

Introduction
to
Electronics and Computers
Vol 1: Hardware

Draft XIV
Brian Collett
Physics Department
Hamilton College
Copyright Brian Collett 1997-2014

Table of Contents

Chapter 1:Introduction	1
Chapter 2:Basic electrical concepts.....	3
2.1 The Properties of Charge.....	3
2.2 The physics of current flow	7
Chapter 3:Simple Components.....	13
3.1 Ideas of components.....	13
3.2 Ground.....	13
3.3 Wires.....	14
3.4 Switches.....	15
3.5 Power Supplies	16
3.6 Resistors	17
3.7 Measuring Instruments	20
Chapter 4:Simple DC Circuits.....	23
4.1 Introduction	23
4.2 Simple Circuits	23
4.3 Resistors in Series	24
4.4 Resistors in parallel	25
4.5 Combinations of Series and Parallel	26
Chapter 5:Formal Analysis of DC Circuits	33
5.1 Introduction	33
5.2 Kirchhoff's Laws.....	33
5.3 Method 1: Node analysis.....	34
5.4 *Method 2: Loop Analysis	36
5.5 Thévenin's Theory.....	37
5.6 Solving Resistor Problems with PSpice	41
Chapter 6:Time Varying Voltages	49
6.1 Introduction	49
6.2 Periodic Voltages	49
6.3 Sinewaves.....	51
6.4 A little helpful notation.....	52
6.5 Time varying voltages and resistor circuits.....	52
6.6 Non-Sinusoidal Voltages	53
Chapter 7:The Capacitor	55
7.1 Introduction	55
7.2 Resistor-capacitor circuits	56
7.3 The Capacitor and the Sinewave	58
7.4 RC Circuits and Sinewaves	61
7.5 The Bode Plot.....	62
7.6 Thévenin and capacitors	64
7.7 Common uses of capacitors.....	64
7.8 AC PSpice Simulations	66
7.9 Real Capacitors.....	68
Chapter 8:R-C Frequency Selective Circuits.....	73
8.1 Introduction	73
8.2 Low-pass Filter.....	73

8.3 RC High-pass Filter	75
8.4 Tone Controls	76
8.5 Bandpass Filter	76
8.6 Multi-section Filters	77
8.7 Studying a New Filter with PSpice.	78
Chapter 9: The Diode	85
9.1 Introduction	85
9.2 The ideal diode	85
9.3 The Real Diode.....	86
9.4 Some Common Diode Circuits.....	87
9.5 Diode Characteristics	88
9.6 Special Kinds of Diode.....	89
9.7 Using PSpice to Study a Diode	91
9.8 The Physics of Diodes	94
9.9 Manufacturing a diode.....	96
Chapter 10: Power Supplies I	99
10.1 Introduction	99
10.2 Transformers.....	99
10.3 Rectifiers.....	101
10.4 Smoothing the output	103
10.5 Simulating the Full-Wave Rectifier	105
Chapter 11: The Field-Effect Transistor	111
11.1 Introduction	111
11.2 The FET	111
11.3 The water model MOSFET	112
11.4 3-terminal device characteristics	112
11.5 Characteristics of an FET	113
11.6 Simple FET circuits	115
11.7 Thévenin Models of an FET	117
11.8 Using PSpice to Study a MOSFET.....	118
11.9 The Physics of FETs	120
Chapter 12: FET switches.	125
12.1 Introduction	125
12.2 Power Switches	125
12.3 Logic Switches	126
12.4 FET Logic Switches	127
12.5 More complicated gates.....	130
12.6 Connecting Switches Together	131
12.7 Integrated Circuit Switches	135
Chapter 13: Digital Logic Theory	141
13.1 Introduction	141
13.2 Truth Tables	141
13.3 Some basic gates.	143
13.4 Multi-gate circuits	144
13.5 Boolean Algebra.....	146
13.6 Logic simplification.....	149
13.7 Logic design	150
13.8 Putting it all together	153

Chapter 14:Combinatorial Functions—Coders, Decoders and Arithmetic	157
14.1 Medium Scale Logic	157
14.2 Coders/decoders	157
14.3 Encoders	161
14.4 Multiplexers/demultiplexers.....	162
14.5 Arithmetic Logic.....	164
14.6 Building Digital Integrated Circuits	168
Chapter 15:Programmable Logic.....	173
15.1 Introduction	173
15.2 The PAL.....	174
15.3 The GAL.....	176
15.4 Programming GALs	177
15.5 A tutorial example using PALCMPL.....	181
Chapter 16:Sequential Logic	185
16.1 Introduction	185
16.2 The Flip-flop.....	185
16.3 The transparent latch	188
16.4 Logic diagrams for flip-flops.....	188
16.5 D-type flip-flops	189
16.6 The J-K flip-flop	191
16.7 Simple counters	191
16.8 Some other Chips	195
Chapter 17:Synchronous Logic	199
17.1 Introduction	199
17.2 Glitches.....	199
17.3 General Synchronous Systems	201
17.4 *J-K Synchronous	205
Chapter 18:Amplifiers	209
18.1 Introduction	209
18.2 Gain	209
18.3 Thévenin Model of an Amplifier.....	214
18.4 Common-Source FET Amplifier	215
Chapter 19:*Multi-Stage Amplifiers	223
19.1 The difference amplifier	223
19.2 The FET Source Follower	226
19.3 The Complementary Source Follower.....	229
19.4 A Complete Multi-Stage Amplifier.....	233
Chapter 20:The Ideal Operational Amplifier.....	239
20.1 Introduction	239
20.2 The operational amplifier	239
20.3 Feedback.....	240
20.4 The non-inverting amplifier.....	241
20.5 The Golden Rules.....	242
20.6 Virtual Ground.....	243
20.7 A few good circuits.....	245
Chapter 21:Real Op-Amps.....	251
21.1 Introduction	251
21.2 Frequency Response.....	251
21.3 Output Current Limit.....	254

21.4 Input Characteristics.....	257
Chapter 22:Comparators	263
22.1 The Open-Loop Comparator	263
22.2 Hysteresis to the Rescue.....	263
22.3 Commercial comparators	266
Chapter 23:Oscillators.....	269
23.1 Introduction	269
23.2 Linear Oscillators	269
23.3 Non-linear Oscillators	273
Chapter 24:Linear Regulated Power Supplies	277
24.1 Introduction	277
24.2 Voltage references.	277
24.3 IC references.....	278
24.4 A simple voltage regulator.....	279
24.5 Current regulation.....	281
24.6 IC voltage regulators	281
24.7 A complete power supply design.....	284
Chapter 25:Digital-to-Analog Conversion.....	287
25.1 Introduction	287
25.2 The DAC	287
25.3 The R-2R Ladder.....	288
25.4 Commercial DAC chips	290
25.5 Imperfections of DACs	291
25.6 Some DAC examples	293
Chapter 26:Analog-to-Digital Conversion.....	297
26.1 Introduction	297
26.2 Flash Conversion.....	298
26.3 A complete 3-bit flash ADC.....	298
26.4 Successive Approximation Conversion.....	299
26.5 Imperfections of ADCs.....	302
26.6 IC Converters	302
26.7 ADC Examples	304
Chapter 27:Power Switches	309
27.1 Power Switches	309
27.2 Moderate Power Switch	309
27.3 Switching capacitive and inductive loads	310

Chapter 1:Introduction

We live our lives surrounded by electronic equipment. We are awoken by digital clock radios, strap digital watches on our wrists, eat breakfast in front of the television, travel to school or work listening to radios and CD players and then sit down in front of computers to do our work. On our way home, the iPod running loudly, we realize that we will be late and use our cell phones to call home. When we get home we eat food cooked in a microwave and then go surf the web for an hour or so before bed.

The scientist is even more surrounded by electronics. The chemistry lab is no longer mainly a place of gleaming glassware. Instead, it is full of electronic balances, pH meters, lasers, and NMR spectrometers. The biology lab has its spectrophotometers, more balances and pH meters, gas chromatographs, SDS gel systems, and electron microscopes. The geologist has electron microscopes, gas chromatographs, X-ray machines and seismographs while practically everything the physicist touches is electronic.

Many of the pieces of apparatus, both in the lab and in everyday life, are black boxes. We understand what they do; we may understand the principles by which they work; but we don't think we need to know the details of their circuits. However, only when we have a good knowledge of that circuitry that we can really appreciate what the instruments do and can understand their limitations.

Consider the process of reproducing music using a CD. In the beginning, there were sound waves traveling through the air. A sound wave is rapid variation in the pressure of the air whose shape depends on the nature of the sound. Those sounds fell on a microphone that translated the varying air pressure into a varying voltage whose shape was just the same (Figure 1-1).

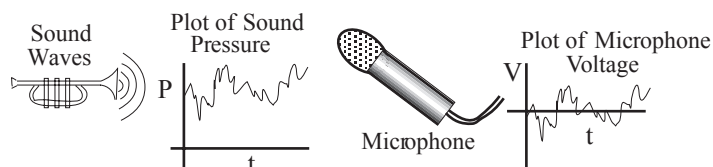


Figure 1-1 Capturing Sound

The microphone is a transducer. It translates the sound pressure information into voltage information, ideally without altering its shape at all. The voltage is then amplified and converted to a stream of numbers by an analogue-to-digital converter (ADC)

The ADC takes a voltage that is continuous in both time and space and samples it into a set of numbers that is discrete in both time and space. Obviously this means throwing away almost all of the information so we have to be careful to keep enough that our ears will not be able to tell that there is anything missing. That means that we have to take many samples per second (more than 40,000 samples per second) and we have to convert each one into a very accurate number, usually 16 bits long. The string of numbers representing the voltage is fed into a computer that performs some mathematical tricks to encode the information. Essentially, the computer adds extra copies of some of the information and then writes it to the CD as a pattern of flat and dented areas.

