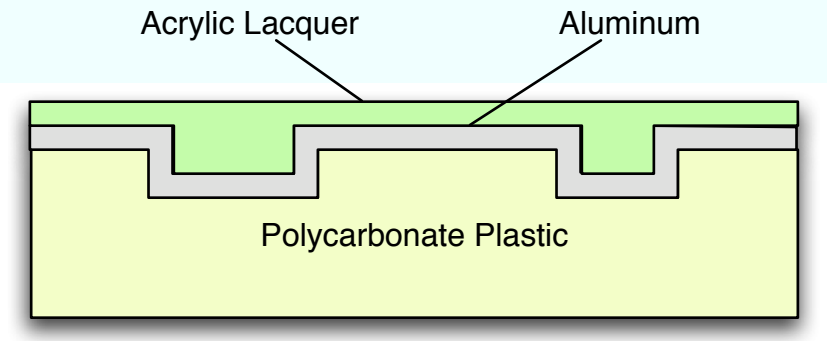
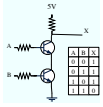


Figure 10.2: A CD Cross Section



The Buzz...

The Buzz on CD Making

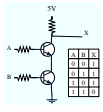
CDs are manufactured in mass quantity from a master source, which is created by a process known as *mastering*. In mastering, a very smooth glass substrate goes through a process that creates a master copy of the CD. From this, negatives known as *nickel stampers* are created which are inserted into injection mold machines. In the machine, hot plastic is injected into the nickel stamper negative at very high pressure and cools into the shape of the final CD.

The peaks and valleys impressed on the CD make up a single spiral data track. The width of the data track is 0.5 microns, which is 0.0005 millimeters. The spacing between the track spiral data is 1.6 microns

Digital Versatile Discs

DVDs are made in a very similar fashion to CDs. They have the same relative shape. The data is also encoded as pits in the polycarbonate layer.

However, DVDs are able to store more information than CDs because they have a smaller track pitch (0.74 microns) and data width (0.32



The Buzz...

Funny Shaped CDs

Data on CDs is stored from the inside spiraling outward. Because of this, it's possible to make CDs that are not 120mm in diameter.

microns). This alone accounts for an increase of about 7 times the storage capacity of a traditional CD.

In addition, DVDs can be double sided. It's also possible to have two layers of data on a single side of a DVD. In this case, below the layer of aluminum pits lies a separate layer of gold pits. The reading laser is able to focus on each of the layer individually, resulting in a DVD that can have twice as much data encoded upon it.

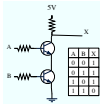
Optical Drives

The job of the optical CD drive is to focus a laser onto the CD and receive a reflection back. This happen while the CD is spinning, faciliting the need for a motor to turn the CD. Some electronics are also needed that can move the optical components radially along the CD's diameter to track the data as the CD spins.

CD-ROM drives come with a speed rating, which is relative to that of a music CD. For example, a drive rated as 1x is able to read data at 150 kilobytes per second. 2x drives can read data at 300 kilobytes per second. Drives are capable of up to about 12x data rate, after which vibration due to excessive speed becomes an issue. Above this speed, some tricks are employed as as the use of special (and more expensive) ball bearings to balance the disk when it spins. The practical limit for CD data rate is about 52x, above which it either becomes too expensive for higher data transfer or the polycarbonate isn't strong enough to withstand the high velocities.

CD Burners

In general, mass manufactured CDs are made from the mastering process discussed earlier. Another CD making process is possible within your own computer using a CD burner.



The Buzz...

CAV vs. CLV

When a circular body (like a CD) spins, its rotation is measured in an *angular velocity*. The read head of the CD is designed to read data at approximately 1.2 meters per second, which is a *linear velocity*. To accomplish this, the servo motor must alter its velocity depending on where the read laser is housed (see Figure 10.3). The CD must rotate at a slower speed as the head moves outward. This is known as *constant linear velocity*.

Hard disk drives rotate at a constant angular velocity.

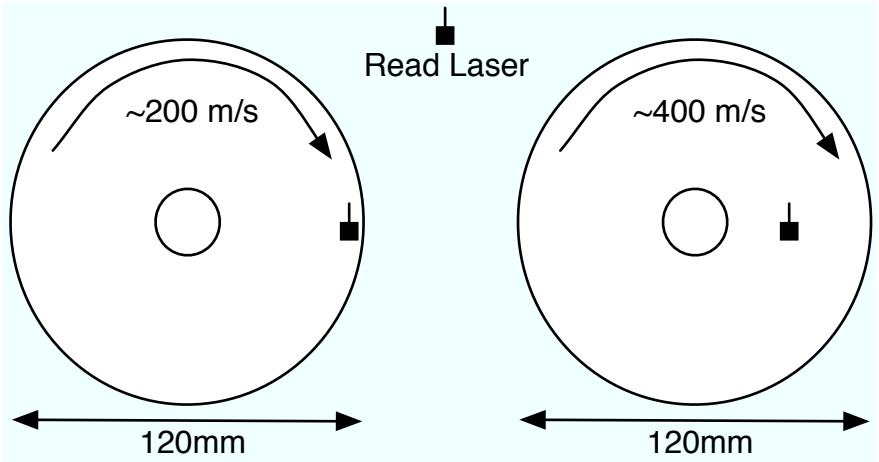


Figure 10.3: CD Velocity

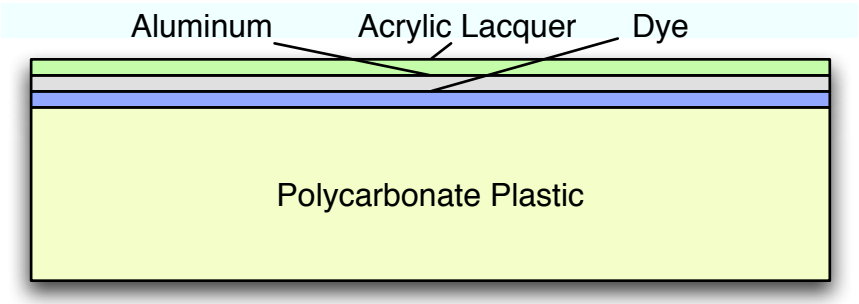


Figure 10.4: A CD-R Cross Section

Recordable CDs (known as CD-Rs) have a different composition than their mass produced master CD counter parts. On a blank CD-R, there are no pits of data. Instead, the aluminum layer is completely smooth. Between the aluminum surface and polycarbonate material is a layer of special dye. In its normal state, the dye is transparent. However, when heated using a special laser the dye becomes opaque.

A CD burner, then, contains a strong *write laser* that is capable of focusing on the dye and turning certain portions of it opaque. This process creates the same effect as the aluminum pits in regular CDs.

One downside to this process is that the discs are not reusable. However, CD-RW (re-writable) discs do exist.

CD-RW Burners

Use a phase change medium, which melts into a liquid at 600 degrees and crystalizes into a solid at 200 degrees.

The phase of the medium locks into place after cooling

A special erase layer

Power calibration area (PCA)

Lasers

CDs are read using a 780nm wavelength laser. The laser shines through the polycarbonate layer onto the aluminum pits (which from this side are now valleys). See Figure 10.5, on the next page for a graphic show-

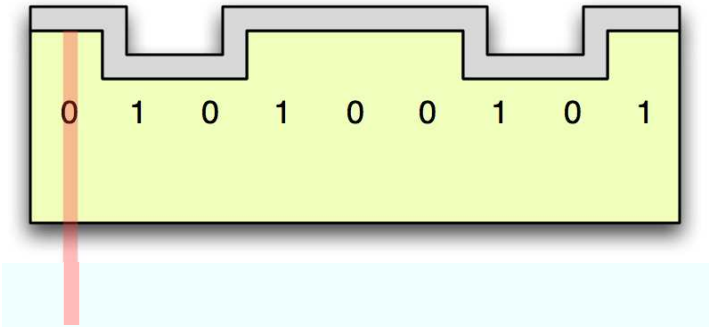


Figure 10.5: A Laser reading CD data

ing how the laser shines through the polycarbonate and onto the aluminum.

As the CD data spins past the laser, the reflection of the light is focused into a photo diode, which is able to register a digital signal based on the intensity of the reflected light. Because the valleys and peaks are at different distances from each other relative to the laser, the reflection of the laser light hits the photodiode different based on whether a peak or a valley is currently in the path of the laser.

Pits and lands do not directly represent a logical 0 or 1. Instead, the change between a pits and land (or land and pit) represents a logical 1. No change between successive pits or lands represents a logical 0.

Data Encoding

Data on CDs is encoded in a variety of special formats. Foremost, all CDs use of *Eight-to-Fourteen* modulation. In this modulation scheme, all 8 bit data structures are transformed into 14 bits of data for encoding on the CD, as shown in Figure 10.7, on page 162.

The reason for using 8-14 modulation is to increase the continuous amount of pits and lands on the CD to allow for better tracking. Without 8-14 modulation, if there were a large number of 0s in a row (represented by a continuously smooth area) on a CD it would be almost impossible for the tracking system to stay fixed on the appropriate data track. 8-14 modulation insures that every logical 1 is separated by at least two zeros, but no more than ten.

Between each 14 bit modulated piece of data is a 3-bit *merging word*. So, a single byte of data is encoded as 17 bits of data onto the CD.

Data Storage

Data on a CD is comprised of 33-byte frames. Each frame has 24 bytes of data, 8 bytes of error correction, and 1 byte of *subcode* data. This means that there are 561 (33-byte frames * 17 modulated bits per byte) bits in each data frame. Another 27 bits are added for synchronization, bringing the total number of bits to 588. Demodulated back to the 8 bit world, this represents 42 8-bit bytes of data.

Each 42 byte frame contains only 24 bytes of actual data. On audio CDs this is represented by six audio samples. Recall that there are two stereo channels and each sample is 16 bits (or 2 bytes). So, (6 audio samples) * (2 channels) * (2 bytes of data) = 24 total bytes of data.

On CD-ROMs, these 24 bytes are used for data storage. Ninety-eight frames of data are put together into *sectors*, containing 2352 bytes of data. Across these 2352 bytes of data are stored:

- 12 bytes: synchronization
- 4 bytes: sector ID
- 2048 bytes: data
- 4 bytes: error correction
- 8 bytes: null data
- 276 bytes: error correction

Subcode Data

Each 33 byte data frame is comprised of a 1 byte subcode data.

10.3 Flash Drives

The first flash drives were invented around 1998 by IBM. They were intended for use as a replacement to floppy disk drives and have quickly become a popular choice for smaller data storage application. They are relatively inexpensive, extremely robust, and very compact. Small scale drives are available as “keys” and come with a USB interface for quick attaching and detaching to a computer. Large scale drives, in similar form factors to magnetic hard drives, are also available though are currently much more expensive than their magnetic counterparts.

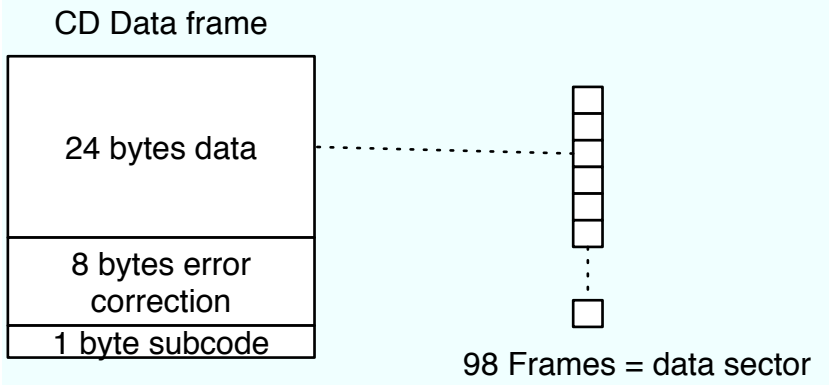


Figure 10.6: A CD data frame

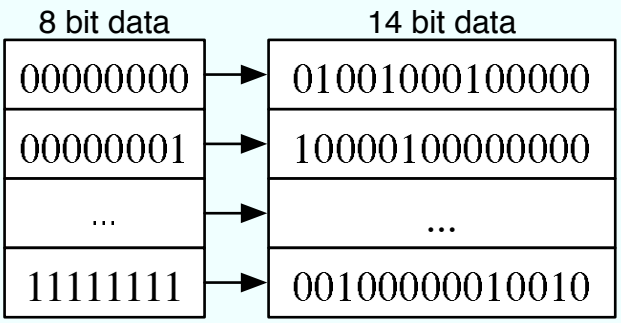


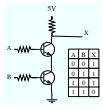
Figure 10.7: Eight-to-Fourteen Modulation

Flash Memory

At the heart of the flash drive is flash memory. It was invented by Fujio Masuoka in 1984 at Toshiba, with Intel later introducing the first commercial version in 1988.

The Floating Gate Transistor

The heart of flash memory is the floating-gate transistor, shown in Figure 10.8, on page 164. A floating gate transistor is constructed similar to a MOSFET, but has an added layer of poly-silicon material on the gate of the transistor, below the normal oxide layer present in MOSFETs. Because the floating gate is insulated by the oxide layer, any



The Buzz...

The Rainbow Books

Specifications for optical media are available in a set of colored books commonly known as the Rainbow Books. For example, the Red Book contains the standards specifications for audio CDs and was originally published in 1980. The book houses the physical specifications for CDs and includes specifics regarding encoding. The Yellow Book defines the format for CD-ROMs.

accumulated charge that gets to the floating gate does not leak even after the power is removed.

In the normal state, where no charge is accumulated on the floating gate, the transistor is said to be in its *1* state. If charge has accumulated in the floating gate, the transistor is said to be in its *0* state.

The floating gate transistor value can be written, read, or erased. The writing process involves putting a high voltage on the control gate of the transistor and a negative voltage at the drain to coerce electrons to flow and collect at the floating gate. The erasing process is basically a reversal of the writing process.

With electrons accumulated (or not accumulated) at the floating gate, the characteristics of the transistor in its normal usage are altered. Without the accumulation the transistor operation normally causes a logical *1* to be read. However, with the accumulation the transistor output is now opposite, and as such a logical *0* is read.

Flash Memory Limitations

Flash memory can be read and programmed one byte at a time, but erasure happens across an entire block. That is, after programming any one bit in order to change its state the entire block it resides in must be erased. This poses some speed constraints for certain applications of flash memory.

Like all media, flash memory also has a limited life. Typically, flash memory cells are designed to withstand over one million write cycles.

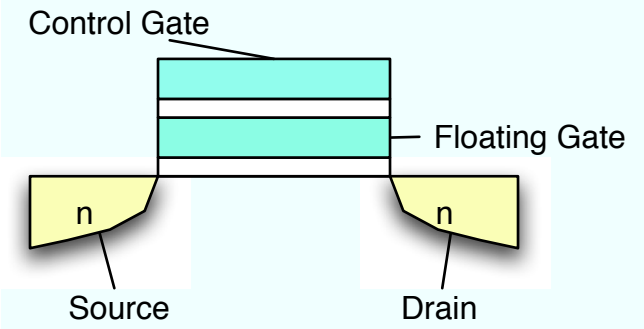


Figure 10.8: A floating gate memory cell

For uses such as USB memory sticks, the one million write cycle lifetime is generally acceptable. However, in hard drive applications where data may be written and rewritten thousands of times in a day, flash memory's write cycle lifetime can be a hindrance.

Chapter 11

Networking

The history of computer networking dates back to the telegraph era of the late 1800s. A stock ticker machine developed in the 1870s allowed the printing of stock quotes over a pair of wires from a long distance using telegraph wires. Later, a device known as a teletype which was a large electro-mechanical typewriter, came into popularity. This device allowed an operator to punch in a message using a special coding scheme on a 5-key keyboard known as the Baudot code. Between the sending and receiving teletype machines was a DC voltage on a wire that would be periodically opened, or interrupted, to indicate the presence of data.

As the mainframe computer grew into existence, so did the need to connect multiple computers together. Early network designers had to incorporate both hardware design as well as software that could utilize the hardware into their systems. Each computer on the network would have to utilize the same type of hardware and software in order to facilitate communications. Without the presence of any standards, individual computer networks were not able to communicate with each other unless they used the same hardware and software.

Research into building formal computerized network was led by the US Defense Department's Advanced Research Projects Agency, which created ARPANet, a precusor to today's Internet. But many other groups worked on computer networks, as well. Some were academic, such as universities that worked on building time sharing systems for their large mainframe computers. Others focused on integration with the telephone company.

In this chapter we look at some history of computer networks and their implementation. The focus of this chapter is on the technology and physical implementation of the networks as it pertains to electronics. In order to maintain this focus, we unfortunately may skip over some interesting aspects of general networking.

11.1 Modems

The word modem is short for MOdulator DEModulator. The main concept is a device that is capable of modulating, or changing, digital information into a signal which can be transmitted via some transmission medium and can then be demodulated back to its original form.

In the computer world, we are most familiar with the modem as an interface to the telephone line. In this case, the job of the modem is to take digital information from the computer and convert it to a signal which can be transmitted over the telephone line and decoded back by a modem on the other end.

POTS Modem History

The POTS (Plain Old Telephone System) was, for many years, controlled by A.T.&T. As computers became popular and the desire to send data over distances started to grow, A.T.&T. began introducing a series of devices that were capable of sending computer data over the telephone network. A.T.&T. maintained a monopoly on these types of devices.

The monopoly, however, was only for devices that were directly (read: electrically) connected to their lines. After a Supreme Court decision in 1968 breaking the A.T.&T. monopoly on certain transmission types, the market quickly opened up to manufacturers who made devices that could send acoustic data through a traditional phone handset.

Throughout the 1970s, manufacturers (including A.T.&T. itself) started the boom in practical data modems. At this time, data transfer rates were 300-1200 bits per seconds (bps).

In the early 1980s, Hayes Communications introduced the *Smartmodem*, which was a normal 300 bps modem but included an integrated circuit that allowed the computer to send commands to the modem that would also directly control the phone line operation. Previously, modem calls were initiated by the sender lifting a handset, dialing a number, then setting the handset into an acoustic coupler. Hayes SmartModem integrated all of these features into the modem itself.

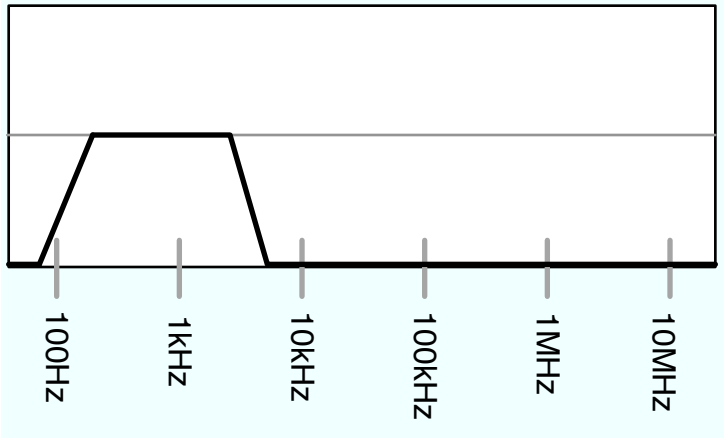


Figure 11.1: The frequency response of the POTS network

Over time, data transfer rates increased, to 2400 bps, then to 9600 bps. Today 56 kbps is most common.

POTS Theory of Operation

The main limitation of any modem lies in the limitations of the transmission medium. In this case, the POTS network main limitation is the transmittion frequency range which is 300Hz to 3400Hz. Any electrical signals that are transmitted over the telephone line must be in this frequency range. The telephone company uses electronic filters to limit transmitted frequencies inside this range. Thus, any frequencies outside of this rage are attenuated. The frequency response of the POTS network is shown in Figure 11.1.

The reason for this electric filtering is simple: the POTS network was designed to carry voice conversations. On average, the human voice creates sound wave between 80Hz and 6000Hz. Enough information conveyed in the human voice is available between 300Hz and 3400Hz — at least, within this range there is enough information to understand the speaker. By limiting the range of frequencies which are transmitted, the phone company is able to fit multiple conversations into a larger data stream. Allowing a larger range of frequencies would make the speaker more intelligible, but would reduce the number of conversations capable of being carried on a single wire.

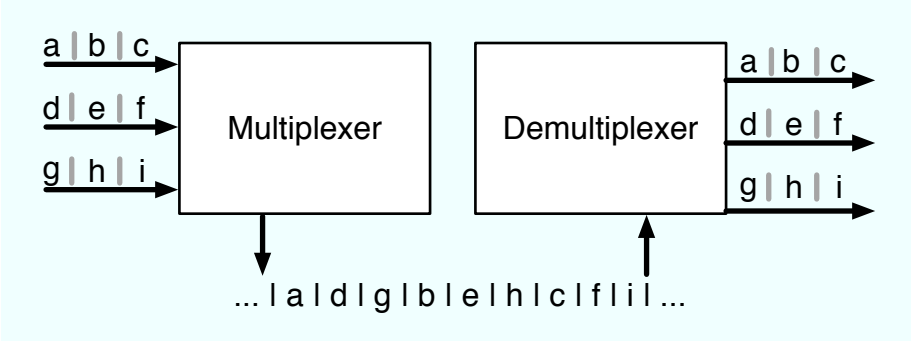
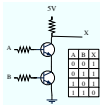


Figure 11.2: Time Division Multiplexing



The Buzz...

Digitizing an analog medium

You may wonder why a modem simply doesn't create a digital signal and send it over the telephone line. Recall from Section 4.1, *Frequency Response*, on page 73 that a square wave signal is simply an additive series of frequencies. The telephone company has filters in place which only allow certain frequencies to pass, namely between 300Hz and 3400Hz. This means that it's not possible to send a digital waveform like a square wave over a telephone line. The filtering that takes place would greatly distort the signal that is received on the other end.

Time Division Multiplexing

The POTS network today uses a technique known as Time Division Multiplexing to transmit the (voice) data. TDM is a technique for transmitting multiple simultaneous channels over a single medium (wire, in this case). A single POTS conversation is composed of 8000 samples per second, each of eight bits. This means the data transmits at 64 kbit/s. The data is then sent to a TDM system where it is multiplexed with many other channels and transmitted in discrete time slices, before it is received on the other end and demultiplexed (see Figure 11.2).

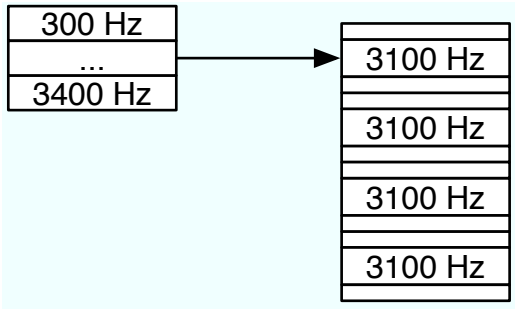


Figure 11.3: POTS Multiplexing

Frequency Shift Keying

A digital modulation technique used in early modems is Frequency Shift Keying (FSK). In FSK, two differing frequencies are used to represent a binary 0 and 1. For example, in early Bell 103 modems, the modem that made the originating call transmitted data at either 1070Hz (0) or 1270Hz (1). The answering modem transmitted data back at either 2025Hz (0) or 2225Hz (1) . These frequencies were chosen because they had the lowest amount of distortion when transmitted across the phone line and they were not harmonics of each other so intereference would be minimized.

An example of FSK data is shown in Figure 11.4, on the next page. Two discrete frequencies are transmitted over certain time intervals. In this case, a lower frequency representing a binary 0 is transmitted first followed by a higher frequency representing a binary 1.

Phase Shift Keying

Soon, a follow on digital modulation technique known as Phase Shift Keying (PSK) appeared. With PSK, the transmitted analog signal always retains the same frequency and amplitude, but will shift by a certain amount in its phase. A common method of PSK used in modems is as follows:

- 0 degrees - 00
- 90 degrees - 10
- 180 degrees - 01

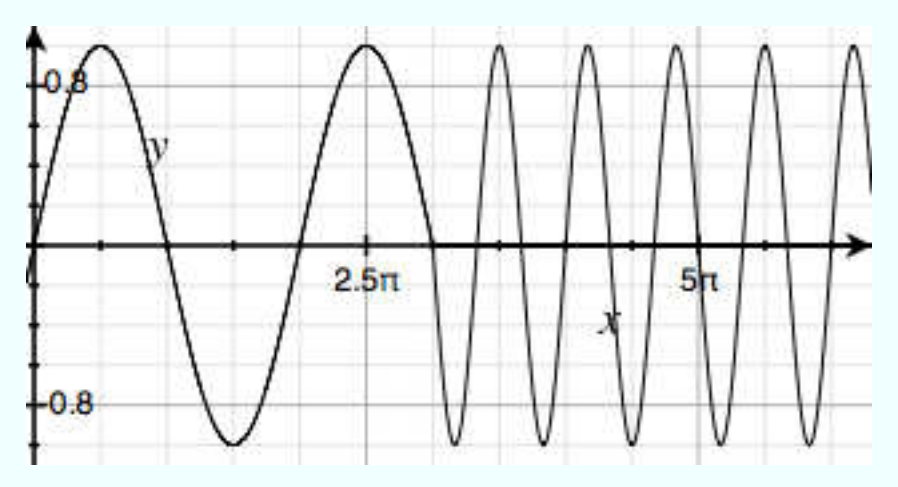


Figure 11.4: FSK Data

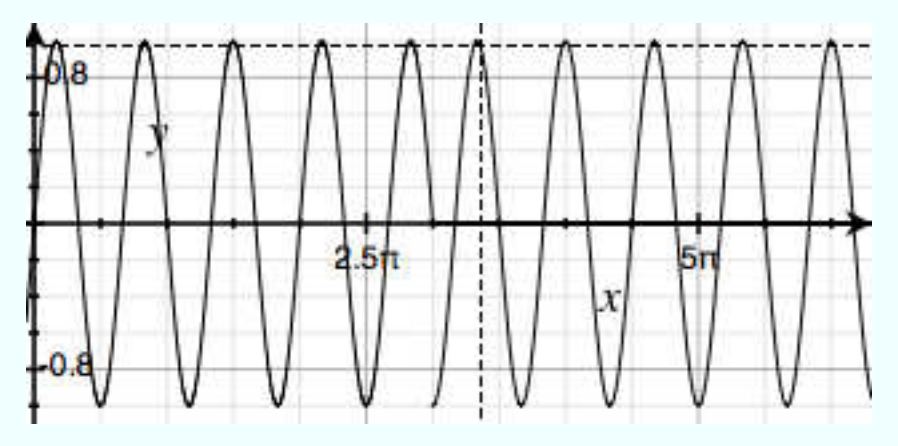
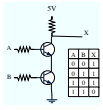


Figure 11.5: PSK Data



The Buzz...

What is a baud?

A baud is a unit of data transmission speed. It represents data transfer rate in relation to a number of events per second. For example, a 1200 baud modem is capable of sending 1200 pieces of information per second.

In general, a baud is the same as a *bit per second*, or bps. However, since modems use tricks like PSK and FSK to encode multiple pieces of data into a single piece of information, baud and bps may not be equal. For example, a modem using Quadrature PSK to transmit data at 600 baud is actually transmitting 1200 bps, because each discrete signal being transmitted encompasses two bits of actual information.

- 270 degrees - 11

This type of of PSK is known as Quadrature PSK since four phases are used. It's possible to use more (8) or less (2) phases for modulation. More phases mean that more information can be encoded, but with a reduced amount of reliability.

Transmission rates using PSK were limited to 600 baud, which meant a transmission rate of 600 samples per second with a full duplex (meaning both modems can talk at the same time). When using quadrature modulation, there are two bits encoded in each sample, so the effective transfer rate is 1200 bits per second. Achieving 1200 baud (and 2400 bps) was also possible in a half-duplex setup, meaning that only one modem could transmit at a time.

An example of PSK is shown in Figure 11.5, on the preceding page. An initial waveform with no phase (representing the bits 00) is first transmitted. At a certain time later, a 90-degree phase shifted waveform is then sent representing the bits 10. This process continues on, each new waveform representing a two bit value.

Quadrature Amplitude Modulation

By the late 1980s, modems began making use of another form of digital modulation known as Quadrature Amplitude Modulation (QAM). QAM

Bit Values	Amplitude	Phase Shift
0000	1	0
0001	2	0
0010	3	0
0011	4	0
0100	1	1/4
0101	2	1/4
0110	3	1/4
0111	4	1/4
1000	1	1/2
1001	2	1/2
1010	3	1/2
1011	4	1/2
1100	1	3/4
1101	2	3/4
1110	3	3/4
1111	4	3/4

Figure 11.6: Quadrature Amplitude Modulation Table

made use of both phase shifting and amplitude shifting. In this case, the four possible phase shifts are used just like in PSK, but in addition four possible differing amplitudes of the waveform are also used. With a 600 baud signal, it is now possible to encode four bits of data for a data transmission rate of 2400 bps.

A table showing the possible states of using QAM to encode binary data is shown in Figure 11.6.

Modem Standards

Many early modems followed formats for data transmission devised by the Bell company for their modem products. Eventually, an international organization know as the International Telecommunications Union (ITU) became involved to help standardize the practice. Before the early 1990s, this organization was known as the CCITT, based on a French acronym.

Modem communication guidelines were issued by the CCITT as part of the V series. For example, V.22 was one of the first standards used. It specified PSK modulation and a 600 baud transmission rate. Later a revision known as V.22bis was released. It specified using QAM at 600 baud which allowed for higher data transmission rates.

Echo Cancellation

A large breakthough in modem communications came with the development of echo cancellation. In voice calls, the phone system routes a small portion of the outgoing voice signal back into the earphone so the speaker hears themself talk. This is known as the *sidetone*.

In the modem world, the sidetone presents a problem because it's not possible to distinguish between whether the information is coming from the sender or the receiver. This was the main reason that the origination modem and the remote modem transmitted data at separate frequencies—the modems could simply ignore and sidetones they were not interested in.

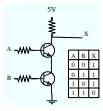
Echo cancellation technology was able to capture the sidetone, and based on its very slight delay from the original signal, calculate if the data it was receiving was actual data or the sidetone.

Once echo cancellation technology became present on the modem, it was possible to move the carrier frequency to 1800Hz, the middle of the frequency band, and have full duplex transmission. This opened up the doors for 2400 baud transmission, and depending on the digital modulation scheme would allow 4800, 9600, or even 14,400 bps data transfer.

This was all encompassed in the v.32bis standard.

Upper Limits

In the late 1990s, the V.90 and V.92 standards were introduced with the expectation that they would probably be the last necessary modem standards, as they caused modems to operate almost at the edge of the channel capacity of the POTS. V.92 specified transmission speeds of 48 kbps and receiving speeds of up to 56 kbps utilizing PCM for both upstream and downstream data transmission. A majority of the data transmission speed relies on the use of data compression before transmission by way of the V.44 standard.



The Buzz...

The T-Carrier System

Though most of us have since moved on from using modems for intra-computer communications, the remote networking that we do is still covered by systems designed by the phone company A.T.&T. (and its research branch Bell Labs). The T-Carrier system is the system used in North America (and Japan) for digital telecommunications systems.