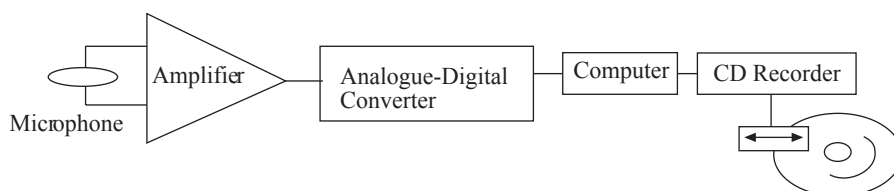


Figure 1-2 Making a Compact Disk

The CD player reads the pattern off the disk. This process is error prone so the computer in the CD player uses the extra information to correct the errors and reconstruct the original string

of numbers. The numbers themselves mean nothing to us so the computer in the CD player sends them to a digital-to-analog converter (DAC), which turns them back into voltages. The reconstructed signal is not quite the same as the original. Where the original was smooth, the copy consists of a lot of little steps so we pass it through a filter to remove the steps. Finally, the signal is amplified and converted back into sound by a loudspeaker, another transducer, and we hear it as sound.

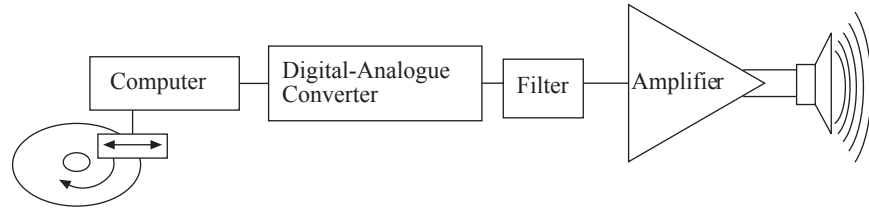


Figure 1-3 Playing a Compact Disk

Most electronic devices follow a similar path. They measure something in the world, turning it into a time varying voltage called a signal. They process that signal in some way, extract information from it, and present that information to us. Table 1-1 gives some examples.

Table 1-1: Some Familiar Electronic Devices

Instrument	Input Signal	Input Transducer	Processing	Output Transducer	Output Signal
CD Recorder	Sound	Microphone	Digitize and Encode	Laser	Light to burn pits in CD
CD Player	Light from CD pits	Photo diode	Decode, convert to analog and filter	Loudspeaker	Sound
Cell Phone	Sound	Microphone	Amplify, digitize, modulate onto RF	Antenna	Radio Wave
Thermostat	Temperature	Thermistor	Compare actual temperature to preset limits, taking time of day into account, and turn heat on/off.	Switch	Power to furnace or A/C
Smoke Detector	Smoke and Fire	Photodiode /Am ion cell	Compare level with preset limits	Siren	Sound
pH meter	ions in solution	pH electrode	Amplify signal, digitize, and display	LCD display	Visual image
Television	Radio Wave	Antenna	Amplify, decode, split into sound and picture	Loudspeaker & LED, LCD, or plasma screen.	Image and Sound

Chapter 2:Basic electrical concepts

2.1 The Properties of Charge

Electronics is the science of making electric charges do interesting and useful work, everything from lighting flashlights to running computers. At the heart of all electronics is the idea of electric charge. There are several ways of thinking about charge, some which are physically realistic and some which are not but which are very convenient because they use our physical intuition about how the world works. In the first half of the chapter we will take the very practical, unrealistic, view that is usually adequate for understanding electronics and leave the more physical view to the second half of the chapter, which can be omitted at a first reading.

2.1.1 Charge

Although physics tells us that there are two kinds of electrical charge, which we call positive and negative charge, and that such charges come only in very small, indivisible packets, that detailed view of charges is very rarely needed to understand or design electronic equipment. Instead we can think of charge as an invisible kind of fluid that can move around inside certain kinds of materials, mostly metals. We refer to charge using the symbol Q and we measure the amount of charged fluid in units called Coulombs. One Coulomb is an arbitrary, fundamental unit in the same way that a meter is an arbitrary, fundamental unit of length—once you define that unit then you can make measurements. Using the unit, two people can compare measurements in a meaningful way but there is nothing that is special about the unit that you choose; it is merely a convenient tool.

We will think of charge as a kind of invisible fluid with certain properties-

1. You cannot make or destroy charge in any electronic apparatus; all you can do is to move it around.
2. Charge can move around more or less freely within some materials (such as most metals or salty water) but cannot move in or pass through other materials (such as air, wood, plastics, etc.). We call the materials through which current can pass Conductors and the ones through which it cannot pass Insulators.
3. When charge moves it does work, either useful work (such as lighting up your room) or useless work (such as the processor chip in a computer getting so hot that it has to have a fan or it will melt itself!).
4. When charge moves, it creates a magnetic field. When a magnetic field moves near a metal wire, it causes charge to move in the wire.
5. Charge comes in two kinds, positive and negative. If you mix the two kinds then they cancel each other out.

2.1.2 Current

Moving charges form what we call Electric Current. We measure current by the amount of charge that passes a point in 1 second. The symbol for current is I and the unit of current is the Ampere—when a current of 1 **Ampere** is flowing, a charge of 1 Coulomb passed each point in 1 second. We have the relationships

$$I=Q/t \text{ and } Q=I \times t$$

where Q is the charge passing a point in time t .

Info The Coulomb is named after the first person to measure the sizes of charges. The Coulomb is to charge what the liter is to a fluid, a convenient reference amount with no particular physical meaning. Like the liter, the Coulomb is tiny when you think about a power station but treasonable when you think about a battery and enormous when you think about a single atom. A moderate power station will generate tens of thousands of Coulombs a second while a portable CD player may run for an hour or so on a single Coulomb. It takes 6×10^{18} elementary charges to make a single Coulomb (compared to 3×10^{25} water molecules in a liter of water).

Info The two different kinds of charge are extremely important in physics but they play very little role in electronics. In electronics we can almost always think of a single fluid that flows from the positive terminal to the negative and not worry about the real details.

Info Like the Coulomb, the Ampere is a moderate, every day, sort of unit. A typical appliance will use a few amperes of current; a vacuum might use 7A, a light bulb 1A, and a hair dryer 10A. However, we find an enormous range of currents in regular use. At one extreme, the US power grid carries millions of amperes around at any time and a bolt of lightning may carry tens of thousands of amps. At the other extreme, \$20 will buy us a component that can measure currents smaller than 1pA

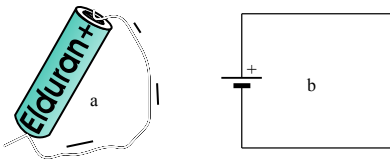


Figure 2-1 Wire and Battery Circuit

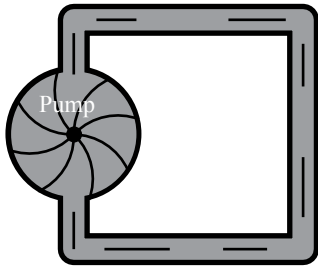


Figure 2-2 Water Model of a Circuit

Info We can measure current in several ways. The most common way is to pass the current through a small resistor and measure the voltage across the resistor with a sensitive electronic voltmeter. Older instruments pass the current through a small coil and use the magnetic field that it generates to move a needle. All of these instruments have to be inserted into the circuit. There are also instruments that can measure the current flowing in a wire without breaking the circuit to insert the instrument. They measure the small magnetic field that the current sets up round the wire.

Note Although we normally speak of conventional current, a current of positive charges flowing from the positive terminal of a battery to the negative terminal, in practice current is carried by negative electrons traveling in the opposite direction. Fortunately we hardly ever have to worry about this. We will stick to general practice and use conventional current.

In order to make charges move there must be a source of energy to act as a pump of the electrical fluid. The usual sources are either batteries, which convert chemical energy into electrical pumping action, or generators, which convert mechanical energy.

Because of property 1, electrical current can flow only in complete circles made of conducting materials, this gives rise to the term Electrical Circuit.

The simplest electrical circuit consists of a battery and a piece of wire as shown here in two different forms. In Figure, a is a sketch of the physical appearance of the system and b is a schematic diagram that uses symbols to show the way the circuit is connected but which ignores the physical appearance.

It can be very helpful to think of the charge as being like water, wires as being pipes, and batteries as being pumps. Using these ideas we can model our circuit as shown in Figure 2-2.

So long as the circuit is complete, the pump can keep the fluid circulating and the current flows. What happens if we disconnect the wire? A disconnected wire is not like a broken pipe—the electrical fluid cannot run out because air is not a conductor—instead it is like a blocked or closed-ended pipe. Now the fluid cannot flow past the blocked end and the current in the circuit is zero.

This analogy helps us understand what happens when we first connect a wire to a battery, before the circuit is complete. This is like connecting a closed end pipe to the pump while the inlet of the pump is blocked. In this case the pump will push water into the pipe until it reaches the blocked and then stop. There will be a very short-lived current and then a new equilibrium with no current flowing anywhere in the broken circuit.

As soon as we connect the free end of the wire to the other terminal of the battery, that is, we connect the end of the pipe to the inlet side of the pump, the current immediately starts to flow all round the circuit. Note that the current starts to flow immediately even though it may take several seconds for the first water that leaves the pump after the connection is made to get to the end of the pipe.