The basic unit of the T-Carrier system is the Digital Signal 0 (DS0) which is used for one voice channel. It offers a transfer rate of 64 kbit/sec, aligning nicely with the 8-bit 8000Hz single conversation sampling used on today's telephone lines.

24 DS0s make up a DS1, sometimes known as a T1. This is the same type of T1 that businesses lease for fast internet connectivity. With 24 channels and an additional bit used for data framing, the total data transmission with a T1 line is 1.544 Mbit/sec.

The V.92 standard seems to be the upper limit of what modems using the POTS network are capable of achieving.

11.2 Local Area Networks

While modems are useful for connecting computers at great distances, computers that are close together need a way to intercommunicate as well. The history of computer networking is vast and complex, so we'll discuss the highlights and some of the more important parts of the developments.

The OSI Model for networks, as described in Section 11.3, *The OSI Model*, on page 178, explains computer network models in a 7 layer approach. This section focuses on the lowest two layers, the Data Link and Physical layers. Some examples of technologies in these layers include Token Ring, Frame Relay, ATM, and Ethernet—the latter of which is so important we devote an entire section (Section 11.5, *Ethernet*, on page 185).

Token Ring

The Token Ring network was developed by IBM in the early 1980s and became IEEE standard 802.5. After the introduction of Ethernet in the early 1990s, the popularity of Token Ring technology started to fade. Today it is found mainly on legacy systems that have been using it for many years.

The nodes of a Token Ring network are laid out like a ring. This means that node 1 is connected to node 2, which is connected to node 3, and so on.

The main concept behind technology is a passing token. Each node, in turn, transmits a special token to its closest neighbor, who then passes the token on to its closest neighbor. The token works its way around the ring. The token is used for arbitration—whoever has the token is allowed to transmit data, and data cannot be transmitted until the token is received.

The Token

As long as no data is being transmitted, an empty token is continually passed around the ring.

Manchester Encoding

Manchester encoding is a common technique that is used for digital data transmission. The key concept is that the transmitted bit information is encoded as a voltage *transition* instead of just a voltage level. For example, a 1 might be represented as a transition from low to high and 0 as a transition from high to low.

There are two big advantages to using Manchester encoding:

- There are no long periods of time without level transitions. If the transmitter wanted to send a long sequence of 1s, for example, traditional encoding could dictate that the transmitted voltage level remain high during these 1s. With Manchester encoding, the levels are always transitioning making it easier to negotiate the data.
- It is self-clocking. This makes the synchronization of the data stream easier.

The Manchester bit transition happens during the middle of the bit definition, as noted in Figure 11.7, on the following page. Because of this, it is sometimes necessary to transition the voltage level at the start of

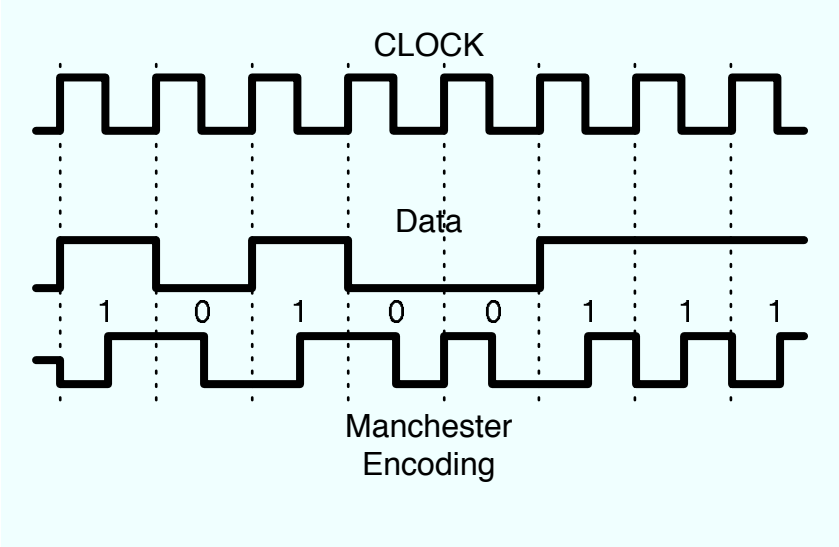


Figure 11.7: Manchester Encoding

a bit as well. The start-of-bit transmission has no impact on the transmitted data; it's simply there to make sure the logic level is properly set up for the Manchester encoded bit transition to work. Because of this overhead, the logic levels have to be able to change at approximately twice the frequency of the data transmission. This may make Manchester encoding undesirable for some forms of communications.

A figure showing an encoded signal is Figure 11.7.

ARCNET

ARCNET is a type of local area network protocol that was developed by the Datapoint Corporation in the late 1970s. It was the first networking type system that allowed expansion and flexibility.

The original system used coaxial cable wired in a star fashion meaning that multiple nodes could connect together via a single central hub.

Access to the physical network was handled by the use of a token, similar to the Token Ring network. With ARCNET, a token continuously cycles around the network. No node can use the network unless it has received the token. Once a node receives the token, it may send a message and then pass the token on to the next machine.

An advantage of ARCNET was that the transmitter received an acknowledgment to its transmission from the receiving end. This allowed for better error recovery from other systems which were based on timeouts.

The use of ARCNET eventually fell in popularity to the revised form of Ethernet and by the late 1980s and early 1990s fizzled out almost entirely.

ATM

ATM, or Asynchronous Transfer Mode, is a networking protocol that encodes data into small packets of a fixed-size. It was created with the intent of helping to bridge the gap between packet-switched networks (such as Ethernet) and circuit-switched networks (like that of the POTS). In fact, it was envisioned as a complete replacement to the POTS network and help the telephone networks integrate with private computer networks.

Many telephone companies have implemented ATM style networks for their equipment and some end users still make use of them today. However, ATM did not catch on as a widespread technology, most notably because the Internet Protocol specification, and Ethernet, became more popular alternatives.

One of ATM's design features is its small data packet size, sometimes known as a *cell*. The small cell size was specified in order to help minimize the variance in total trip time between packets. Since ATM was primarily designed as a replacement for POTS networks, the original intent was to carry voice conversations. This required evenly spaced packets to arrive at their destination, or a choppy conversation would be heard instead of a smooth fluid one.

Today, as data transfer speeds have greatly increased, ATM's small data packet size is actually more of a hinderance than a feature.

Frame Relay

Frame relay is a form of data transmission that sends data as a series of digital frames to any number of destinations. Most notably, Frame relay networks were design to connect multiple local networks together as endpoints within a *wide area network*.

Frame relay provided an alternative for business to use something other than a *dedicated* line. In this sense, a dedicated line was leased from

the telephone company and was used to directly connect two endpoints together. Instead, with frame relay, network equipment would be used in the middle to route traffic between a source and a destination. This allowed the provider of the frame relay network (the telephone company, in this instance) to provide frame relay access to multiple customers, all of which would share the cost of the network infrastructure.

While frame relay networks are still in existence, they have largely been replaced by the widespread use of Internet Protocol based networks.

11.3 The OSI Model

When the popularity of personal computers started to take off in the early 1980s, many vendors began creating their own proprietary hardware to sell to the mass market. The area of inter-networking hardware was no different.

The OSI, or Open Systems Interconnect, model was conceived to help standardize the concepts of computer networking. It allows components to work together regardless of the manufacturer.

The OSI model is made up of 7 layers:

Application Layer

As the top most layer of the OSI model, this layer is the one most visible to the end user. An example of this is the Hypertext Transfer Portocol (HTTP) which is used for web page access on the internet. This protocol specifies how the end application receives and transmits data that it is interested in, such as the use of the “GET” command to retrieve a specific piece of information. Note that nothing is specified in how to make the connections nor how data is physically routed or handled. These specification lie in lower levels with other protocols.

Presentation Layer

The presentation layer handles requests from the application layer and passes them on as requests to the next layer, the session layer. This layer handles things like encryption, which the end application may not be concerned with from a global point of view. It also handles some of the encoding aspects of the data that is transmitted and received, such as ensuring strings are encoded in the proper format for the particular application.

Session Layer

The session layer handles the conversation between the two endpoints of a networked application is handled. The handles the connection and disconnections of applications and manages the flow of data between them.

Transport Layer

The transport layer handles data transmission issues between applications. It is responsible for determining the best method of transmitting data. Two of the most common transport protocols used today are TCP and UDP.

Network Layer

The network layer handles the network aspects of routing: error control and flow control. It also handles the important aspects of translating logical destination addresses into the physical addresses of the machines at the other end. The Internet Protocol (IP) is the most well known network layer protocol used today.

Data Link Layer

The data link layer allows for the transfer of data between two nodes on the network. Ethernet is the most common data link protocol used today.

Physical Layer

The physical layer specifies the hardware and electrical aspects of the networking implementation. It provides the means for transmitting the digital bits that the data link layer uses for the transmission.

11.4 Cabling

With the exception of wireless networks, which is covered in detail in Chapter 13, *Wireless*, on page 205 all networks need some form of electrical or optical cabling for data transmission. In this section, we'll take a look at some of the options for cabling and explain reasons behind the design of those cables.

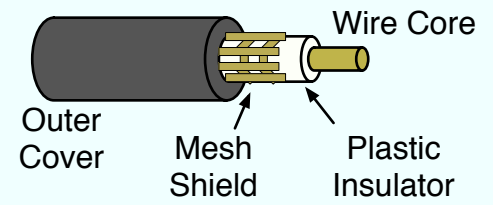


Figure 11.8: Coaxial Cable

Coaxial Cable

Coaxial cable is one of the most widely used medium for electrical signal transmission. A coaxial (or simply, coax) cable has a single copper conductor in the middle that transmits the signal. Wrapped around this conductor is a plastic layer that helps provide insulation. Around the plastic layer is a braided metal mesh that is used for shielding. Finally, an outer plastic coating envelops the whole cable. The whole diagram of a coaxial cable is shown in Figure 11.8

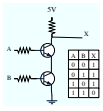
There are two advantages that coax cable has over a normal piece of wire.

- The shield - The mesh wire shielding layer protects the signal wire from receiving outside interference from external electromagnetic radiation. It also helps to confine the signal within the cable minimizing the amount of electromagnetic radiation leaving the cable.
- Flexibility - The ability to easily bend coaxial cable allows it great maneuverability in routing it to its destination.

Coaxial cable is especially useful for high frequency applications, from about 1 MHz to 3 GHz. While coax works well for frequencies below 1 MHz (all the way down to DC power), the use of the dielectric insulation between the signal and shield conductors generally means that the power losses in coax cable are higher than using other cables. For these lower frequencies, using a different type of cable may be required.

Connector

The Bayone-Neill-Concelman (BNC) connector is the most common type of connector used on coaxial cable. Different types of BNC connectors exist such as a T shaped connector, a termination connector, and the mating connector.



The Buzz...

The Faraday Cage

A Faraday cage is a metallic enclosure that protects the inside from electromagnetic fields. It's named after Michael Faraday (as first mentioned in Section 3.1, *Magnetic Motion*, on page 35) for a demonstration he performed in 1836 explaining the phenomenon.

The Faraday cage works on the principal that charge exists only on the surface of a conductor. As an electric field is applied to a conductive surface, charge accumulates on the surface of the conductor. This means that charge cannot penetrate the inside of the cage, protecting the contents from any electric and electromagnetic interferences. A side effect to the behavior of the Faraday cage is that any charge that is created inside the cage also cannot escape to the outside.

A humorous example of a Faraday cage is the tinfoil hat that some people may wear to protect their thoughts from being read via electromagnetic waves. Since electromagnetic waves are unable to penetrate through the metallic conductor, some feel their personal thoughts are well protected from potential thieves.

The BNC connector attaches to the cable by force either through a crimping tool or by the use of a screw-on connector. An example of a crimped on BNC connector is shown in Figure 11.9, on the next page.

Coaxial cable types

Coaxial cable is designed by a specification known as the RG# or Radio Guide Number. Each designation has an associated set of parameters such as diameter of the signal wire, type of shielding braid, and internal impedance number.

Today, there are two commonly used impedances of coax cable.

- 50 ohm - This type of coaxial cable is commonly used with radio transmitters. Many of the transmission antennas are 50 ohms and matched impedances are important in signal transmission. The two most common 50 ohm coax cables are RG-8 and RG-58.

Figure 11.9: A BNC connector on a coaxial cable

- 75 ohm - Commonly used with video and audio transmission systems. Through experimentation it was found that about coaxial cable with about 77 ohm provides the lowest power loss for the transmitted signal, so a 75 ohm standardization was formed. The most common 75 ohm coax cables are RG-6, RG-11, and RG-59.

Twisted Pair Wires

Twisted pair cables are another very common type of networking cable used today. The name *twisted pair* comes from the fact that pairs of wires are twisted around each other for the entire length of the wire. A graphic showing a single piece of twisted pair wiring is shown in Figure 11.10, on the following page.

Twisted pair cables can be shielded or unshielded, the latter being less expensive. Typically, twisted pair wires come in multiple sets. For example, category 5 twisted pair wire has four individual sets of twisted pair wires for a total of eight wires.

Crosstalk

The reason for the twists in the wires is to help reduce a phenomen known as *crosstalk*. Crosstalk happens when a signal transmitted on a communication wire causes some external, and undesired, effect on

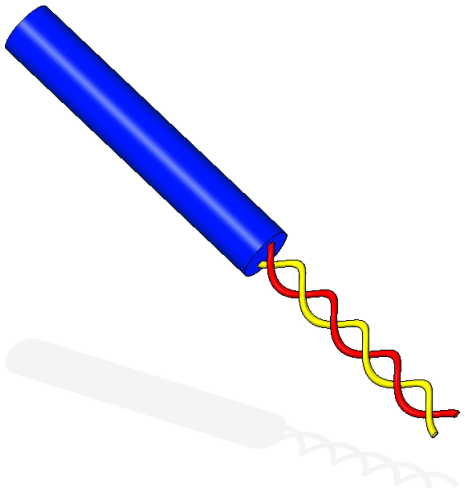


Figure 11.10: A twisted pair set of wires

another wire. The twists in the twisted pair wire help to eliminate this phenomenon.

The signals transmitted over twisted pair are commonly digital signals known as balanced differential signals. This is shown in Figure 11.11, on the next page, along with a normal 0-5V digital signal for comparison.

The use of balanced differential signals with twisted pair wires is important. Because the voltage in one wire is increasing while the voltage in the other wire is decreasing their currents flow in opposite directions and any electromagnetic fields generated by one wire due to this change counteract with the other wire. Because of this, emitted electromagnetic radiation from the wire pairs is highly reduced.

Also, the transmission of the differential signal allows for something known as *common-mode* rejection. This is a fancy way of saying that since there are two wires and they are run together in parallel, any electromagnetic noise that may affect one wire will affect both wires in approximately the same way. Because our goal is to look at the *difference* between the two wires, all common noise between the two won't affect our measurement.

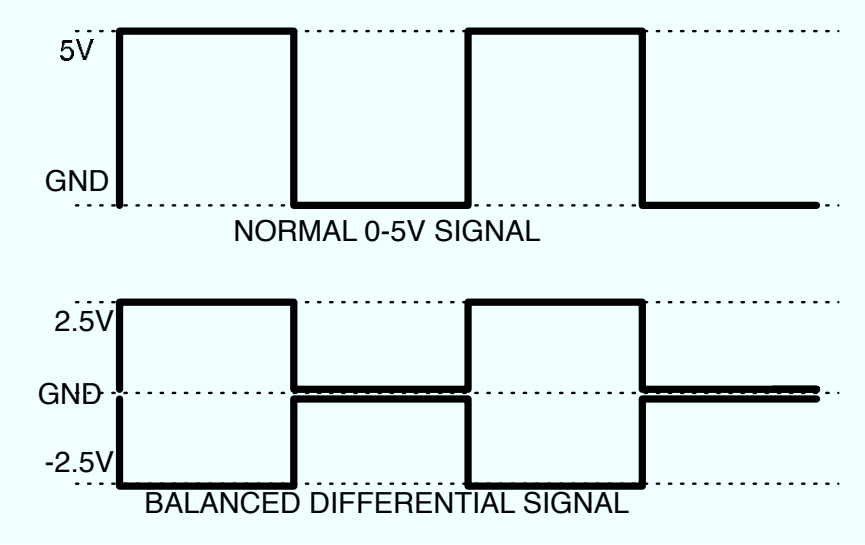
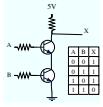


Figure 11.11: A normal 0-5 volt signal and a balanced differential signal



The Buzz...

The Right Hand Rule

Look at your right hand. Give yourself a thumbs-up, but loosely curl your other finger so your hand is cupped. Your thumb represents a piece of wire with current flowing in it in the direction you are pointing. Your fingers represent the magnetic field curling around the piece of wire as a result of this current. This little trick is known as the right-hand rule, and it's helpful way to remember the direction of both current and its resulting magnetic field.

Twisted Pair Wire Categories

Twisted pair wires are sold by category type. Some of the common categories include:

- Category 1: Original formal twisted pair cableing used for carrying voice conversations for the POTS network.
- Category 2: Used for Token Ring networks with transfer rates up to 1 Mbps.
- Category 3: Capable for data transmission up to 16 MHz and used for data transmission up to 10 MBps.
- Category 5: Rated for data transmission up to 100 MHz. Used commonly today for most network data transmission.
- Category 5e: An enhanced version of Cat5, recommended as the minimum category of wire to use for new installations.
- Category 6: A relatively new rating, with data transmissions up to 200 MHz. It gives the best possible performance using today's wiring standards.

11.5 Ethernet

By far, the most common type of inter-computer networking used today is Ethernet. Ethernet is actually a protocol that works on the data-link layer of the OSI model. It commonly gets used with other higher level protocols like TCP and IP as well as top level protocols like HTTP or FTP. The underlying data-link and physical protocols for most of these layers is ethernet.

The Progression of Ethernet

Ethernet started as a network utilizing a coaxial cable, known as the ether, in which multiple nodes were able to communicate using radio-frequency communication methods. The two earliest OSI physical layer specifictions were 10BASE2 and 10BASE5.

10BASE-2/5

The first Ethernet specifications used coaxial cable for data transmission. 10BASE5, sometimes called Thicknet, used a thick cable similar to RG-8 while 10BASE2, sometimes called Thinnet, used the thinner RG-58. The nomenclature of 10BASE5/10BASE2 referred to:

- 10 - represents transfer speed, in this case 10 MBit/sec.
- BASE - refers to BASEband as the type of signalling used on the wire.
- 5 or 2 - refers to the maximum cabling length, in hundreds of meters.

The 10BASE2/5 Ethernet schemes had some limitations. They both required termination resistors and special transcievers which could only be placed at certain intervals of the cable length. 10BASE-5 also only allowed nodes to be installed linearly along a single piece of cable while adding a node to a 10BASE-2 network meant shutting service down to the entire network during the upgrade.

StarLAN (1BASE-5)

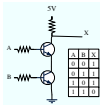
The first use of ethernet on a set of twisted wires was known as StarLAN. It ran at a speed of 1 Mbps and was developed as part of the IEEE 802.3 specification. Developed in the early 1980s, it provided the concept known as *link-beating* in which a heartbeat signal was sent out by the devices on the network so that it could easily be determined if they were currently connected to the network or not.

At its inception, one of the main focuses of StarLAN was to be compatible with existing telephone wiring systems. The idea was that the physical aspects of the network could be used over pairs of wires in bulk telephone cables. In order for this to work, the specification had to allow for both the ability to use the wires existing in the telephone cables AND to make sure network communications did not interfere with other sets of wires in the same cable.

10BASE-T

The development of StarLAN led to the development of 10BASE-T, also known as IEEE 802.3i. The main physical difference between 10BASE-T and its predecessors is in the cabling. In this case the T stands for Twisted Pair, in which the use of a twisted pair of conductors is specified over the use of coaxial cable.

The 10BASE-T specification does not specify many requirements as earlier versions did. Instead, it specifies certain characteristics that the physical medium must meet such as noise and signal degradation requirements. By doing this, the designers were allowing the use of existing twisted pair wiring network that may have already been



The Buzz...

Termination Resistors

In some networked applications, *termination resistors* are used. These are resistors that are placed at the end of a network segment and are used to match the impedance (or resistance) between the transmission line and the unconnected end of the network cable.

Impedance matching in transmission lines is important because impedance mismatch causes some of the electromagnetic wave that is transmitted to be reflected back over the transmission line. This can cause interference in the communication. The matched impedance that is created from the addition of the termination resistor minimizes this reflection.

installed as long as they met the specifications. A quick test with an electronic cable tester would determine if existing building wiring, that may have been run originally as telephone wire, met 10BASE-T specifications.

Data transmission in 10BASE-T (and earlier Ethernet standards) was completed via a Manchester Encoded signal.

RJ-45

The 10BASE-T specification called for RJ-45 jacks at the end of each wire segment. An RJ-45 is a connector with 8 electrical connections and is shown in Figure 11.12, on the following page. It is similiar, though wider, to the RJ-11 jack used in telephone jack connections.

10BASE-T is commonly used with CAT5 wire, which has 4 pairs of wires. In 10BASE-T, only two pairs are used — one for transmitting and one for receiving.

100BASE-T

A number of 100BASE-T Ethernet implementations exist though few became widespread. Of them, the 100BASE-TX standard was the most popular. This standard provided 100 Mbit/sec data transmission capabilities. 100BASE-T is almost physically identical to 10BASE-T.

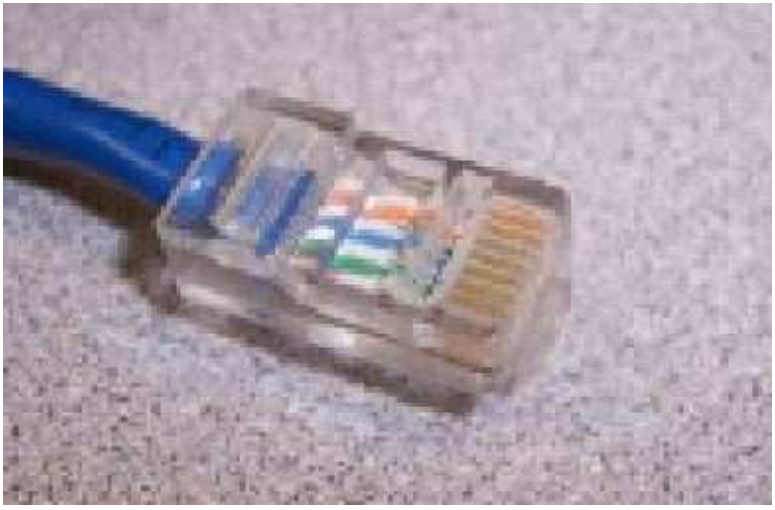


Figure 11.12: An RJ-45 connector at the end of a CAT-5 cable

A big difference between 10BASE-T and 100BASE-T(X) is in the data encoding. 10BASE-T used Manchester encoding for its data transmission while 100BASE-T(X) used an encoding scheme known as MLT-3, shown in Figure 11.13, on the next page. MLT-3 (Multilevel Transmission with 3 levels) encoding offers improvement of Manchester encoding because it is less bandwidth intensive (that is, there are less data transitions than Manchester for the same information). It is also bipolar, alternating between a positive, a 0, and a negative voltage state. This feature helps reduce electromagnetic interference.

1000BASE-T and beyond

More recently, the push has been to move Ethernet into the 1 GBit/sec transmission category known as 1000BASE-T and very recently into the 10 GBit/sec category. In order to accomplish these higher transfer rates, some of the following techniques are being utilized:

- The use of optical fiber instead of copper cable. Fiber optic cables are basically immune to electromagnetic noise and allow very long distances of transmission versus copper cable.
- The use of PAM (Pulse Amplitude Modulation) encoding of MLT-3. Where MLT-3 offered 3 bit levels of transmission, PAM offers 5 (PAM-5) or 8 (PAM-8) or more.

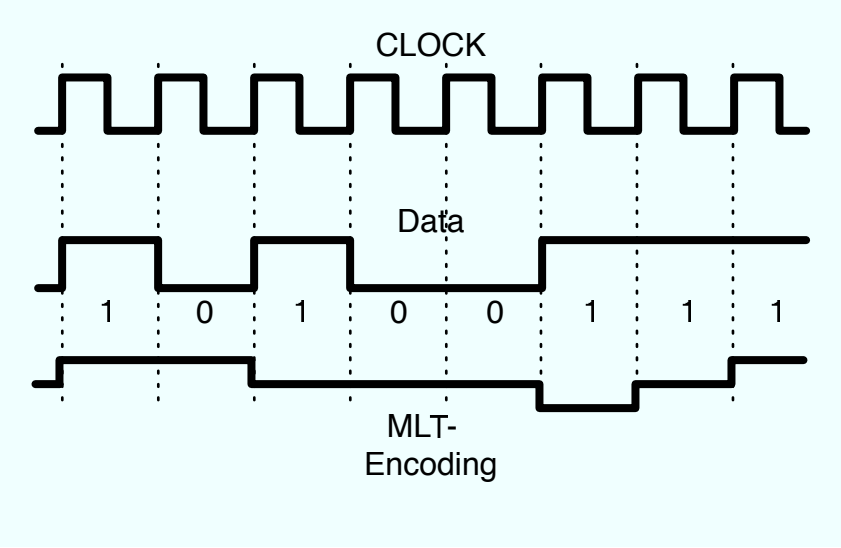


Figure 11.13: An example of MLT-3 encoding

- The other two pairs of wires. The CAT5 and CAT6 cables that are commonly used for Ethernet have thus far only used 2 of the 4 available pairs of wires. By making use of the other two pairs, more data can be transmitted between the transmitter and receiver.

Chapter 12

External Devices

12.1 Display Devices

The Cathode Ray Tube

For many years the Cathode Ray Tube (CRT) style of computer display was the style used for computer monitors, though its use extends into many other applications including television, radar displays, and oscilloscopes. The original idea of the CRT was invented by Karl Bruan in 1897 and later revised by Philo T. Farnsworth in the late 1920s for what would become the television.

In a CRT, an electron *gun* emits a stream of *cathode rays*. These rays are electrons that are moving through a vacuum tube from the cathode (negative electrode) to the anode (positive electrode). The cathode is heated and radiates electrons which move through the vacuum to the anode. Just behind the anode is a glass wall that is coated in a *phosphorescent* material that glows when hit with electrons.

Early experiments with the CRT involve placing a shape between the cathode and the phosphor screen. The electrons that were directed at the screen were blocked by the shape and would not hit the screen so the shape would cause a shadow to appear on the screen. It was also discovered that the beams of electrons could be manipulated by the use of magnets; that is, placing a magnet near the electron beam would cause it to deflect from its straight line course.

Modern CRTs use this magnetic deflection to control the path of the electron beam. First, a wire coil known as the *focusing* coil is used to create a magnetic field around the electron beam as it leaves the gun. The focus coil is used to manipulate the shape of the beam into

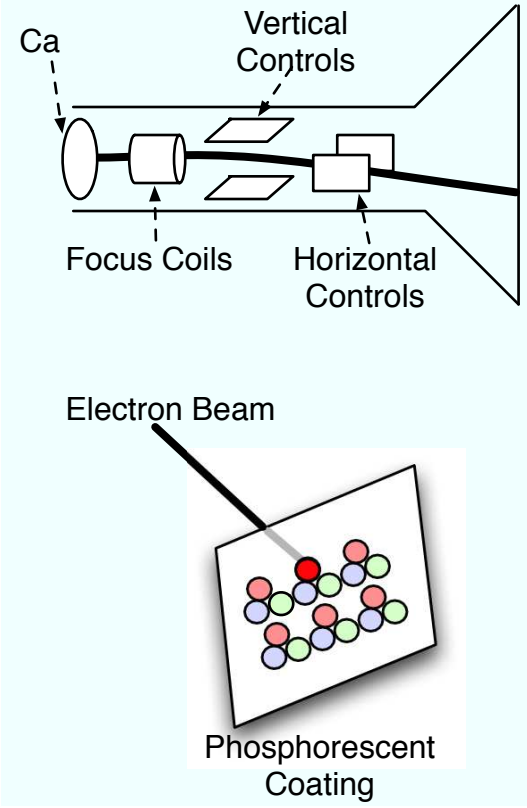


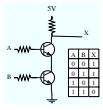
Figure 12.1: A Cathode Ray Tube

a focused area of electrons. Without the focus coil, the beam would spread out.

Next, a second set of coils known as the *deflection coils* are used to aim the beam at a specific part of the glass screen. By varying the current in the coils, the amount of magnetic deflection can be controlled and thus the end location of the beam can be targeted.

Finally, the intensity of the electron beam is controlled by the incoming data signal.

Color CRTs use three different phosphorescent materials emitting red, blue, and green lights. They are set very close together to give the impression of a single color. Color CRTs also use three separate electron guns, one for each of the primary colors.



The Buzz...

Phosphorescence

Phosphorescence is a form of *photoluminescence*. A photoluminescent material is capable of absorbing a specific type of radiated energy, generally of the form of visible light, and re-emitting that energy as a photon, also of visible light. Another example of photoluminescence is *fluorescence*.

What makes phosphorescence unique is that the absorbed energy gets transformed into a different atomic energy state that is difficult to transform back into its original state. The transformation back to emitted energy happens slowly, meaning that a phosphorescent material emits its absorbed energy slowly over a (relatively) long period of time. Phosphorescent materials are thus used commonly as "glow in the dark" materials.

Monochrome Displays

Early PCs came with monochrome, or single color, displays. These were typically either green or amber in color. The displays were initially used to only display text due to the limitations of the video cards used for the displays. Later, with the invention of the Hercules and CGA video cards, bitmapped graphics also became available for display.

Liquid Crystal Displays

The history of the Liquid Crystal Display (LCD) started in Germany in 1904 with Otto Lehmann. In 1936, a patent was granted to Marconi's company for an LCD "valve" which allowed the selective passing or filtering of light through an electrically controlled gate. However, the most important work in modern LCD technology was performed by George Gray at the University of Hull in England in the late 1960s. Gray, though support of an English group known as the Royal Radar Establishment, ultimately discovered the chemical compound used in today's modern liquid crystal displays.

LCD Concepts

An LCD is made up of a number of layers that are used for the reflection or absorption of light, as shown in Figure 12.2, on the following page.