2.1.3 Voltage

The only way to make water flow round a set of pipes is to exert a force or more properly a pressure, on it. Similarly, the only way to make an electrical current flow in a circuit is to exert an electrical pressure on it. We call that electrical pressure Electric Potential and we measure it in units of Volts.

Electrical potential, like pressure in a fluid system, is a relative quantity—we can only measure the potential difference between two points in a circuit. Thus an instrument for measuring potential, a Voltmeter, has two terminals (usually colored red and black) and measures the potential difference between the terminals. If the pressure tries to push charge into the red terminal and out of the black one then we say the potential difference is POSITIVE and if the pressure tries to push the charge into the black terminal and out of the red then the potential difference is NEGATIVE. In order to make a measurement you MUST connect both terminals of the meter because it can only measure the potential difference between two points.

Having said that, in practice we often speak of voltage at a point in the circuit, just as we often speak of the pressure at a point in a fluid system. We can do this by fixing on a single point in the circuit as a reference and making all measurements between that point and the point of interest as shown below. This means that the point is automatically at zero volts since the potential difference between that point and itself must be zero. We call this point ground and it is usually chosen to be at the negative terminal of the battery or other voltage source powering the circuit.

2.1.4 Power

When current flows it does **work**, either usefully or producing waste heat. We call the rate at which work is done **power**. The power dissipated in an electronic device depends on both the amount of current that flows and on the voltage driving it according to the formula

$$P=I \times V$$

where I is the current through the component and V the voltage drop across it.

We measure energy in **Joules** and power in **Watts**.

$$1 \text{ Joule} = 1 \text{ Watt} \cdot 1 \text{ sec} \quad \text{or} \quad 1 \text{ Watt} = 1 \text{ Joule} / 1 \text{ Sec}$$

so a current of 1 Amp flowing through a potential difference of 1 Volt generates a power of 1 Watt and so does work at the rate of 1 Joule per second.

Example

A radio that draws 30mA = 0.03A of current from a 9 volt battery uses energy at a rate of

$$P=0.03\text{A} \times 9\text{V}=0.27\text{Watts.}$$

2.1.5 Resistance

A potential difference causes a current flow. In the mid 19th century Georg Ohm found that, for most common materials under normal conditions, the amount of current that flows is proportional to the voltage difference that makes it flow. If we double the driving voltage then the current flowing will double, if we halve the driving voltage then the current halves. In such a material, we can make a graph of the relationship between the current flowing and the voltage drop making it flow. The graph will be a straight line like that in Figure 2-3.

This kind of graph, called an **I-V curve**, is one of the most useful tools for describing the behavior of electronic components. This is the simplest kind, a straight line. A component with this behavior is called a Linear Component and a circuit made up only of linear components is a Linear Circuit. Notice that we plot I as a function of V and not the other way round because in most circuits we directly control the applied voltage (for example, we choose the battery) and then we measure the current that results.

Such a component system is described mathematically by the relationships

$$V \propto I \text{ or } V=k \times I$$

where the constant of proportionality, k, is a characteristic of the particular circuit. We call the constant **Resistance** (symbol R). This linear relationship is called **Ohm's Law** and it can be written in several different forms depending on what information we have and what we want to know. If we know the current flowing in a known resistor and want the voltage across the resistor then we use

$$V=I \times R$$

If we know the voltage across the resistor and we know its value then we can find the current flowing through the resistor using

$$I=V/R$$

Finally, we can find the resistance if we know the voltage and the current using

$$R=V/I$$

The unit of resistance is called the **Ohm**, symbol Ω . A resistor of 1 Ohm lets through 1 Ampere when there is 1 Volt across its terminals.

Example

For example, if we connect a 220 Ω resistor across the terminals of a 1.5V battery then we can find the current that flows in the resistor using the second form of Ohm's law.

$$I = V/R = 1.5\text{V} / 220\Omega = 0.0068\text{A} = 6.8\text{mA}$$

Note The Joule is another of our standard units of measurement. It is the amount of energy that it takes to lift a 1 kg mass through a vertical distance of 10 cm. It is thus a fairly small unit of energy. For example a small container of yogurt supplies about 920,000 Joules of energy! The Watt is a similarly small unit of power. A single electric light bulb uses 60-100 Watts of power; a hairdryer about 1000-1200 Watts. Another common unit of power is the HorsePower. 1 HorsePower = 750 Watts.

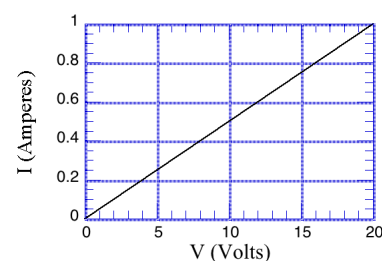


Figure 2-3 I-V plot for a linear resistance

Note There is an old trick for remembering the various forms of Ohm's law. You write the law in a triangle as shown in Figure 2-4. Then, to find any one term, you cover that term up with a finger and what is left is the formula. So, e.g., if you cover up the I (circled in the figure) then you are left with V/R so that $I = V/R$ as shown in the figure.

Figure 2-4 Ohm's Law



The resistance, R , is related to the slope of the line on the I-V graph but beware; it is not itself the slope. Instead, we have

$$\text{Slope} = \frac{\text{Change in } I}{\text{Change in } V} = \frac{I}{V} = \frac{1}{R} \text{ so that } R = \frac{1}{\text{Slope}}$$

Thus a nearly horizontal line corresponds to a very HIGH resistance while a nearly vertical line corresponds to a very LOW resistance.

Example

The current through the resistor shown in Figure 1-6 changes by 1A when the driving voltage changes by 20V. The slope is

$$\text{Start} = \frac{\text{Change in } I}{\text{Change in } V} = \frac{1A}{20V} = 0.05 \frac{A}{V} \text{ so that } R = \frac{1}{0.05} = 20\Omega$$

Remember The voltage in Ohm's law is the voltage across the resistor; the difference between the voltages at the two ends.

Example

If one end of a 100Ω resistor is at 4.5V above ground and the other end is at 2.3V above ground then the voltage across the resistor is $4.5V - 2.3V = 2.2V$ so that the current through the resistor is $I = 2.2V/100\Omega = 22mA$.

2.1.6 Power in resistors

The voltage drop V across a resistor that is passing a current I causes the current to do work on the resistor so that power is dissipated. This power causes the resistor to heat up and we call the effect **Joule Heating**. Because Ohm's law relates the current and voltage in a resistor, we can calculate the power dissipation if we know only the voltage and the resistance or only the current and the resistance. If we apply a voltage V to a resistance R then a current $I = V/R$ flows so that the resistor dissipates power at the rate of

$$P = I \times V = I \times V/R = V^2/R$$

Similarly, if a current I flows through a resistor R then there must be a voltage $V = I \times R$ across the resistor and so it must dissipate power

$$P = I \times V = I \times I \times R = I^2 R$$

Example

A loudspeaker has a resistance of 8Ω . Calculate the voltage that must be applied to make it dissipate 10 Watts of power.

Since we know the power and resistance, we can say

$$10W = V^2/8\Omega$$

so that

$$V^2 = 10W \times 8\Omega = 80 \text{Volts}^2$$

which means that

$$V = \sqrt{80} = 8.9V$$

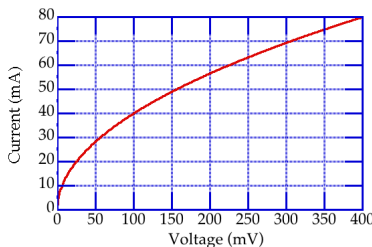


Figure 2-5 Non-linear Resistance I-V

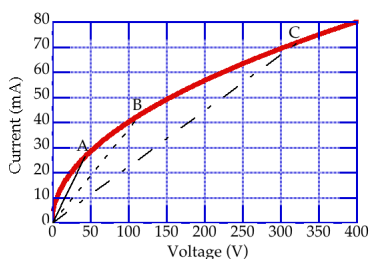


Figure 2-6 Chord Resistance

2.1.7 Non-linear resistances

There are some components that have I-V curves that are not straight lines. These are called **non-linear** components and they require that we extend our idea of resistance to take account of these more general cases. Look at the I-V graph of Figure 2-5. This is the I-V curve for a small light bulb. Obviously, no single value describes this resistance, but it can still make sense to talk of resistance at different applied voltages, if we are careful about how we do it. There are two different ways to define the resistance of a non-linear component. Both are useful in different ways.

The simpler of the two, sometimes called the **chord resistance**, is useful if you want to find the total current that will flow when a given voltage is applied. To find the chord resistance, we just apply Ohm's law directly to the straight lines as in Figure 2-6.

The chord resistance is not used very often; usually you would just use the curve itself to find the current at a given voltage.

The more useful resistance is called the **slope resistance** and is important when you want to know how small changes in the applied voltage will affect the current. First, look at what happens when we make small changes in the voltage across a linear resistor. We know that

$$V=I \times R$$

so if we increase V by an amount ΔV then a new current flows. We shall write the new current as the sum of the old current I and some small change in I , ΔI , produced by that change. Then Ohm's law says

$$V+\Delta V=(I+\Delta I) \times R$$

If we subtract the original V from both sides we are left with

$$\Delta V=\Delta I \times R$$

So, for a linear resistance, we get exactly the same resistance value if we write Ohm's law in terms of the changes as when we write it in terms of the voltage and current. We use this equation as the definition of the **slope resistance**, often called r rather than R .

$$r=\Delta V / \Delta I$$

Now we can apply this to our non-linear resistance. The slope resistance is

$$r = \frac{1}{\text{Slope of tangent line at point of interest}}$$

By drawing tangent lines carefully, we can measure the values of the resistance at various points as shown in Figure 2-7.

Example

At point A in Slope Resistance, the solid tangent line runs from $V=0V$, $I=12mA$ to $V=225mV$, $I=80mA$ so we have

$$\Delta V=0.225V-0V=0.225V \text{ and } \Delta I=0.08A-0.012A=0.068A$$

so the slope resistance is

$$r=\Delta V / \Delta I=0.255 / 0.068=3.3\Omega.$$

At point C the dashed tangent line runs from $V=0V$, $I=32mA$ to $V=0.38V$, $I=80mA$ giving us a slope resistance

$$r=0.38 / (0.08-0.032)=0.38 / 0.048=7.9\Omega$$

2.2 The physics of current flow

Everything in nature is made up of atoms and the atoms themselves are made up of still smaller particles. At the heart of every atom (Figure 2-8) is an incredibly tiny ball of mass and charge called the nucleus. Inside the nucleus there are two kinds of particle, some of them have no charge and are called neutrons, some carry a positive charge and are called protons. Flying round the nucleus is a cloud of negatively charged electrons, one electron for each proton in the nucleus.

Every charged particle carries round with it an electric field, an intangible thing that affects any other charge that comes into it. The field round a positively charged particle acts on other positive charges to push them away but it acts on negative charges to pull them in towards the particle. The field around a negatively charged particle acts in reverse, attracting positive charges and repelling negative ones. The forces that these fields exert are enormously powerful, strong enough to hold the electrons in orbit around their atomic nuclei and to glue atoms together into molecules. For example, the electric force between two protons is 10^{36} times as large as the gravitational force between them.

Electrostatic forces would dominate the universe if there were only one kind of charge. Instead, the positive charge on the proton is exactly the same size as the negative charge on the electron; a complete atom has no net charge. The positive and negative charges exactly cancel and there is no electric field to affect the world outside the atom. Since almost all matter is electrically neutral, the electric forces that play such a large role in the atomic world are not important for objects as large as us.

Remember We need the $1/\text{slope}$ in order to get units of Volts / Amps=Ohms.

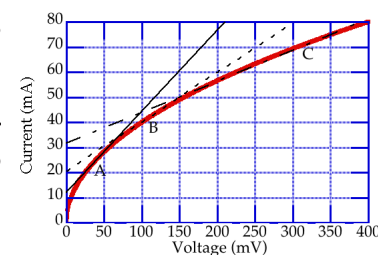


Figure 2-7 Slope Resistance

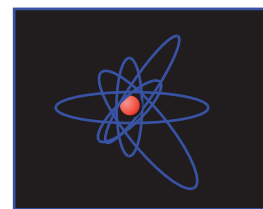


Figure 2-8 Semi-classical Atom

Note Any picture like this can only be a poor approximation of reality. First, the nucleus is so small compared to the total size of the atom (only about 10 millionths of the radius of the atom) that it could not be seen on the same scale. Second, quantum mechanics says that you can't ever really know where the electrons are so drawing well-defined orbits is misleading. Since we can't draw anything much closer to the reality we stick to images that give our minds a view that has proved useful over the years and don't try to interpret them too literally.

Info The existence of two kinds of charge that cancel each other out makes the electric field very different from the gravitational field. Put several charges together and on average you have a weaker electric field because on average half of all charges are negative and half positive. Put several masses together and you always have a stronger gravitational field because there is only one sign of mass.

2.2.1 Solids

At room temperature the atoms of most elements stick together with electric forces to make rigid arrays of atoms, solids. There are many kinds of solid. Some kinds are made up of regular arrays of one, or a few, kinds of atoms. We shall mostly be concerned with these crystalline solids. Other kinds are made up in a much more complex fashion. First the atoms join together into large assemblies called molecules and then the molecules form arrays that make up the solids. A substance such as wood has an enormously complex structure in which arrays of molecules make up fibers and then these fibers are held together by other kinds of molecules. Since complex molecular solids are almost always non-conductors we shall not be concerned with them.

A crystalline solid is built from regularly repeating, rigid blocks of atoms. In an elemental solid such as pure copper all the atoms are the same. In a compound solid, such as table salt, there may be several kinds of atom but they come in groups and all the groups are identical Figure 2-9.

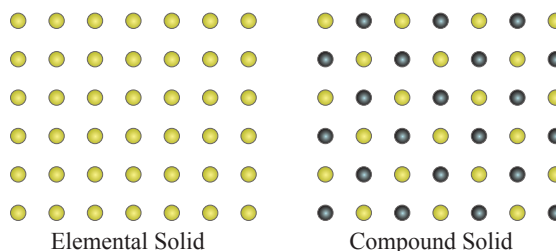


Figure 2-9 Atomic Structure of Solids

Inside most compound solids the atoms are very, very tightly attached to their electrons. You can apply a voltage to the solid and the electrons stay put; no current flows. They are insulators. Inside most elemental solids, in particular inside the metals, a few of the electrons are much less tightly bound. These electrons are called the conduction electrons. There are usually one or two of them to each metal atom and they can be pulled away from their parent atoms by an electric field.

If we put a metal crystal in an electric field then the electrons in the metal will move around. We get an idea of how this happens in the following figures.

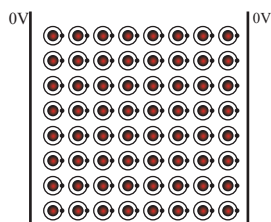


Figure 2-10 Atom in Zero Field

Figure 2-10 shows the metal crystal before we apply an external field. Each atom is shown as a little blob with a circle and a dot round it. The blob is the atomic core—the nucleus and those electrons that are tightly bound to it. The circle is the orbit of the single loosely bound conduction electron and the dot is that electron.

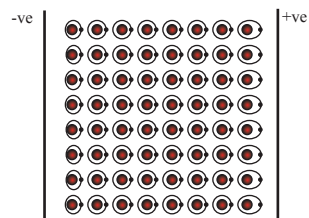


Figure 2-11 Atom in External Field 1

In Figure 2-11, the crystal has been placed in the external electric field. The field has distorted the electronic orbits. This is shown schematically in the picture by the expansion of the orbits on the positive side of the field and the contraction on the negative side. The field is pulling the electrons towards it.

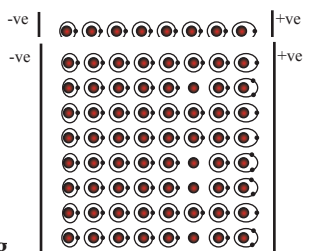


Fig 2-13 Atom in External Field 3

In Figure 2-12 the electric field is stronger. Now the distortion is large enough that a few of the electrons in the second column jump into the oversize orbits in the last column. This leads to a few extra electrons in the far right column and leaves behind some atoms without outer electrons in the second column. These atoms have orbits for electrons but no electrons to put in them. Such an orbit without an electron is called a hole. Note that the atoms with holes have a positive charge so that the hole behaves a little like a positive particle.

As Figure 2-13 shows, the previous situation is unstable. The second column holes pull on the electrons on either side of them. Those to the right are held by the external field but those to the

left are not and so electrons have jumped from the third row into the second. Now the second row is happy but the third row has atoms with holes.

Obviously, the holes are no more stable in the third column than they were in the second. Thus, electrons move from the fourth row to fill the holes, then electrons move from the fifth row into the fourth and so on. As the electrons move from left to right, the holes appear to move from right-to-left until they end up in the left most column. At the end, there are extra electrons at the right edge, which has a negative charge, and some orbits with holes at the left-hand edge, which has a positive charge (Figure 2-14). The whole process takes only a matter of nanoseconds.

We have seen how charges can move around inside a metal as electrons jump from atom to atom and holes move in the other direction. So far, this has been a transitory process with charges building up on the surfaces of the solid and current flowing for a very brief time. If we connect wires (other pieces of metal in which the same processes are going on) to the crystal then we can bleed off the extra electrons on from the right and supply electrons from the left so that a steady current flows. That current is carried in the same way, with electrons hopping through the metal lattice leaving holes into which other electrons can jump.

2.2.2 Resistance

In a perfect crystal the electrons can hop from atom to atom without any loss of energy; current flows with no voltage loss. This phenomenon is called superconductivity and it is observed in a number of materials at very low temperatures. Most real crystals at room temperature contain various kinds of imperfection. There are impurities, atoms of the wrong kind, and atoms out of place in various ways as we see in Figure 2-15.

Even if there are no impurities or displaced atoms, the individual atoms are rarely in exactly the right places because they jiggle around. When an electron tries to jump to an out-of-place atom or to an incorrect atom it has problems. The effect is that the electron loses a little bit of energy. That means that the outside world has supply energy to keep the current flowing; there has to be a voltage to pump the current through the material. The more electrons are trying to get through the crystal at any time, the more energy it takes to push it through. This is the origin of Ohm's law. Even more, the higher the temperature is, the more the atoms jiggle around. That means that the atoms are more out of place at high temperatures and the resistance is higher.