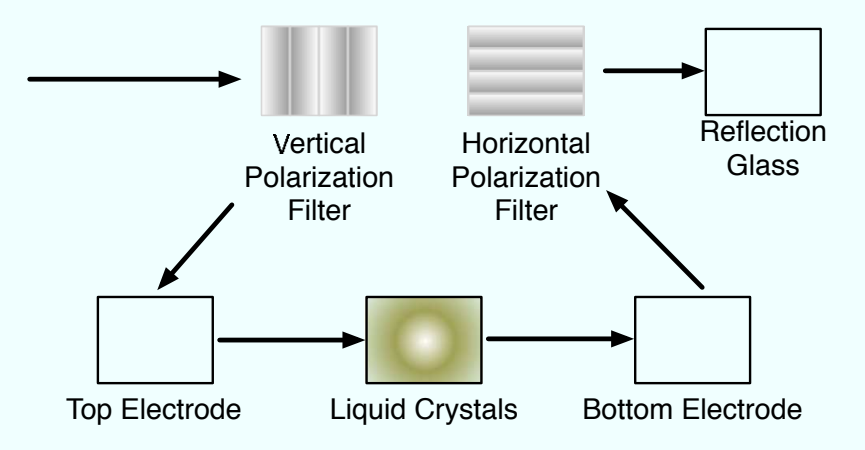


Figure 12.2: How the LCD works

The liquid crystals themselves reside on a flat plate. On both sides of this plate are flat glass plates with the electrodes that can be used to energize the crystals. In their normal state, the crystals form helixes which rotate light as it passes through them. All light passing through the top vertical polarization filter is aligned in the vertical direction. As it goes through the crystals, it is rotated into the horizontal direction. As it gets to the horizontal polarization filter it passes through and is reflected back on the back surface.

When energized, the liquid crystals align themselves in the direction of the electric field meaning that the rotation of light they exhibited earlier is now lessened. This means that light that enters the vertical filter will not pass through the horizontal filter. Instead of being reflected, the light is absorbed.

The amount of current into the electrode can be controlled and as such, the intensity of the reflected or absorbed light can also be controlled.

Colored LCDs

Colored LCDs are similar to colored CRT displays. Each pixel is made up of three individual red, blue, and green primary colored units situated very close to each other. To obtain the color, an added filter is used to only allow a certain frequency of light to pass through, thus meaning that only that color of light will be reflected from the rear surface.

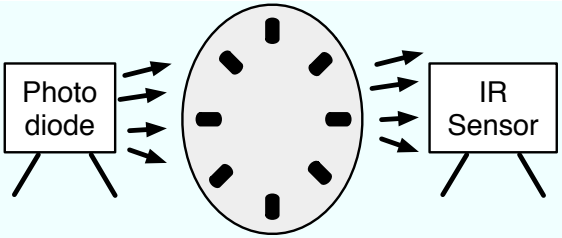


Figure 12.3: The wheel inside of a ball mouse

12.2 Input Devices

Mice

The basic computer mouse come in two types: a ball mouse and an optical mouse.

Ball Mouse

The ball mouse is made up of a small ball inside that rolls around against two wheels, one in the x direction and one in the y direction. As the ball rolls, it causes the wheels to rotate. A sample wheel is shown in Figure 12.3.

The wheel has a number of slots in it which allows a photo-diode on one side of the wheel to send an infrared signal to a sensor on the other side of the wheel. As the wheel rotates, this signal is completed and broken by the slots in the wheel. The mouse uses this information to determine which direction the wheel is turning and how fast the wheel is turning, which can then be translated by the computer.

Optical Mouse

The optical mouse replaces the ball with a red Light Emitting Diode (LED). The light from the LED is sent downward to the desk surface and reflects back up onto an optical sensor. The optical sensor is able to take a digital "snapshot" of the surface below the mouse. It sends this information to a small processor which watches the snapshot over time to determine how it is changing. It interprets the changes as moves of the mouse and sends the appropriate signal to the computer.

Touch devices

Touchpads

Almost all laptops today come with built in touchpads. These input devices allows you to use pressure created by one or more fingers to manipulate the movement of the cursor on the screen, much like a mouse.

Touchpads can use a variety of technologies to determine finger placement and movement. Most commonly, a touchpad senses the electrical capacitance between a finger and a grid of electrodes within the touchpad. Since the human body is a conductor, the contact of the finger to the top of the touchpad, which is made of an insulating material, to the electrode creates a capacitor. The touchpad has sensors in each of its electrodes that can sense this capacitance increase and deduce the location of your finger.

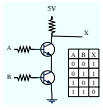
While capacitive touchpads cannot directly measure the pressure being applied, the capacitance does vary with the amount of skin in contact with the surface. Because of this, more pressure generally creates more capacitance, which allows the touchpad software to determine some level of applied pressure.

Touch screen displays

Touch screen displays are becoming popular ways of inputting data without the need for using a keyboard. Coupling the input to a dynamic display like a monitor allows a programmer to change what kinds of input the program will allow and lets the user simply point to their selection on the screen.

Capacitive technology, like that used in many touchpads, is common in touch screen applications. It operates in exactly the same way, with a network of electrodes spread over the area of the screen. However, one downside to using capacitive technology in this regard is that it requires the use of a conductive pointing device. This means that other pointing devices, such as a plastic stylus, won't work with this technology.

An alternative to capacitive touch displays is resistive touch displays. With resistive touch displays, the outer layer of the display is coated first in a then resistive material, and then in a conductive material. When pressure is applied to the screen, the outer conductive layer presses through the thin resistive layer and makes contact with an inner conductor. Electrodes that are located around the screen sense



The Buzz...

Bouncy Switches

Every time a switch closes, something known as *bounce* occurs. This is because a switch isn't a perfect electrical device and as the mechanical portion of the switch makes contact, it may make and lose contact rapidly a few times during this period. This back and forth electrical noise can cause some unwanted action if a resulting electrical action also *bounces* with the switch. Thus, many mechanical switches that are tied to sensitive electrical circuits employ some form of de-bounce circuitry that is able to filter out the bounce noise caused by the switch. In Figure 12.4, on the following page, a graph is shown displaying what actual switch bounce may look like when a switch is pressed.

the change in resistance and register the touch event. Resistive touch screens are able to work with any device able to apply pressure to the screen.

Resistive touch screens are less expensive to manufacture than capacitive screens, but provide significantly less clarity due to the additional material applied over the screen. It's also possible to damage the resistive screen if a sharp object cuts through the layers.

Keyboards

Computer keyboard technology has changed very little over the years. The basic premise is that below each key is a small switch that when pressed completes a circuit that is read by a small processor in the keyboard.

What has changed over time is the complexity of the switches. Various computer models make use of varying pressures on the switches allowing a more tactile feedback that some people enjoy. In addition, some switches provide an audible response to help associate the actual action of the keypress as feedback to the user.

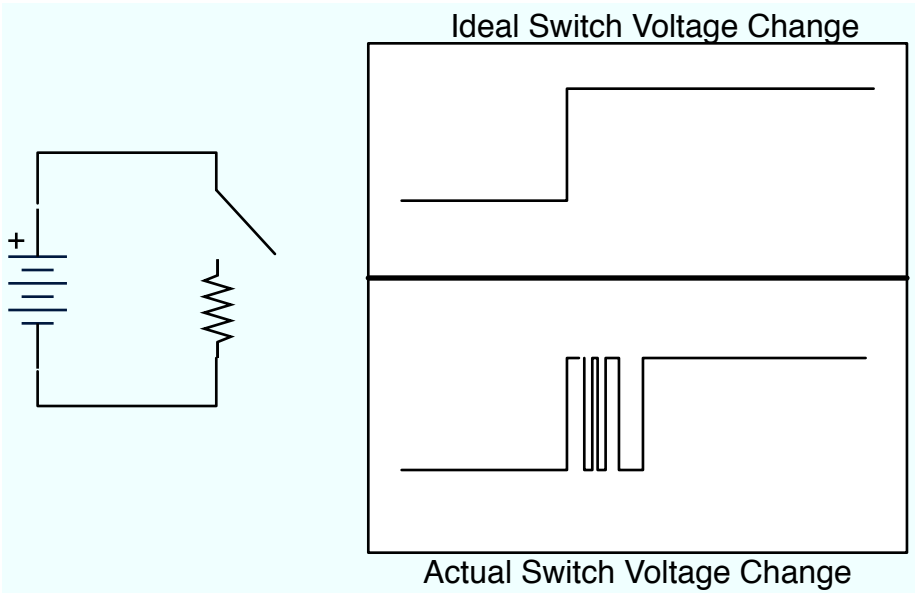


Figure 12.4: A graph showing the difference between an ideal switch and an actual switch

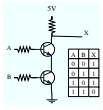
12.3 Connections

There are a number of common connection types used for communication with external devices. These connection types fall into two classes: serial and parallel. Serial data connections transmit their information one bit at a time over a single line. Parallel data connections transmit their information over multiple data lines simultaneously. Serial has the advantage of requiring fewer data lines, and thus fewer wires and connections. However, it requires more time to transmit the same amount of information versus parallel.

Sound

Audio from the sound card or integrated sound system is transmitted via 3.5mm jacks over the following lines:

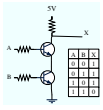
- Pink — Microphone input.
- Light Blue — Analog line level input.
- Light Green – Analog line level output.



The Buzz...

Colored Connectors

Most modern PCs now use colors when distinguishing the various connectors on them. These color codes are part of a standard known as PC 99 which was developed by Microsoft and Intel in 1998. The idea was to help users unfamiliar with the connectors establish a visual identity as to which cable went into which connector. Eventually the color coding was adopted by most motherboard and PC manufacturers.



The Buzz...

What is line level?

In the world of audio electronics, *line level* refers to the signal strength of an audio signal as it is transmitted between audio devices.

The voltage level that is output by microphones is very small, so it is amplified before being transmitted to an audio device. This stage is commonly referred to as *pre*-amplification. In contrast, final amplification takes place before the audio signal is sent to an output device like speakers.

The pre-amplified signal is passed around from audio devices at line level.

In addition, some sound cards also add black and brown colored connectors as outputs for systems with multiple speakers. Newer sound cards may even support digital input and output interfaces.

PS/2

Until very recently, the common method of connecting a keyboard and mouse to a computer was through a 6-pin mini-DIN

connector commonly referred to as a PS/2 connector as it was introduced on the the original IBM PS/2 computer. The data is sent serially. For any event such as a mouse button press, key press, or motion, a set of data bytes are transmitted indicating what happened.

The 6 pins on the mini-DIN connector are used as follows:

- 1 - Data
- 2 - Unused
- 3 - Ground
- 4 - +5V
- 5 - Clock
- 6 - Unused

Serial Port

The serial port is an interface for communicating with external devices in a serial fashion. Early computers used a circuit on the motherboard known as the Universal Asynchronous Receiver Transmitter (UART) that was capable of transmitting and receiving data characters over the serial line, as well as handling other important aspects such as ensuring proper timing.

Most computer serial ports implement a standard known as RS-232 which is used for sending binary between two devices. It specifies the electrical characteristics of the signals being used in addition to the connector types and cable lengths. It does not, however, specify the elements of character encoding nor does it specify data transmission speeds.

The original RS-232 standard was implemented in a 25-pin D-type connector (so named because it looks like the letter D). Early computer serial ports came with this 25 pin connector. As a trade off to size and expense, many manufacturers instead began using a 9-pin D-type connector to implement the serial protocol by dropping some of the lesser used signals.

1. A 5-pin DIN connector, known as the AT connector, was used in early PCs but is not commonly used anymore. It is electrically equivalent to the PS/2 style connection (though physically different).

While it is slowly being phased out by USB, the DB-9 connector and the RS-232 protocol are still prevalent on many of today's computers.

The pin definitions for the connector are:

- Pin 1 — Carrier Detect (DCD)
- Pin 2 — Received Data (RD)
- Pin 3 — Transmitted Data (TD)
- Pin 4 — Data Terminal Ready (DTR)
- Pin 5 — Common Ground
- Pin 6 — Data Set Ready (DSR)
- Pin 7 — Request To Send (RTS)
- Pin 8 — Clear To Send (CTS)
- Pin 9 — Ring Indicator (RI)

The specification and use of RS-232 began as a common connection between a TTY device and a modem. Over time, it became a popular choice for communication between devices that were neither a TTY or a modem. Many of the data lines that were originally used for the setup of modem communications are implemented for many devices. In fact, many devices only make use of lines 2,3, and 5.

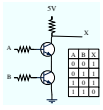
Data Framing

Data that is to be transmitted over the serial line is first *framed*. Commonly, a one byte (8 bit) piece of data will be sent at a time. Data framing includes the addition of the following to the data:

- Start bit - To notify the receiver that the message has started
- Stop bit - To notify the receiver that the message has ended
- Parity bit - An additional bit used to notify the receiver of the intended number of 1s or 0s in the transmitted message to help for error detection.

Parallel Port

Along with the serial port, the parallel port was a commonly used interface for connecting external devices to the computer; though, the ubiquity of USB today has largely rendered the parallel port obsolete.



The Buzz...

RS-232 Connections

The RS-232 standard considers the connection between two distinct devices, a DTE (Data Terminal Equipment) and a DCE (Data Communications Equipment). This can be problematic when designing a new device because of the choice as to which type of device to emulate. For example, consider that a DTE device implements its data transmission line on pin 3. This means that the DCE has to implement its receive line on pin 3 in order for a straight cable connection to work.

To connect a DTE device to another DTE device (or, similarly a DCE to a DCE) one must make use of a *null modem* adapter. This adapter provides a wiring crossover between the transmit and receive pins that allows two DTEs to communicate. Sometimes the other lines are crossed over as well, depending on the required implementation.

Null modems initially served their purpose to allow two personal computers (each implementing a DTE style RS-232 port) to communicate with each other. This connection was commonly used for file transfers before network cards became common.

In its prime, the parallel port was most commonly used to communicate with a printer. Some hard disk drives and external storage devices such as ZIP drives were also connected via the parallel port.

The original version of the parallel port, sometimes known as the printer parallel interface (PPI), was developed by the Centronics company in the 1970s. Though for many years the Centronics connection was the standard, no formal standard was ever put into place. Because of this, newer versions were created leading to the creation of the IEEE 1284 specification in 1994. The IEEE 1284 standard called for modern ports to support five modes covering the most common specifications including the legacy Centronics connection.

Some early computers implemented the parallel port with a 36-pin Amphenol connector[2] as described in the Centronics standard. The

2. This connector is usually just referred to as a Centronics connector, though it was

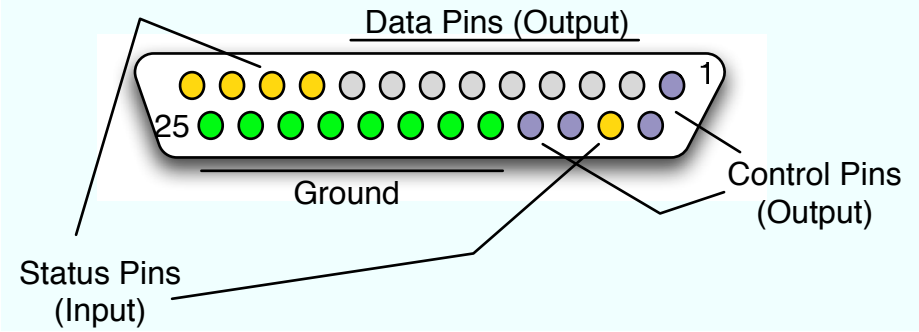
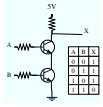


Figure 12.5: The 25-pin D connector used for the parallel port



The Buzz...

Parallel Port Data Lines

The parallel port drivers use the data lines of the parallel port to transmit digital data out to remote devices. These digital lines can also be used for other things. For example, if you were to take control of the parallel port from the operating system you could individually control each of these data lines (as well as a few other lines in the parallel port) and use them for any type of digital control you like. Your author has done this very thing and, along some added circuitry and the use of electrical relays for control of high current, created a computer controlled holiday light display.

IEEE 1294 specification also allowed for a DB-25 D-type connector which quickly proved to be a more popular choice.