2.2.3 Semiconductors

Towards the right hand side of the periodic table there are a few elements that sit on the border between the conductive metals and the insulating non-conductors, elements such as selenium, germanium, and, most important, silicon. The outer electrons in these solids are more tightly bound to their atoms than those in the metals but not quite so tightly as those in non-metals. These materials are called semiconductors. Very pure semiconductors, called intrinsic semiconductors, are insulators under normal conditions. However, if small amounts of impurities are added to the crystal, it can be made to conduct. By controlling the impurities, we can make semiconductor devices with a wide range of fascinating properties. These semiconductor devices, such as diodes, transistors, and integrated circuits, lie at the heart of modern electronics. They are the amplifiers and switches that make all our electronic gadgets work.

Let us examine the conduction mechanism in silicon, by far the most important of the semiconductors. Each atom of silicon has four electrons in its outer orbit but they are too tightly bound to move around in the crystal. If we can knock a few electrons out of their places then they can move around in the silicon. It takes quite a lot of energy to knock an electron out of its orbit and the only source of energy is the thermal motion of the atoms. At room temperature the thermal energy available is so low that there are very few of these conduction electrons around. Indeed the electron density in a pure semiconductor is about 10^{-12} times the density in a conductor. Thus, although these electrons are free to move, they can carry very little current

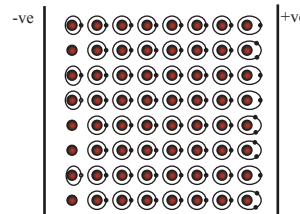


Figure 2-14 Atom in External Field 4

Note This description of what the electrons do is only a model. Quantum mechanics tells us that we should not really think of individual electrons but should describe how waves of electron probability act within the solid. However, most of us have minds that don't work that way and so we use models to gain insight into the behavior of the electrons. So long as we don't take the details of the models too literally, these models can give us a useful working insight into the way the atomic world operates.

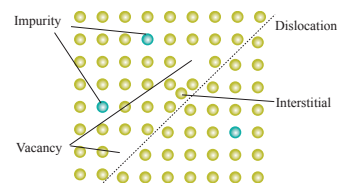


Figure 2-15 Imperfections

and the material is almost an insulator. As you warm the crystal up, there is more energy available to knock electrons out but even at 100°C there are too few electrons around to carry any significant current. So we have a material with almost no moveable electrons (or holes) but one in which electrons (and thus holes) can move around quite easily.

If we apply an electric field to a piece of semiconductor, that is we put a voltage across it, then there is an extra source of energy. If the electric field is strong enough then it can pull electrons out of their orbits and drive them through the crystal. However, the fields needed to do this are very large compared to the usual fields that we encounter in most devices. Only in rather special circumstances do you find fields that are strong enough to pull electrons out of their orbits and make them conducting electrons. However, those circumstances do occur in some semiconductor devices as we shall see in chapter 12.

If we add a very tiny amount of an element that is similar in size to silicon, then the new atoms can take the place of a very few of the silicon atoms without altering the structure of the crystal. If we add in an element that has five electrons in its outer orbit, such as phosphorus, then we get a crystal with a few sites that have too many electrons but are otherwise very similar to their neighbors. Something very interesting happens. The extra electrons are very loosely bound to their parent atoms and can hop around. They can't hop from impurity to impurity because the impurities are much too far apart, but the nearby silicon atoms are sufficiently similar that they can accept an extra electron for a little while and then pass it on. Thus the impure semiconductor can conduct by passing the extra electrons along through the crystal. This is called an n-type semiconductor because it has extra negative charges that carry the current. We say that the semiconductor has been doped with the impurity and call the impurity a donor impurity because it donates extra electrons to the bulk crystal.

If we use as our impurity an element such as aluminum, with only three electrons in its outer orbit, then we get impurity sites that have one electron too few. Each impurity atom has a hole where the fourth electron should be. It is very easy for an electron from a nearby silicon atom to fall into the impurity site and so make the hole jump from the impurity onto the silicon. Again, we have introduced some free charges that can carry current around, if we apply an electric field to move them. We call such a material a p-type semiconductor because it conducts with positive holes. We call the impurity an acceptor because it accepts electrons from the silicon crystal to leave behind mobile holes.

Note that there is one very important difference between conduction in a metal and in a doped semiconductor. In a metal there are potentially at least as many charge carriers as there are atoms so there is no practical limit to the current that can flow. In a semiconductor there are only as many charge carriers as impurities. Since only about 1 in 10^6 atoms is an impurity there is a much lower limit to the maximum current that can flow. It is quite possible to saturate a semiconductor, that is to reach a point at which the current cannot increase no matter how high you raise the voltage because you have run out of charge carriers.

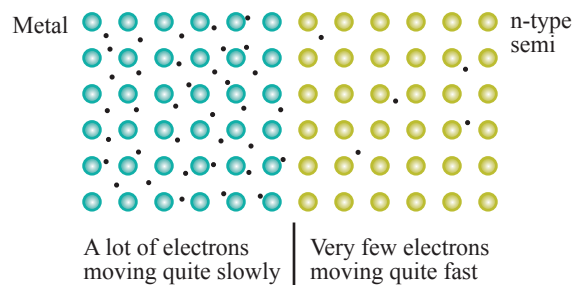


Figure 2-16 Metal-n Type Semiconductor Junction

If we connect a metal wire to a piece of n-doped semiconductor (Figure 2-16) then we can get electrons to cross the junction very easily. The metal has floods of conduction electrons available and can pass them around quite freely. At a junction between an n-type semiconductor and a metal these electrons pass freely across the junction and current flows from metal to semiconductor or vice versa. In the semiconductor, the current is carried by a very small

number of electrons moving quite rapidly. In the metal, there are enormously more electrons around and so they move along much more slowly to carry the same current.

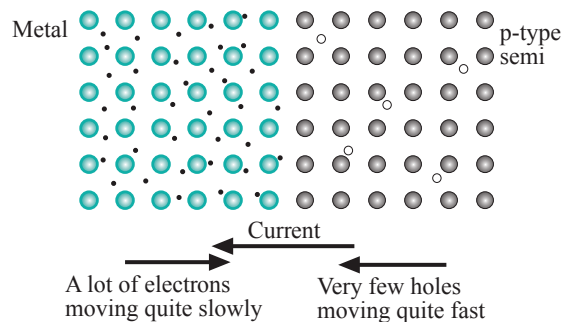


Figure 2-17 Metal-p Type Semiconductor Junction 1

At a junction between a p-type semiconductor and a metal, things are a little more interesting. If the current is flowing from the semiconductor to the metal then the holes are flowing towards the metal (Figure 2-17). At the junction, each hole eats up an electron from the metal—that is an electron from the metal falls into the hole. This means that more electrons must flow from the metal toward the junction. So the electrons and holes both flow towards the junction, where they destroy each other.

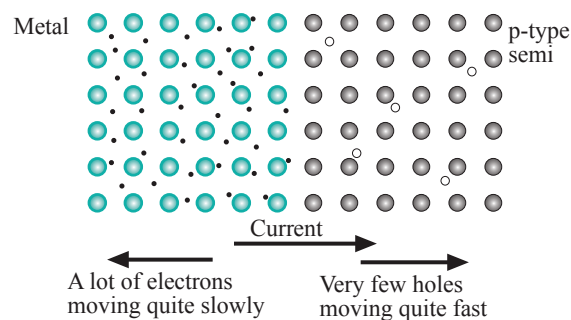


Figure 2-18 Metal-p Type Semiconduction Junction 2

If the current flows from the metal to the semiconductor then the electrons in the metal are flowing away from the junctions (Figure 2-18). That means that electrons are being sucked out of the semiconductor, creating holes. The holes migrate through the semiconductor carrying the current with them. Thus the electrons and holes both migrate away from the junction, the holes widely spaced and moving quickly, the electrons tightly packed and moving slowly.

Summary

Electronics deals with movement of **charge**, Q , measured in **Coulombs**. When charge moves it forms a **current**, I , measured in **Amperes**.

$$I=Q/t \text{ and } 1 \text{ Ampere} = 1 \text{ Coulomb per second}$$

Current is driven by electrical pressure or **potential**, V , measured in Volts. When a charge Q moves through a potential difference V it does work, W , measured in **Joules**.

$$W=Q \times V \text{ and } 1 \text{ Joule} = 1 \text{ Coulomb} \times 1 \text{ Volt}$$

When a current I flows through a potential difference V , the work it does emerges, usually as heat. The rate at which work is done, the **power**, P , is measured in **Watts**.

$$P=I \times V \quad 1 \text{ Watt} = 1 \text{ Joule/sec} = 1 \text{ Amp} \times 1 \text{ Volt}$$

For most conducting materials the voltage, V , needed to drive a current, I , through the material is proportional to the current. We call the constant of proportionality **resistance**, R , measured in **Ohms**.

$$\text{Ohms Law } V=I \times R \quad \text{and } 1 \text{ Volt} = 1 \text{ Ampere} \times 1 \text{ Ohm (symbol } \Omega \text{)}$$

If the current is not proportional to the voltage then we can still define the slope resistance or incremental resistance

$$r = \Delta V / \Delta I$$

When current flows in a resistor of value R it dissipates **Power** P at a rate given by

$$P = I^2 \times R \text{ or } P = V^2 / R$$

In a conventional plot of I vs. V the slope of a line is the inverse of resistance so that the slope resistance r is given by

$$r = \frac{1}{\text{Slope of tangent line at point of interest}}$$

Exercises

1. As a person walks across a room, her rubber-soled shoes pull electric charges off the nylon carpet and build up a charge of $1 \mu\text{C}$ on her body. When she touches the light switch there is a small spark and the current flows to ground. If the spark lasts 0.1 mS what current must it contain?
 2. A small light bulb draws 10 mA from a 3 V battery when it is lit. What is its resistance?
 3. How much current does a 100 Watt light bulb draw from a 115 V power source? (Note: Real household light bulbs operate from household current that actually changes its value continually, as we shall see in chapter 6, but the answer turns out to be same.)
 4. What resistance must the lamp in problem 3 have to draw that current?
 5. A typical person with fairly dry hands has a resistance between one hand and the other of about $100,000 \Omega$ to $1 \text{ M}\Omega$. If such a person can feel a tingling sensation when a current of 0.1 mA flows, what is the highest voltage of battery that they could handle without noticing any such effect?
 6. At what rate does the light bulb in question 2 dissipate power?
 7. A car stereo must operate off the 12 V power produced by the car's alternator. What is the largest amount of power that you can get out of an 8Ω loudspeaker in such a case?
 8. Given the difficulty of hearing the sound over the noise of the engine, would you expect a car stereo to be more likely to use 8Ω loudspeakers or 4Ω loudspeakers? Why?
1. Find the chord resistance at point C on the graph of Figure 2-6.
 2. Find the slope resistance at point B on the graph of Figure 2-7.

Chapter 3: Simple Components

Every electrical circuit is made up from smaller pieces that we call components. Some of these are very simple—for example, pieces of wire—and some are quite complex, for example power supplies. When we are studying a circuit, we usually choose to ignore all the details of what goes into a component. After all, even a wire is quite complicated if we start to think about exactly what it means for charge to flow inside a piece of metal, while a power supply may have tens of physical pieces inside it. Therefore we pick a convenient level of detail and try to understand the circuit at that level. For example, we treat a power supply as a source of voltage with particular characteristics regardless of what is inside it, even though, at another time, we might be very interested in studying the details of the power supply's construction. Similarly, we usually think of a transistor as a component with certain electrical characteristics and do not worry about the details of charge transport inside, although at some point in our studies we must look at the transistor in much more detail.

3.1 Ideas of components

We can think of the generalized component as a black box with terminals that can be attached to other components and whose behavior we understand in terms of the effects of applying voltages to those terminals. Then we characterize each component by the number of terminals that it has. Most simple components have two terminals, a number of interesting ones have three, while most of the interesting large scale components have four terminals arranged as an input pair and an output pair.



Figure 3-1 n-terminal devices

The terminals are arranged in pairs because current has to flow in complete circuits so that the input current has to flow both into and out-of the device. Similarly, the output current has to flow both into and out-of the device. The **three**-terminal device is often thought of as a special case of the four-terminal device that has the bottom pair of terminals fused into a single terminal. This means that if we understand four-terminal devices then we understand three-terminal devices as well!

3.2 Ground

I can think of only one practical 1-terminal device, **ground**, something that has no simple physical existence. It is just a single terminal with a symbol on it that is used to mark the spot in a circuit that is designated as zero volts. In many cases, this point is physically connected to earth ground through the power supply. This is a point that defines zero volts and is assumed to be able to sink or source any amount of current without its voltage changing.

The symbol in Figure 3-2 with three lines is the usual symbol but some European circuits use the other symbol. The other symbol, with the little down-facing triangle, is also used in some precision circuits where it is important that wires that are connected to ground and that carry significant currents be kept separate from other wires that are also connected to ground but are sensitive to tiny variations in voltage. In such a case all the points marked with the triangle symbol could be connected together but they and points marked with the three lines would be joined at exactly one point. This helps minimise noise in sensitive circuits..

Any two points that are connected to ground are also connected to each other. Ground acts like a piece of wire connecting all grounded points. This is particularly important when using wall-powered equipment such as oscilloscopes and signal generators. They are usually

More complicated configurations are possible, as we shall see later. However, these are sufficient for our needs for quite a while.

Ground

The earth itself is a fairly good conductor of electricity. The power companies make use of this and all power company power stations are electrically connected to the earth through large metal rods driven into the ground.

Large metal rods are also driven into the ground where the power reaches a house and the electrical system of the house is connected to the rod. The conductor connected to ground is the big round pin in a 3-pin plug. All devices that use 3-pin plugs are inherently connected to that common ground point! This is important when connecting two such devices together as the grounds **MUST** be connected to each other and not to any other point since they are already connected at the other end.

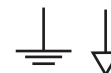


Figure 3-2 Ground Symbols

grounded, that is they have one side of their input or output connected to ground through the wall connection. Thus we must be sure that we connect the grounded sides of all leads going to such an instrument to the same point in the circuit! Otherwise we can create an inadvertent short circuit—connecting a point which wants to be at some voltage to ground and possibly damaging the device.

Note In some high precision circuits it may not be possible to treat more than one point in the circuit as 0V because of real currents flowing in real wires. Great care is sometimes needed to avoid problems caused by incorrect grounds

3.3 Wires

The simplest, and most pervasive, 2-terminal component is a wire. The ideal wire, which we represent on circuit diagrams by a simple black line, is a component that can pass any current without a voltage difference appearing between its ends; it has zero resistance. Thus any two points connected by a wire are always at the same potential. This means that the shape of a circuit is irrelevant, only the way that it is connected affects its behavior.

Remember Any two points connected by a wire are at the same voltage!

Real wires do not meet this ideal and can carry only a limited amount of current before suffering damage and they do have a small voltage difference between the ends because their resistance is only very small (usually thousandths of an Ohm or less). However these imperfections can be neglected in all but the most precise circuits or those that carry very large currents.

Most wires used in electronics are made of copper, because it has the lowest resistivity of any material except silver. They are often coated with a thin layer of tin, which provides corrosion resistance, since bare copper oxidizes quite quickly in the air. While most wires use a single piece of copper, wires such as microphone cords or lamp cords that must be flexible are often made from many strands of thin wire. No matter whether the wire is solid or stranded, the total thickness of the wire determines its resistance and its current carrying capacity. A thick wire has a much lower resistance per foot than a thin wire does and so can carry a larger current without heating up.

Simple components such as resistors usually come with bare wire leads already attached so that the component can be soldered into a circuit, although there is an increasing trend towards leadless components that are soldered directly to the board. Such bare wires are not used to connect different circuit boards together because the wires could touch and make connections where none should be. Instead, interconnecting wires are usually covered, or **jacketed**, in an insulator such as rubber or plastic. The best interconnection wires are jacketed in Teflon, which is both physically strong and very heat resistant, so that an incautious soldering iron will not melt holes in the insulation.

If several wires have to go more than a few inches together then it is often a good idea to use a **cable**, which collects several insulated wires together and puts them inside an outer insulating sheath (Figure 3-4). If this cable has to go more than a few feet or has to pass through an area that is full of electrical noise then a conducting **shield** is often put round the wires inside the sheath. This shield is connected to ground at one end of the cable and prevents the wires inside from picking up the electrical interference. This is most often used for low-level signals such as microphone signals.

One common form of low noise cable is the **coaxial** cable (Figure 3-5). Here the two wires carrying a signal and its ground are arranged one inside the other as a pair of coaxial cylinders separated by insulation. The outer conductor is often made of hundreds of thin wires braided together for flexibility. It is called the **shield** and is connected to ground. It protects the signal carrying wire from interference. Coaxial cables are used for very low level signals and are especially common in radio frequency circuits.

Note Table 3-1 shows some common wire gauges with their diameters and resistances in Ohms/1000 feet

Table 3-1: Some Common Wires

AWG	Diam (mm)	$\Omega/1000\text{ft.}$
10	2.5	1.0
12	2.1	1.6
14	1.6	2.5
16	1.3	4.0
18	1.0	6.4
20	0.8	10
22	0.6	16
24	0.5	25
26	0.4	40
28	0.3	65

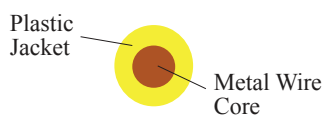


Figure 3-3 Insulated Wire

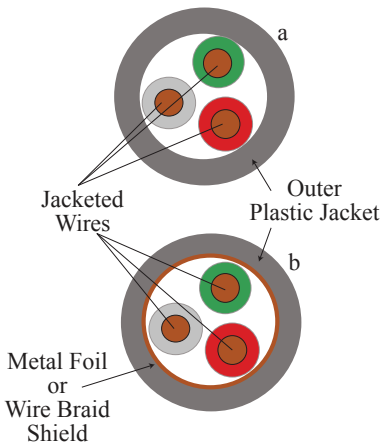


Figure 3-4 a) Unshielded and (b) Shielded Cables

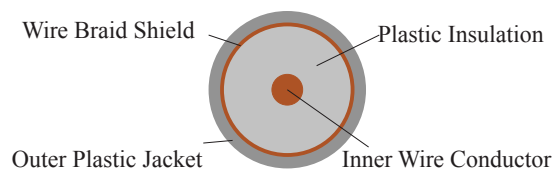


Figure 3-5 Coaxial Cable

3.3.1 Fuses

A fuse is a special piece of wire that is designed to self-destruct if too much current passes through it. Fuses protect more expensive pieces of equipment (a fuse only costs a few ¢) by burning out before the excess current can damage the equipment. Once the fuse has burnt (or **blown**) out it must be replaced before the equipment will function once more. Of course, you should make sure that the problem which caused the fuse to blow is corrected before you replace the fuse. Otherwise you will not only waste another fuse but also risk further damaging the equipment.