The pin definitions of the 25-pin D-type parallel port are defined in Figure 12.5.

technically made by the Amphenol company

Firewire

Also known as IEEE 1394, Firewire was a serial bus interface that allowed devices to connect and disconnect from the computer in real time. The IEEE standard was created in the late 1980s as a replacement to the SCSI bus. It was popular in Apple brand computing products but due to per-product licensing costs never gained widespread popularity mainly due to the popularity and inexpensive nature of USB.

USB

Today, the most prevalent form of connection to external devices is via the Universal Serial Bus (USB). The USB specifies a host controller, typically residing within the computer, and allows multiple connected devices. USB was design to allow the connection of devices to the PC without the need for adding expansion cards to the motherboard.

The USB standard is overseen by a board of implementers including companies such as Intel, Microsoft, and Apple. The first specification, known as USB 1.0 was released in early 1996. A revised USB 1.1 specification followed in late 1998. A higher speed standard, known as USB 2.0 was released in 2000.

Connections

USB standard connectors are four-pin connectors with the following pin-outs:

- +5V
- D-
- D+
- Ground

The D+ and D- pins are the data transmission pins, which operate differentially. A display of the pins are shown in Figure 12.6, on the following page.

Data Transmission

Transmitted USB data is encoded in a Non-Return to Zero, inverted (NRZI) fashion—the same style used in compact disc encoding. A logic 1 does not change the signal from its current state while a logic 0 creates a transition of the signal from its current state.

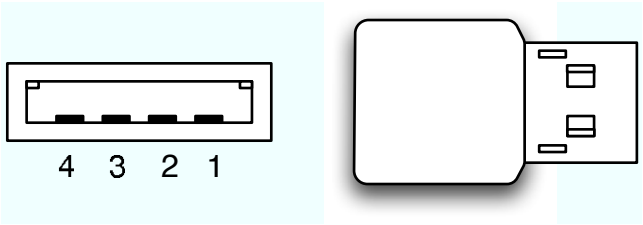
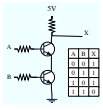


Figure 12.6: A USB connector



The Buzz...

USB Connectors

Special care went into the design of the USB connectors after a thorough study of the existing forms of connectors used in many other computing applications. The connector is specifically designed to have a good gripping force. It is very difficult to insert into a receptacle the wrong way. Special equipment like gender changers need not be used. Even the outer metallic sheath of the connector is specified in the design, in order to guarantee that it makes contact with the receptacle sheath before the internal pins make contact with their counterparts. This helps to mitigate any static buildup that could otherwise damage internal electronic components.

To be happy in this world, first you need a cell phone and then you need an airplane. Then you're truly wireless.
► Ted Turner

Chapter 13

Wireless

13.1 Wireless Fundamentals

Wireless communications is accomplished by using low powered radio frequency waves for data transmission between two or more devices. In this case, a radio wave means an electromagnetic wave between about 3Hz and 300GHz. These radio waves are categorized into groups based on their frequency as shown in Figure 13.1, on the next page.

The History

Early experiments performed by Faraday (see Section 3.1, *Magnetic Motion*, on page 35) concluded that transmission of electromagnetic waves through the open air was possible. Until this point, all electromagnetic wave transmission was observed through conductors. However, the equations which explained the waves indicated that some wave presence would also be found in the space surrounding the conductor.

In 1873, James Maxwell wrote a paper titled *The Dynamical Theory of the Electromagnetic Field* which, amongst other things, described a set of equations that explained the behavior of electric and magnetic fields and their interactions with matter. This began the theoretical search for the existence of transmitted waves through the air.

In 1888, Heinrich Hertz created the first demonstration of electromagnetic wave transmission through the air with a device that was able to create radio waves. He was able to reformulate Maxwell's equations into an equation known as the *wave equation*. This equation mathematically describes the properties of transmitted electromagnetic waves.

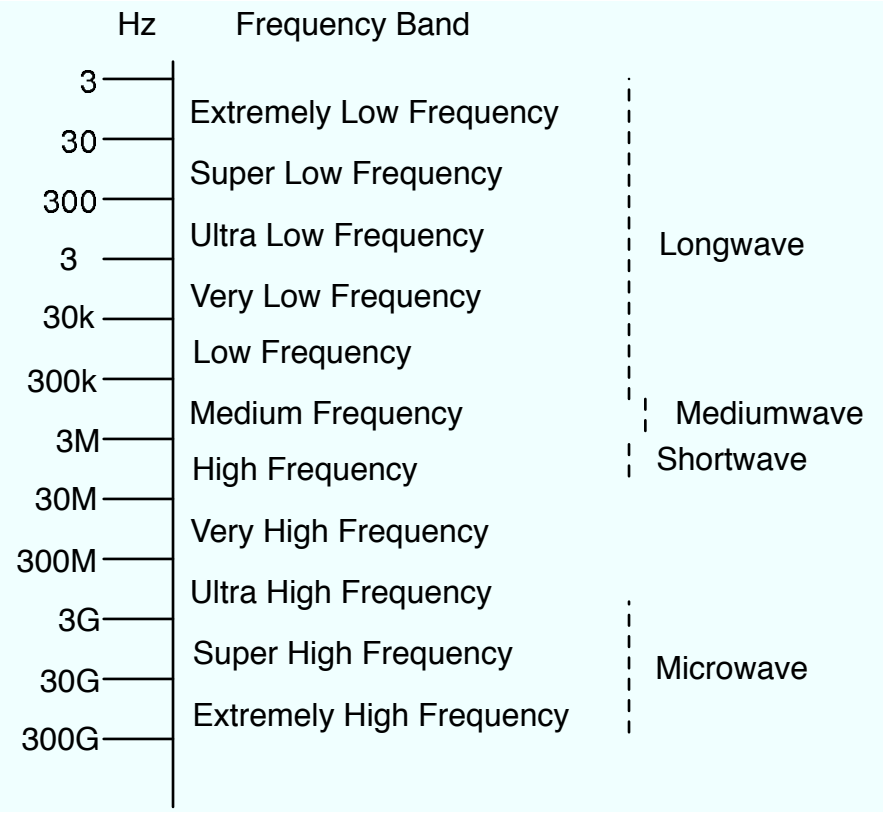
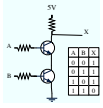


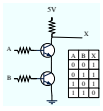
Figure 13.1: Frequency Bands



The Buzz...

Do higher frequencies exist above 300GHz?

Certainly, but they're not considered to be radio waves anymore. Beyond 300Ghz (which corresponds to a wavelength of 1mm), the next grouping of frequencies corresponds to light. As the frequency increases it first becomes infrared, then the visible spectrum of colors, and finally ultraviolet. Beyond ultraviolet light are X-rays and gamma rays.



The Buzz...

Hertzian Waves

Hertz discovered the first electromagnetic wave, sometimes called a Hertzian wave.

The electromagnetic wave is a type of *transverse* wave. This means that it oscillates in the direction perpendicular to the direction that it advances. That is, if the wave was traveling from west to east, it would be oscillating from north two south.

Furthermore, electeomagnetic waves have two perpendicular directions of oscillation because they are made up of both electrical and magnetic field components. Thus, if the wave travels in the x-axis, then the electric field component would be in the direction of the y-axis and the magnetic field component would be in the direction of the z-axis.

Radio Heats Up

After scientific demonstration by Hertz, research into radio transmission grew. Early demonstrations only showed the possibility of radio wave transmission. However, focus quickly turned to using radio waves to transmit information—such as a wireless telegraph. More than just scientific credibility was at stake—money was at stake too.

In 1893, Nikola Tesla demonstrated the first wireless transmission system. At this time, he was heavily focused on his work with AC power (see Section 3.1, *The Ultimate Power Battle*, on page 35). Tesla's main focus was on wireless AC power transmission, though his initial patent on wireless electrical energy transmission described other uses such as communicating messages.

In 1895, Guglielmo Marconi incorporated some of Tesla's ideas and derived a system that could be used for wireless communications. Marconi is often attributed as being the father of radio even though the work of many other people had been involved in his discovery. In fact, Marconi himself is said to have admitted that he borrowed ideas from patents filed by Karl Braun.

The Electromagnetic Spectrum

The range of all possible electromagnetic wave frequencies makes up the *spectrum*. Of interest to use are the radio frequencies, which use the band between 3Hz and 300 GHz. As seen in Figure 13.1, on page 206, these frequencies are generically grouped into frequency classications from extremely low to extremely high. There are also four classifications of waves as being microwave, shortwave, mediumwave, or longwave. In theory, the only difference between any of the names is simply the frequency at which the wave oscillates. However, when it comes to actual transmission there is more to the story.

ELF, ULF, VLF, LF

Electromagnetic waves at low frequencies like ELF, ULF, and VLF are used mainly for surface to submarine data transmission as they more easily penetrate the depths of ocean water than higher frequencies. They're less suited for surface communications because low frequencies generally correspond to low data rates. As well, these low frequencies mean high wavelengths. Antennas required for signal transmission in the ELF range are many kilometers long.

LF broadcasts are generally used in military applications and as an alternative to higher frequency bands which may already be crowded with conversation. Some practical applications include LORAN, a navigational system that can track aircraft position using time intervals between radio transmissions. This is a precursor to today's satellite based Global Positioning Systems.

MF

Mediumwave frequencies make up a majority of broadcast communication frequencies. For example, AM radio stations broadcast within this frequency range.

These frequencies have a characteristic known as *groundwaves*. This means that they tend to follow along the surface of the earth even as it curves. Because of this, they made for ideal long distance transmission frequencies. Furthermore, in the evening they are highly reflective off of the ionosphere, resulting in even further signal transmission.

HF

High Frequency transmissions (sometimes called shortwave) are commonly used for terrestrial communications at long distances. This is because the ionosphere, the part of the atmosphere that is full of ions

due to radiation from the sun, reflects these frequencies. This makes it possible to transmit these frequencies to the other side of the world. However, the ability of the atmosphere to help carry these frequencies is highly dependent on atmospheric conditions.

Because the shortwave band is well suited for transmission, it is a very popular band for data. This part of the spectrum is divided into many pieces and tightly allocated by the various countries' regulating bodies (the FCC in the United States).

VHF

VHF frequencies are used for short to medium distance communications. In this band are FM radio stations (between 88 and 108 MHz) and television stations 2 through 13.

Throughout this part of the spectrum are licensed bands for hand-held radio sets, cordless telephone, military applications, air traffic navigation units, and some remote control devices.

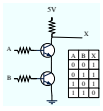
UHF

As frequencies progress into the UHF range the effects of using the ionosphere for signal propagation begin to degrade. UHF transmissions propagate less readily than lower frequencies. Thus, UHF is suited for medium distance communication. It houses bands for television channels 14-69, cellular phones, and some amateur radio operations.

SHF, EHF

In the super and extremely high frequency bands are some cellular phones, and most notably radar systems. In these bands, radio signals are highly susceptible to atmospheric effects including attenuation by water droplets (such as precipitation). Some communication systems make use of these frequencies for high speed data transmission but they are much more effective in flat terrain.

The UHF, SHF, and EHF bands contain the set of frequencies known as *microwaves*. Microwave frequencies are generally broken down into letter designations suck as the "C band" (4 to 8 GHz) and "K band" (18 to 26 GHz) ranges.



The Buzz...

Microwave Ovens

Microwave ovens make use of microwave frequencies for cooking food. Inside, a device known as a *magnetron* emits radio wave energy at a frequency of about 2.45 GHz, corresponding to a wavelength of about 12.24 centimeters. The radio wave is propagated through a guide that aims the wave into the cooking area and causes the waves to reflect off of the surfaces in somewhat of a random pattern (a fan is usually employed to help with this). At this frequency, water and fat in the food absorb the energy created from the wave and begin to heat.

The cooking area is a metallic structure that creates a Faraday cage to keep the microwaves inside the chamber. The viewing holes in the door are sized so that the 12.24 cm wavelength microwaves cannot pass through but visible light with a much smaller wavelength can pass through.

13.2 Wireless Fundamentals

The fundamental concept of wireless communications is that of a radio wave being transmitted between a sender and a receiver. From the sender's side, some circuitry must be in place that is capable of generating the frequencies needed for broadcast. On the receiver's side, circuitry is needed that is capable of picking up the transmitted signal.

Modulation

Early wireless systems carried voice information, which contains frequencies from 80Hz to 6kHz. These frequencies are picked up by a microphone and converted from sound waves into electrical waves.

A problem quickly develops, however, when more than one person tries to transmit these frequencies at the same time. Interference results, and the transmitted information quickly becomes distorted.

To counteract this, radio signals are *modulated* with a *carrier wave* which changes their wave characteristics. Modulating is done to allow more conversational room within a frequency range and to change the

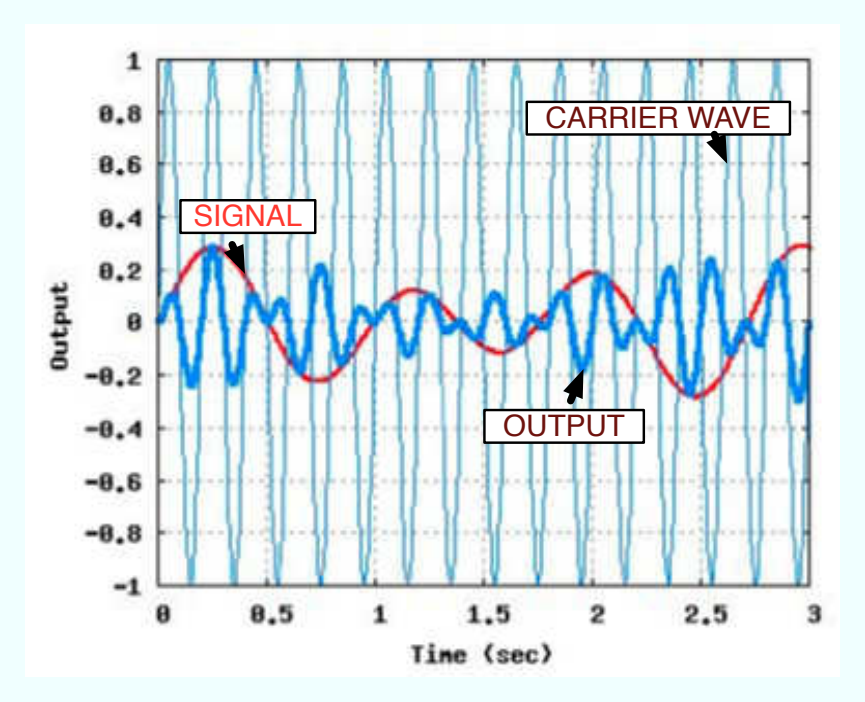


Figure 13.2: Amplitude Modulation

transmitted signals more compatible with how they are being transmitted.

Amplitude Modulation

One form of modulation is Amplitude Modulation, or AM. In AM, an informational signal (such as voice or music) is modulated with a single frequency carrier wave via multiplication. The output signal is the same frequency as the carrier, but with changes in amplitude based on the input signal. In general, the carrier wave frequency is much higher than the signal frequency. A graphic of amplitude modulation is shown in Figure 13.2.

Frequency Modulation

Another common form of modulation is Frequency Modulation, or FM. In an FM signal, the carrier wave is modulated through changes in frequency (as compared to changes in amplitude like in AM).

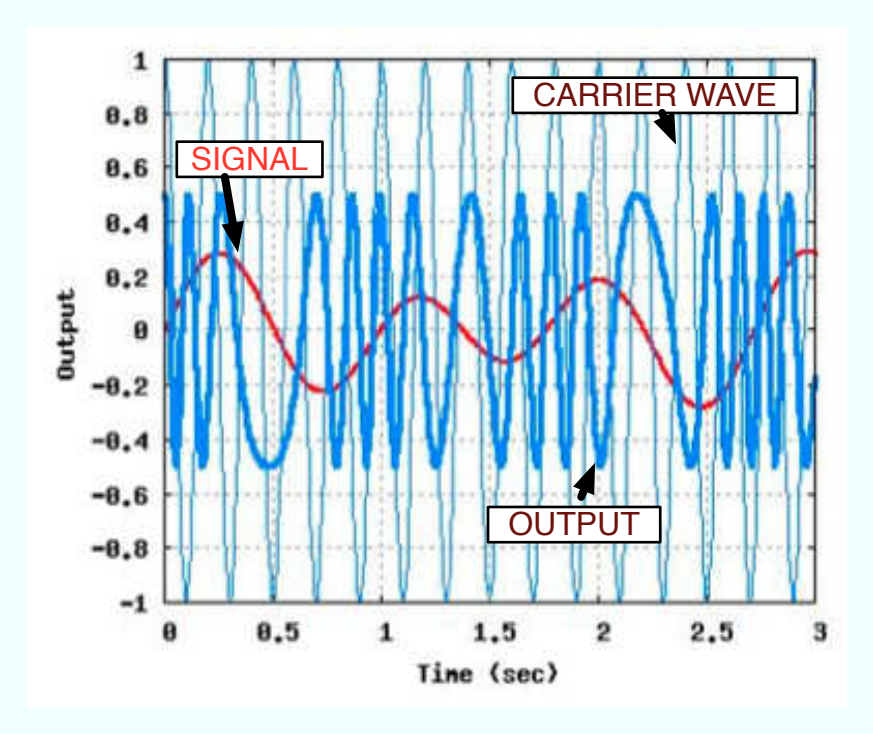


Figure 13.3: Frequency Modulation

Demodulation

When a transmitted modulated signal is received, it must be demodulated. This can be accomplished through a diode rectifier in the case of AM, or with a *phase-locked loop* in the case of FM. A phase-locked loop (PLL) is a circuit that uses internal feedback to lock on to a signal and maintain a phase relationship. The PLL generates a frequency from an internal oscillator and this frequency is compared to the incoming signal. If the frequencies do not match, the PLL increases or decreases its frequency to try and match the incoming frequency. Once locked, the PLL is able to match the incoming frequency very precisely. The PLL signal is then able to be used to demodulate the signal.

Radio Stations

The names AM and FM are generally thought of in reference to radio stations that play music or talk shows. AM was the first modulation

technique used for radio station broadcast and was popularized after World War I with the first commercial service beginning in the 1920s. In the United States, AM radio stations broadcast at carrier frequencies between 520 and 1710 kHz spaced every 10 kHz.

FM broadcasts in the US began in the 1940s in the band between 42 and 50 MHz though today are located between 88.1 and 107.9 MHz. Carrier spacing every 0.2 MHz.

Antennas

Wireless signal communication at any appreciable distance would not be possible without the use of antennas. An antenna is simply an electrical conductor whose shape is designed to help radiate an electromagnetic field when transmitting a signal. On the receiving side, the antenna is placed in the path of an electromagnetic field and a current is induced within the conductors of the antenna.

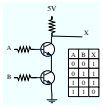
There are two basic forms of antennas: omni-directional and directional. Omni-directional antennas radiate energy equally through their transmission plane where as directional antennas are designed to try and focus their energy in a certain direction. Both styles are generally designed to operate well within a certain narrow range of frequencies, which means an antenna used for transmitting one type of signal may not be appropriate for transmitting another signal of a different frequency.

The simplest type of antennas is a rod antenna, and is nothing more than a straight conducting rod. The rod antenna is an omni-directional antenna radiating its signal out in a cylindrical shape, as well as receiving incoming signals in the same way. One limitation, then, is that the antenna does not transmit or receive in the direction of the rod.

In general, rod antennas are constructed to be the length of a quarter wavelength of their transmission frequency.

13.3 Wireless Technologies

By far the most pervasive wireless standard for computing systems is the 802.11 standard by the IEEE, also known as the Wi-Fi standard. Within the 802.11 standard there are six techniques for modulation. All six use the same protocol.



The Buzz...

The ISM Bands

Some frequency allocations are set aside for use to the industrial, scientific, and medical communities (ISM). This means that while they are regulated,they are reserved for non-commercial use. Recently, wireless protocols like the 802.11 specification have made use of these frequency allocations.

While frequency allocations are controlled by individual countries, many of these countries are members of the International Telecommuncation Union (ITU) which helps guide the decisions of such allocations.

802.11

The original 802.11 protocol specification was released in 1997. Two data rates were specified: 1 or 2 mega bits per second (Mbps). The specification called for data transmission to be done either via infrared or as a radio frequency at 2.4GHz; this is a frequency that is reserved for ISM purposes (see The Buzz for more info).

CSMA

The 802.11 standard utilizes a method known as Carrier Sense Multiple Access (CSMA) for data transmission. With CSMA, a node that wishes to transmit first listens for the carrier signal to determine if nobody else is already transmitting. If a node is currently transmitting, the one that wishes to begin transmission must wait.

Once the data channel becomes available, the new node can begin communicating. However, it's also possible that two nodes can try and communicate at the same time. To counteract this, additional measures can be employed. One such measure is known as CA, or Collision Avoidance. With CA, a node informs other nodes when it wants to transmit so that they're all aware that a transmission is about to occur. Obviously, the same issue can happen as before: two nodes can try and use CA to notify at the same time. However, with CA, the amount of node contention is substantially reduced.

802.11a

The first amendment to the original 802.11 specification (sometimes known as Wi-Fi) was titled 802.11a. It was approved for use in 1999. It uses the same protocol as 802.11, but the specification calls for use in the 5GHz carrier band as opposed to the original 2.4GHz band. The advantage to this is less interference with other devices already working in the 2.4GHz band, but the downside is the transmission distances are significantly reduced. The maximum data rate, however, was increased to 54Mbps.

802.11b

The 802.11b amendment was also approved for use in 1999, with the maximum data rate increased to 11 Mbps. 802.11b offered a great speed increase over the original standard and was very closely aligned to the original. Manufacturers with products conforming to 802.11 found it much easier to migrate to 802.11b than 802.11a.

802.11g

The third amendment to the 802.11 standard, 802.11g, was released in 2003. Similiar to 802.11b, the maximum data rate was increased to 54 MBps. It is also directly compatible with 802.11b, meaning that new 802.11g equipment can talk with 802.11b devices at a reduced speed.

802.11n

The next anticipated amendment to the 802.11 standard is 802.11n. Speculation thus far is that the data rate will be increased to 540 Mbps making it 10 times faster than 802.11g and 100 times faster than 802.11a or b.

Bluetooth

Bluetooth is a wireless specification for small personal networks. It is specified by IEEE standard 802.15.1. Bluetooth devices are categorized into three classes of power transmission levels allowing for either 1, 10, or 100 meters of transmission range.

WiMAX

The IEEE 802.16 standard is the basis for WiMAX, or Worldwide Interoperability for Microwave Access. The major advantage that WiMAX has over 802.11 Wi-Fi is in its quality of service, or QoS. Wi-Fi stations must

compete with each other for data transmission through a central access point through a contention process. When large numbers of stations try to transmit, it can cause bottlenecks. WiMAX instead uses a scheduling algorithm that allows it to register itself with a central access point and from that point on it is issued a guaranteed time slot for its data transmission. This allows QoS to be better maintained during periods of heavier data traffic.

Sooner or later every one of us breathes an atom that has been breathed before by anyone you can think of who has lived before us—Michelangelo or George Washington or Moses.

► Jacob Bronowski, ***Biography of an Atom—And the Universe, New York Times,*** October 1968

Appendix A

The Low Level

Most of us do not have much more than a black box perspective on electronics. We tangibly interact with electronics, but do not need (or sometimes want) to know the lower level aspects that remain hidden from our view. To really understand electronics, however, we need to go beyond just the visible aspects of our physical world and focus on the world at a much smaller level—the atomic level.

In this chapter we'll explore the aspects of atoms related to electronics. We'll look at the basic structure of the atom, what creates electrical charge, and how that charge becomes electricity.

A.1 The Atomic Level

If you had high school chemistry, you should recall that atoms have three parts - protons, neutrons, and electrons. In the basic model of the atom, protons and neutrons make up the nucleus, or the center

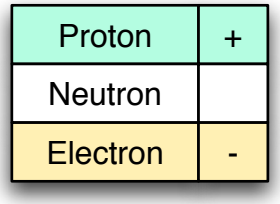
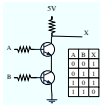


Figure A.1: An Atom's Parts and its Charges



The Buzz...

Subatomic particles

Long before we knew anything about atoms, ancient Greeks were referring to the concept of atoms (meaning "indivisible") as the smallest pieces of matter. With the discovery of the electron 1890s, scientists quickly found that atoms were actually made of orbiting electrons and fixed nucleons (protons and neutrons).

Throughout the early 20th century, scientists began finding even smaller subatomic particles. Quarks and neutrinos were soon discovered as even smaller subatomic particles contained within protons, neutrons, and electrons. This led to discoveries of classes of subatomic particles, such as bosons, leptons, and gluons. The sheer number of particles and interactions is staggering. Niels Bohr once commented that "a person who wasn't outraged on first hearing about quantum theory didn't understand what had been said".

portion. The electrons float around this nucleus, like in Figure A.2, on the next page.

This planetary model was created in 1904 by Japanese physicist Hantaro Nagaoka and was based on little more than clever guesswork. It was later refined by physicist Niels Bohr in 1913. Bohr found that electrons resided in well defined orbits, and that the electrons had the orbits had discrete energies, meaning that only certain electron orbits were allowed. According to Bohr's theory, an electron that moved between orbits would disappear from one and reappear in another without visiting the space in between (known as a *quantum leap*).

While revolutionary for its time, Bohr's model is little more than basic classical mechanics with some specifics related to the quantization of electrons. Today we know that the atom is more sophisticated. However, Bohr's model still gives a good enough picture of what's happening at the atomic level as it relates to the everyday world that it's still taught at the introductory level.

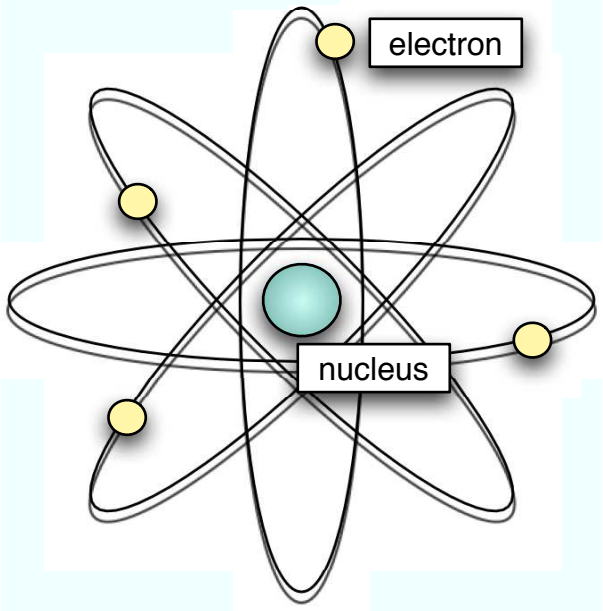


Figure A.2: Basic Atomic Model

The Charge Property

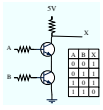
Subatomic particles posses a fundamental property known as *electrical charge*. One such property is electrical charge. Protons exhibit positive charge, electrons exhibit negative charge, and neutrons exhibit no charge. For the purposes of the study of electricity we can effectively ignore the presence of neutrons and focus solely on protons and eletrons.

Scientists have studied the proton and decided to call its amount of charge the *elementary charge* (abbreviated *e*). The actual amount of charge is defined to be 1.602e-19 [1] Coulombs, but that's not a very fun number to remember. It's easier to just say *e*.

It also just so happens that an electron has -*e* charge. That is, protons and electrons have equal but opposite in sign amounts of charge.

When grouped together, a proton and an electron's total charge cancel each other out. As a pair, they are electrically neutral.

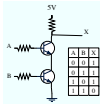
1. That's 0.00000000000000000001602



The Buzz. . .

What is electrical charge?

Electrical charge is a fundamental property of nature that is possessed by some subatomic particles. It not something that can be felt or seen; instead, it is a concept that is the result of observation of how these particles interact with each other. Placed in close proximity, subatomic particles exert physical forces on each other. This interaction is explained by the charge property contained in the particles.



The Buzz. . .

Protons are Positively Charged

The decision to call electron charges negative and proton charges positive are the result of a historical convention.

The static charge produced on rubber was known as negative charge while static charge produced on glass was known as positive charge. Once atoms were discovered and studied, scientists found that electrons had the same polarity as the charge on rubber and protons had the same polarity as the charge on glass. So they considered electrons to be negatively charged and protons to be positively charged.

Electric charge is considered a *firstuse* property. This is a fancy way of saying that e amount of charge is the smallest amount we can possibly have, so we can base everything off of this smallest discrete amount. Since we can only have a discrete number of protons and electrons in matter, so we will always have a whole multiple of e charge.

A.2 Elementary Education

The most basic way of distinguishing types of atoms from each other is through the number of protons the atom contains. This property

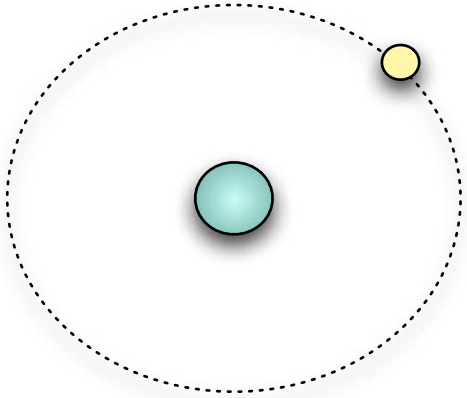


Figure A.3: Hydrogen Atom

is known as the *atomic number*. The most simple atom, with an atomic number of one, is Hydrogen. It has a single proton and a single electron, as seen in Figure A.3.

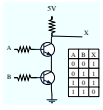
Helium, with an atomic number of two, has two protons and two electrons in its atomic structure. This trend continues through the approximately 116 different atomic types.

Electron Shells

We already know that electrons float around the nucleus of the atom. But one interesting fact is that they tend to float around the nucleus in shells. As the atomic number goes up, an atom having many electrons will have multiple electron shells.

The closest shell to the nucleus can have up to two electrons. Hydrogen, with atomic number 1, has one lone electron in this shell. Helium, with atomic number 2, has two electrons in this shell. The next element, Lithium, with three electrons, has its inner shell (with two electrons) full, and has an outer shell of one electron.

The maximum number of electrons in a shell can be seen in Figure A.4, on page 223. Unfortunately, the electrons don't always fill in the shells completely before creating a new shell. The reason is due to differing



The Buzz...

What holds the atom together?

One thing you probably do remember from high school chemistry (or Paula Abdul music videos) is that like charges repel and opposite charges attract. This means that protons and electrons are attracted to each other - and this is good, because it helps hold the atom together. However, in atoms with multiple protons it seems reasonable that the protons would repel each other and force the atom apart. So what gives?

Electromagnetic and gravitational forces are two of the four fundamental forces in nature. The other two are the strong nuclear force and the weak nuclear force. The strong nuclear force is what holds the protons together. There are differing theories as to how this is accomplished, all of which are too complicated to think about in this book. But there is something that holds it all together. For more information about the nuclear forces, I recommend referencing a good high school or college level physics textbook. The website http://particleadventure.org also has a lot of great information about this very topic.

energy levels in different shells types. Thankfully, it is unimportant for us to know in detail. [2]

Instead, we're really only interested in the *outer* most shell. This outer shell is also know as the *valence* shell.