Fuses come in two kinds, fast acting and slow blow. Fast acting fuses are just thin wires in a convenient package. When too much current flows, the wire gets so hot that it vaporizes very quickly. Slow blow fuses are a little more complicated and wait a second or so before vaporizing. Quite a lot of pieces of equipment take more current for a very short period while they are starting up, usually to charge a capacitor or to get a motor started. A fast blow fuse that would protect the equipment when it was running would blow during turn on so we use slow blow fuses.

Fuses almost always come in little glass cylinders with metal ends (Figure 3-6). These fuse cartridges are designed to make it easy to replace a fuse, which is a delicate piece of wire. The cartridge is usually marked, either on the glass or on the metal caps, with the rating of the fuse. Each fuse is rated with the maximum current that it can pass without blowing up and with the maximum voltage at which it is designed to operate. Almost all common fuses are rated at 250V.

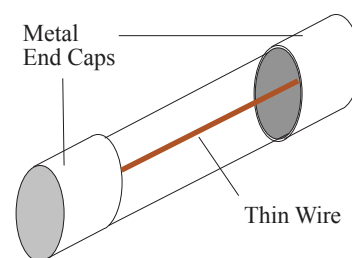


Figure 3-6 Typical Fuse

3.4 Switches

A switch is a component that controls the flow of current. It corresponds to a valve in our plumbing analogy. The simplest kind of switch has two positions, one that lets current flow and one that blocks it. Such a switch is just like an on-off valve or stopcock in a plumbing system. It works by either keeping two pieces of wire apart so that no current flows or by bringing them together to allow current to flow. Figure 3-7 shows the symbol for such a switch. This kind of switch is called a **Single-Pole, Single-Throw (SPST)** switch. This is a notation that tells us that the switch has a single moving element (a single **pole**) and that moving element has one contact position (a single **throw**). Such switches come in a variety of physical forms including toggle, push-buttons, and slide switches.



Figure 3-7 SPST Switch

More complicated kinds of switch exist that can send current into one of several different paths. These are called multiple throw switches and are used in such places as the input selector of a stereo system. For example, a quad-throw or 4-way switch (Figure 3-8) can connect the input of the stereo to one of 4 different sources such as Tape, FM, CD, and AUX. One common form of this switch works by sliding a piece of metal along, usually with a plastic fitting, so that it connects only one of the inputs to the output (or vice-versa) at any time another common form is the rotary selection switch..



Figure 3-8 4-way switch

Even more complicated switches are made by mechanically coupling several of these basic kinds of switch to give a switch that can switch current in several circuits or paths at once. Each moving element is called a pole so we get switches such as the one in Figure 3-9, which is called a double-pole double-throw (DPDT) switch. It connects each of two inputs to one of two outputs so that if circuit 1 is connected to output A1 then circuit 2 is connected to output A2. Note that the dotted lines in these symbols stand for a mechanical connection. They are

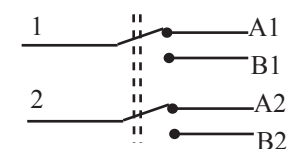


Figure 3-9 DPDT Switch

not part of the circuit and they do not pass current. DPDT switches are usually toggle switches but larger multi-pole multi-throw versions are normally rotary switches.

3.5 Power Supplies

All circuits need to get power from somewhere. Ultimately most of our electrical power comes from large generating stations run by power companies. They use large machines that rotate coils in a magnetic field to turn mechanical energy from waterfalls, wind turbines, or steam turbines into electrical energy. They then distribute that energy over the power lines that you see all around, from the great pylons that march over long distances to the wires hung from telephone poles that reach most houses. We normally access that power through wall sockets that provide a source of 115V RMS alternating current (see Chapter 6).

3.5.1 Power Supplies

Straight wall current is not suitable for most electronic devices and must be converted into a more useful form by a **power supply**, which takes power from a wall socket and change the voltage and current to values appropriate for a particular piece of equipment (see Chapter 10). Such power supplies are almost always intended as **constant voltage sources**. An ideal constant voltage source would supply power at a constant voltage regardless of the amount of current that is drawn. Real supplies approximate this more or less well. In practice there are always limits to the amount of current that a particular supply can draw and there is often some reduction of voltage as more current is drawn.

A power supply is basically a box that plugs into the wall and that has some **terminals** to connect wires to. Sometimes it has a control to set the output voltage but most power supplies produce only fixed voltages. A power supply is quite complex internally (see later) but is given a simple symbol to disguise the complexity and show only its function (Figure 3-10). The operating voltage is often shown inside the circle although it can equally well be shown beside it.

If the supply is one that can be varied then the voltage range is usually shown. For example, Figure 3-11 shows a 0-to-15V power supply (a fairly common item).

In both of these examples I have marked the positive terminal with a plus sign. It is quite common to omit the plus sign when the voltage source has its connections emerging from the top and bottom. In such cases the top lead is always assumed to be the positive terminal.



Figure 3-10 5V Power Supply

Info Strictly, the circle symbol is a **Voltage Source**, an ideal component that will supply whatever current is necessary to hold the voltage across its terminals a given voltage. A real power supply only approximates this ideal and has limitations on the maximum current that it can supply without suffering damage or without the voltage changing.



Figure 3-11 0-15V Variable Supply

Real vs. Ideal

The power supply symbols that we draw represent ideal devices. An ideal voltage source will hold the voltage between its terminals fixed regardless of how much current flows; the ideal current source will force the current to flow no matter how much voltage it takes. Real power supplies have real limitations. Real voltage sources can hold their outputs fairly constant up to some rated maximum current and above that the voltage will fall, possibly drastically. Similarly, real current sources have limits on the maximum voltage that they can apply to force current through an external circuit. The people who design and build systems are responsible for making sure that the real power supplies that they use to implement the ideals are capable of working under the conditions in the real circuit.

Power Supplies and Ground

Even though we don't show it on the symbol, real power supplies are at least 4-terminal devices. Two terminals bring power in from the wall socket and the two that we show deliver it to the circuit. Many simple wall-wart power supplies follow this model exactly. They have 2-pin wall plugs and a 2-pin output connector. With only 2-pin wall plugs, they are not connected to the ground of the house wiring and so the output is completely isolated from ground. This would create safety concerns if they were limited to low voltages and currents.

The slightly more sophisticated power supplies that we usually find in electronic instruments usually do use 3-pin plugs and so do connect to house ground. This means that one of the output terminals is connected to ground, the negative terminal for a positive power supply and the positive for a negative power supply. Because one terminal is connected to ground we have to be very careful when connecting them to circuits and instruments that also connect to ground, instruments such as oscilloscopes and signal generators.

Still more sophisticated supplies have three output terminals, one that is ground and then two that are isolated from ground. The two isolated terminals are the actual power supply terminals. You can use the supply as either a positive or negative supply depending on how you connect ground or you can omit ground altogether and **float** the supply. In that case neither terminal is at ground and could be at any voltage. All the power supply does is keep the positive terminal at a fixed voltage above the negative terminal.

3.5.2 Batteries

Another common source of electrical power is a **battery**, which converts chemical energy to electrical energy. Batteries are also more or less good constant voltage sources. They do well at small currents but most are limited to currents of a few hundred milliAmps, though car batteries can supply several hundred amps for a few seconds.

A battery is made up from one or more cells each of which supplies a constant fixed voltage. The voltage of a single cell depends only on the chemistry of the materials that make it up. Common types include the 1.5V carbon or alkali-manganese cell, the 2V lead acid cell (think car battery), and the 3V Lithium cells found in many cameras and computers. The symbol for a single cell battery is shown in Figure 3-12. The longer crosspiece marks the positive terminal and the battery voltage is usually indicated in text somewhere nearby.

A multi-celled battery, such as a car battery or the 9V battery pack with several separate cells that you find in radios and cassette players, has a more complex symbol made by stacking the simple symbols (Figure 3-10).

3.5.3 Constant Current Supplies

There is a second kind of power supply that is much less common than the voltage source and that is the Constant Current Source (Figure 3-14). As its name implies, this is a device that will force a constant, preset, current to flow between its terminals. That is, the source will set the voltage between those terminals to whatever value is necessary to make the desired current flow. The symbol for a constant current source has two circles to distinguish it from the voltage source.

As with the voltage source, we mark the positive terminal with a plus sign and write the value of the current somewhere nearby. Similarly, it is quite common to omit the plus sign when the symbol is shown vertically. In such a case the top terminal is always the positive terminal.

3.6 Resistors

A resistor is a 2-terminal component designed to have a constant, well-determined resistance. Resistors are the most common components and they are described by three characteristics, the resistance value, the precision, and the amount of power that the resistor can dissipate without blowing up. The most common resistors are quite small, somewhat inaccurate, and can dissipate only $\frac{1}{4}W$ of power but resistors are made with much larger powers or higher precisions for special situations.

There are two standard symbols used to denote a resistor, the first (Figure 3-15a) is the most common symbol in the US and the second (Figure 3-15b) is mostly used in Europe, but there is quite a lot of mixing these days.

In circuit diagrams we always specify the resistance value for each resistor in the circuit. We only mark the precision or power if there is something unusual about the resistor. Otherwise we assume that all resistors can be $\frac{1}{4}W$ 5% carbon film types. We give the value either using the obvious decimal point notation such as 4.7k for a 4700Ω resistor or using a compact notation that puts the range letter in place of the decimal point. In this notation a 4700Ω resistor would be labelled 4k7, a $2,400,000\Omega$ resistor labelled 2M4, and a 3.3Ω resistor labelled 3R3. This notation is also sometimes used on the bodies of larger resistors themselves. However, most small resistors are instead marked using a system of color coded bands (see the box at the top of the next page).