Valence Shell

The valence shell will never hold more than eight (8) electrons. When the valence shell fills up, either a new valence shell starts to form or an inner shell starts to get more electrons.

We can refer to an atom by its *valence number*, which is simply the number of electrons in its valence shell. Copper, for instance, has a

2. There are good resources on the web explaining how to figure the electron shell configuration out. There's also a good cheat sheet periodic table at http://www.chemicalelements.com/show/electronconfig.html

Shell n	Maximum # of Electrons
1	2
2	8
3	18
4	32
5	50
	$2n^2$

Figure A.4: Shell and Electron Count

valence of +1 because there is one electron in the outer shell. Hydrogen also has a valence of +1. Carbon has a valence of +4.

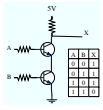
With the exception of Hydrogen and Helium, the objective of each atom is to reach a valence number of 8. Atoms with valence number 8 tend to be very stable. Because of this, the valence number can also be considered to be the number of electrons needed to make 8. We know that copper has a valence of +1, but we could also refer to it as a valence of -7. Carbon has a valence of +4 or -4. All of the noble gases (see The Buzz, on the following page) have a valence of 0.

The valence number indicates how easily the atom gains or loses electrons. A valence number of +1 means the atom can easily lose the single electron in the valence shell. A valence number of -1 means that the atom tends to want to take an electron to fill its shell in.

A.3 Materials and Bonding

Atoms rarely tend to stick around by themselves. In general, they like to bond up with other atoms. In fact, they tend to try and fill up their valence shells by sharing electrons with other nearby atoms.

The most common bond, the covalent bond, happens when an atom



The Buzz...

Noble Gasses

The *noble* gasses consist of helium, neon, argon, krypton, xenon, and radon. They are considered inert because they don't easily react with other elements to form compounds. Their lack of reactivity is directly related to their full valence shells. Since they have no tendency to gain or lose an electron, they tend not to react with other elements to form into larger compounds.

with a less than full valence shell pairs up with another atom with a less than full valence shell. It turns out, valence electrons like to join up to make electron pairs—and in the process they bond the atoms together. For example, in Figure A.5, on the next page we see that the four valence electrons from a carbon atom join up with four hydrogen atoms, each with one valence electron, to form CH_4, methane. The carbon atom filled up its valence shell, to eight electrons, and the hydrogen atoms filled up their valence shell to two atoms each.

Another bonding form of interest to our electricity knowledge is metallic bonding. Elements that are typically known as metals have few valence electrons. Most metals have just one or two lone electrons in their valence shell. This relatively empty valence shell causes metals to bind very tightly, because each atom is trying to grab other atoms to fill in the empty spots in the valence shell. When metals form into solids, the structure of their bonds is such that these valence electrons are loosely bound to any particular atom.

As an example, take a look at the element copper's electron configuration in Figure A.6, on page 226. It has just one lone electron in its valence shell. When copper atoms bond together to make a solid piece of metal, they join up via metallic bonding and each atom in the structure has this loosely bound valence electron. With just a little added energy, it's possible to move one of these electrons to another atom, which then moves to another, and another. These easily liberated electrons are what make metals good electrical conductors, as we'll see in Section 2.2, *Conductors and Insulators*, on page 17.

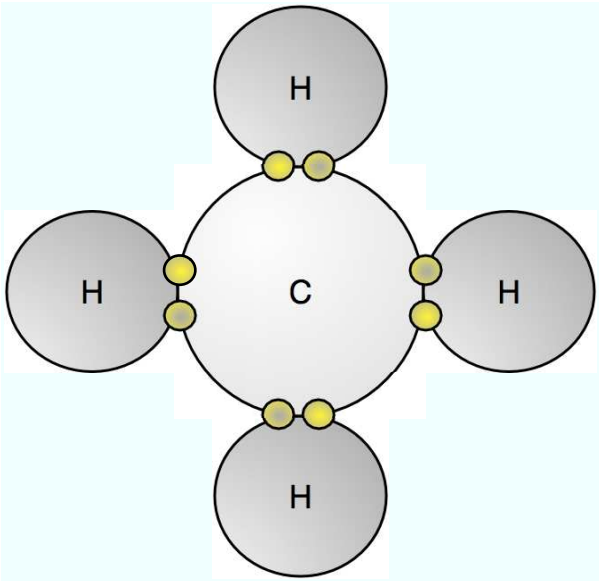


Figure A.5: Covalent bonding in Methane

A.4 Just a little spark

When we talk about electricity, what we are actually talking about is electrical current. Electrical current is the flow of electrical charge; that is, it's nothing more than moving some charge from one place to another - just like the flow of water.

So far, at least from this book, we know of two items that have electrical charge—protons and electrons. This means that we can create an electrical current by either moving protons or electrons. Based on atomic structure, it's relatively difficult to grab a proton out from the nucleus and move it somewhere else. In materials like metals, however, it's relatively easy to move an electron.

Now, don't misunderstand. Electrons, for the most part, don't just move around on their own—at least not in a controllable way. In order to make them move we have to give them a reason. That is, we have to give them some energy in order to cause them to move. This energy can come in a number of forms - light, heat, or something called an electric field.

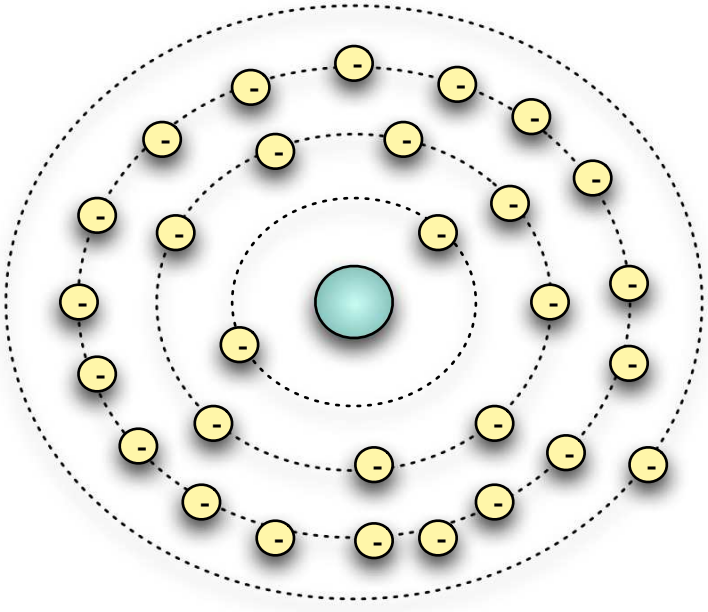
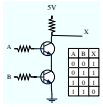


Figure A.6: Copper Atom Electron Configuration



The Buzz...

I moved some atoms. Did I make electricity?

Based on the notion that electricity is the same as moving electrical charges, for some of us it may seem logical that by moving some atoms, say for example a glass of water, from one location to another we've just moved electrical charges and thus created electricity.

However, recall that protons and electrons have the same, but opposite in sign, amounts of electrical charge. Also recall that atoms have the same amount of protons and electrons. This means that an atom has a total net charge of 0. Physically moving an atom does not create electrical current.

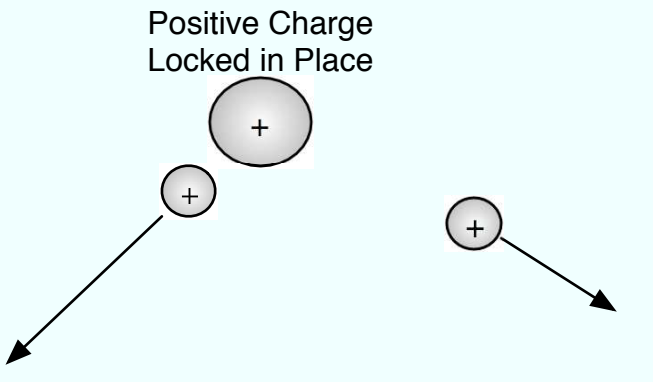


Figure A.7: Reaction of charges near a fixed charge

A.5 Electric Fields

Earlier we discussed the fact that like charges repel and that opposite charges attract, or create a force on one another. The amount of force is dependent on how much charge (quantity) there is and how far apart the charges are (distance).

If we were able to pin down some charge so that it doesn't move, we could then analyze the effect it would have on other charges nearby. In Figure A.7 we see that for a fixed piece of positive charge, other positive charges nearby are repelled. Closer charges are repelled with more force that charges farther away. This force is what creates an electric field.

Another possible model is a small tunnel, in which one of the ends has excess positive charge lined up. In this case, as shown in Figure A.8, on the next page other positive charges in the tunnel are repelled. If there were negative charges, they would be attracted to the end of the tunnel.

Electric Potential

We could go on all day with different pictures and scenarios, but the point is that if we were able to put some fixed electrical charge somewhere it would cause a reactionary force on other electrical charges nearby. Thus, this fixed electrical charge has electrical *potential*. Its presence creates an electrical field around itself that has the potential to move other electrical charges around, if they were present.

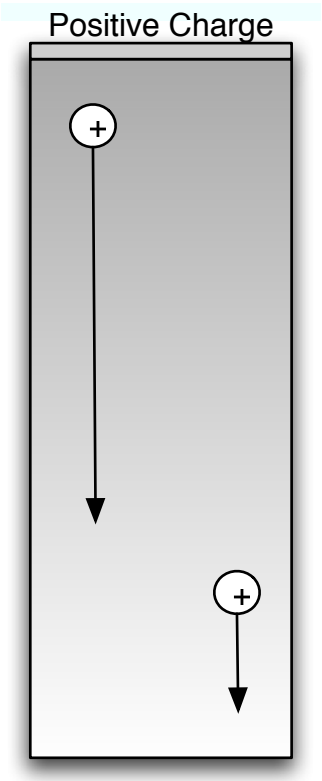


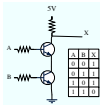
Figure A.8: Electrical Charges in a Tunnel

This concept of electric potential is very important. Remember, potential means that something *could* happen. In this case, if we put a piece of charge somewhere it creates an this electrical field around itself that would potentially have an effect on any other charges nearby. The electric field is stronger closer to the source charge and gets weaker as the distance from the source increases. The strength of the electric field is also directly proportional to the total amount of charge at the source.

A quick review

Based on all of our knowledge, we now know enough to conclude that:

1. In a piece of metal, like copper, there are electrons that are loosely bound to the nuclei of the atoms in the structure.



The Buzz...

Quantum Leaps

The concept of electron shells is a simplification of what's really going on in an atom. Electrons exist in states known as *orbitals* which form more of an electron cloud than sets of rings. Electrons fill in orbital positions in a known way, which follows a structure of increasing energy levels in the atom.

It's possible to bombard the atom with a certain amount of energy and excite an electron into moving from its "rest" state to a new state. In doing so, the electron makes a *quantum leap* to its new position. The new home of the electron may only be temporary, as the electron may release some energy back outside of the atom and move back to its original position.

What makes quantum leaps so fascinating to scientists is the fact that the electron cannot exist in any state in between the two it moves between. That is, there is no continuity between the two states; the electron is either in one state or the other. Furthermore, the amount of energy required to cause the electron to move is also quantized. If not enough energy is added, the electron will not move at all.

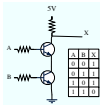
2. These loosely bound electrons can be moved if they are provoked with a little energy.
3. One way of providing energy to move the electrons is subjecting them to an electric field.

And that is how electricity is born.

A.6 Magnetism

A closely related phenomenon to electricity is *magnetism*. We've all experienced magnetism, and probably played around with horseshoe magnets, iron filings, and compass needles. But what is magnetism? Where does it come from?

Electrons have a property called *spin*. Spin is rotational momentum of a body about itself. For example, the earth spins about its own axis every



The Buzz...

The Earth's Magnetic Field

The earth produces a magnetic field which is what allows compasses to align themselves to the magnetic north pole. origin of the field is still somewhat of a mystery, but it is believed to be the result of electrical current due to the motion of liquid metals in the earth's core.

twenty four hours. Distinguishing this type of rotation is important, because the earth also revolves around an external point (the sun).

Electron spin is somewhat related to this idea. It is an intrinsic property of the electron, much like charge. The idea of spin is a bit difficult to explain; in fact, for many years physicists were quite perplexed on how to explain the math behind what they observed with electromagnetism. It wasn't until Einstein introduced his Special Relativity theory in 1905 that an explanation was finally brought forward.

Suffice to say that at the electron level, magnetism is a result of the spin of the electron. But magnetism goes much further than this. Electric fields, which we discussed previously, are very closely related to magnetic fields. In fact, changing electric fields create changing magnetic fields—something we'll see more of in Section 3.1, *Magnetic Motion*, on page 35. Magnetic fields are simply the result of the motion of electrical charges.

There are also substances, like iron, which exhibit properties of magnetism. This is because the electrons which are capable of producing magnetic dipoles, and thus magnetic fields, are aligned with each other.

So, in summary, know that electric fields and magnetic fields are closely related and in some cases energy can transfer between the two.

A.7 Sources of Electricity

In order to be useful in electricity, we must separate electrons from the nucleus in order to make the potential difference that causes current to flow. This can be accomplished in a number of ways.

Chemical Conversion

As we've already seen, batteries are one way we can create a potential difference. Inside, a chemical reaction creates charge on two different terminals. This chemical process is explained in more detail in Chapter 3, *Electrical Power*, on page 34.

Ionic Conversion

Ions are substances that have net electric charge. They can be as simple as a single electron or proton. Negatively charged ions are *anions* while positively charge ions are *cations*. Anions have more electrons than protons while cations have fewer electrons than protons.

The process of becoming an ion is known as *ionization*. In one form of ionization, the electron is taken away from the element leaving behind a cation. In fact, since most metals have a very loosely bound outer electron they are generally viewed as a grid of cations with excess electrons floating around.

Electromagnetic Conversion

Magnetism and electricity are closely related. It's possible to create magnetism via electricity and electricity via magnetism as we'll see in Section 3.1, *Magnetic Motion*, on page 35.

Static Electricity

By using friction, we can cause electrons on an insulator to become separated. These electrons can accumulate and result in a net charge on the surface of an object. During a discharge, the electrons flow and create electricity.

Light Conversion

We can use light to bombard some materials and cause them to emit electrons when hit.

Thermal Conversion

When heated, some materials will cause electrons to flow in a controllable way.

The Pragmatic Bookshelf

The Pragmatic Bookshelf features books written by developers for developers. The titles continue the well-known Pragmatic Programmer style, and continue to garner awards and rave reviews. As development gets more and more difficult, the Pragmatic Programmers will be there with more titles and products to help you stay on top of your game.

Visit Us Online

A Peek at Computer Electronic's Home Page
http://pragmaticprogrammer.com/titles/ctelec
Source code from this book, errata, and other resources. Come give us feedback, too!

Register for Updates
http://pragmaticprogrammer.com/updates
Be notified when updates and new books become available.

Join the Community
http://pragmaticprogrammer.com/community
Read our weblogs, join our online discussions, participate in our mailing list, interact with our wiki, and benefit from the experience of other Pragmatic Programmers.

New and Noteworthy
http://pragmaticprogrammer.com/news
Check out the latest pragmatic developments in the news.

Contact Us

Phone Orders:	1-800-699-PROG (+1 919 847 3884)
Online Orders:	www.pragmaticprogrammer.com/catalog
Customer Service:	orders@pragmaticprogrammer.com
Non-English Versions:	translations@pragmaticprogrammer.com
Pragmatic Teaching:	academic@pragmaticprogrammer.com
Author Proposals:	proposals@pragmaticprogrammer.com