Figure 3-12 Single Cell Battery



Figure 3-13 4-Cell Battery

Note You don't normally see the individual cells of which a battery is constructed. However, if you carefully open the metal case of a 9V battery you will find inside 6 small, oblong, black packages. These are the individual 1.5V cells making up the battery. These contain mildly caustic chemicals and should be treated with care. Wearing gloves is a good idea and it is important to wash carefully after handling the cells.



Figure 3-14 Constant Current Source

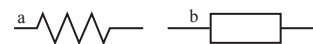


Figure 3-15 Resistor Symbols

Note The precision of a resistor is a measure of how much faith you can put in the value marked on the resistor. Resistors are made by the thousand in machines and there is some variability in the process so that two nominally identical resistors will not have exactly the same resistance. The more careful the process, the closer the values will be and the more expensive the resistors will be. The precision expresses the range of real values as a percentage of the nominal (printed) value. Thus a 100Ω 5% resistor is guaranteed to have a value between 95Ω and 105Ω while a 100Ω 1% is guaranteed to have a value between 99Ω and 101Ω.

The Resistor Color Code

Standard 5% and 10% resistors are normally marked with their values using a scheme of colored bands. The marking is made up of two parts. First there is a set of three colored bands giving the value of the resistor then there is a gap and at the other end of the resistor there is a single band (or occasionally a double band) giving the tolerance of the resistor—silver for 10%, gold for 5%, and pink for 1%. The value uses three bands, one for the first digit, one for the second digit, and one giving the number of zeros after the second digit. Each position uses the following code

Black	0	Brown	1	Red	2	Orange	3	Yellow	4
Green	5	Blue	6	Purple	7	Gray	8	White	9

So a resistor marked Red Purple Brown has a value 2, 7, 1 zero or 270. Note that this is not quite scientific notation, the last band is not the power of ten but the number of extra zeros. It is very convenient to remember that

- All resistors that end with a brown band have values in the hundreds of Ohms.
- All resistors that end with a red band have values in the thousands of Ohms.
- All resistors that end with an orange band have values in the ten thousands of Ohms.
- All resistors that end with a yellow band have values in the hundred thousands of Ohms.
- All resistors that end with a green band have values in the millions of Ohms.

This scheme dates from the time before machinery was available to print little tiny numbers on the resistors and is rarely used for higher precision resistors. They usually just have the value printed in numbers but they still use the weird convention that the last digit is the number of extra zeros. For example, a resistor marked 1001 is a 1000Ω 1% resistor.

3.6.1 Real resistors

The simplest resistor is just a piece of thin wire—the longer and thinner the wire the greater the resistance, just like a plumbing pipe. You can push a lot of water through 2m of 10cm diameter sewer pipe with very little pressure while it takes a lot of pressure to force the same amount of water through 20m of 1mm tubing. Such **wire-wound** resistors are usually wound on a ceramic core and then coated with enamel and have the value printed on the outside. Wire-wound resistors have quite high power ratings, 5W to 30W being common and large resistors above a hundred watts available.

Resistors are available in a range of precisions, from extremely cheap 10% resistors to fairly expensive 0.01% ones. The most common types of resistor are made from thin films of carbon or tin oxide deposited on a ceramic core and then cut into a spiral (see box on next page). These are low power components that are available in a range of precisions.

Carbon film resistors with a tolerance of $\pm 5\%$ dominate the industry. They are cheap, rugged, and available in a wide range of standard values, from small fractions of 1 Ohm up to about 10 MOhms. For circuits that need more reproducible resistor values there are 1% metal film resistors available with a similar range of values. There have to be many more different standard values of 1% resistors in order to make sure that every possible value is available.

Until recently, most resistors were small ceramic cylinders with metal wires coming out of the ends. The wires were passed through holes in a circuit board and soldered to connections on the back of the board. Many resistors now are made as small brick shaped components with little metal pads rather than the metal wires. They are designed to make connections on the side of the board on which the resistor. These are called **Surface Mount Resistors**.

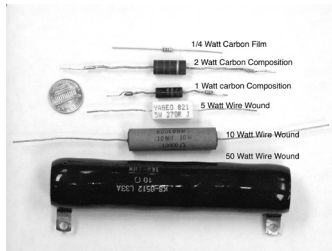


Figure 3-16 Some Real Resistors



Figure 3-17 Surface Mount Resistors

Note How Big?

The text above is set in 9pt Palatino. The smallest of these resistors is about 2x the size of the period at the bottom of its question mark. And yes, they are *very* hard to work with!

Standard Values of Resistors

The higher the precision of the resistor, the more digits it takes to specify the value and the closer you can put the standard values. For example, 10% resistors come in a range of values that puts 12 different values in each decade of resistance. So that between 10Ω and 100Ω there are 12 standard values, between 100Ω and 1000Ω there are 12 standard values, and so on. Here are the 12 different values (called the E12 series) each of which is given to only 2 digit

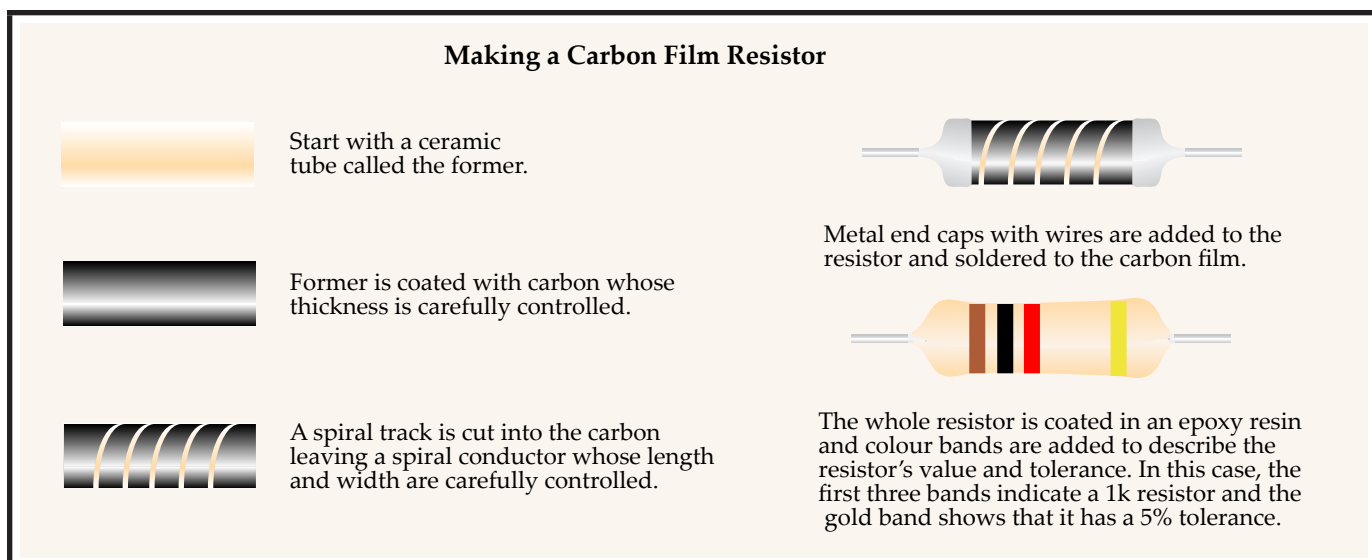
1.0 1.2 1.5 1.8 2.2 2.7 3.3 3.9 4.7 5.6 6.8 8.2

By contrast, 5% resistors can be spaced twice as closely so that there are 24 standard 5% resistor values per decade. These make up the E24 series of values shown here. Note that these are still 2 digit numbers.

1.0 1.2 1.5 1.8 2.2 2.7 3.3 3.9 4.7 5.6 6.8 8.2

1.1 1.3 1.6 2.0 2.4 3.0 3.6 4.3 5.1 6.2 7.5 9.1

Higher precision resistors such as the 1% and 0.1% metal film types, can be spaced yet more closely and so even more different values have to be made. Two digits no longer suffice to specify the values so precision resistors may need 3, 4 or even more digits.



3.6.2 Variable Resistors

In addition to fixed value resistors there are several different kinds of variable resistor that are in use. These are almost always arranged so that there are three connections, two fixed and one moving. This means that they behave like two resistors connected together (in series, see below) with the resistance values depending on the position of the moving connector.

The most common form uses a carbon track, just like a large carbon film resistor, with a moveable metal contact that is controlled by a knob. Some other forms replace the carbon track by a tightly wound metal wire along which the moveable contact slides but the principle is the same for all of them. This general construction form is reflected in the symbol shown in Figure 3-18.

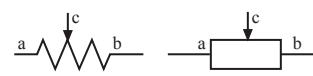


Figure 3-18 Variable Resistors

The resistance between points a and b is fixed and equal to the value of the variable resistor. For example, if the resistor is labeled 10k then that is the resistance between points a and b. Point c is the moveable contact that slides along the track so that the resistance between points a and c increases from about 0 to 10k as the slider moves from the bottom to the top of the track while the resistance between c and b decreases from 10k to about 0. When the component is used as a simple variable resistor, connections are made only to points a and c and b is either left unconnected or is connected to point c (which gives better noise behavior).

Note There is no problem leaving point b unconnected since a wire that is not connected at one end acts like a blocked pipe and no current flows in that part of the circuit.

One of the most common uses of variable resistor is to extract an adjustable fraction of some signal. This is how a volume control works. A variable resistor is connected between the signal source and the circuit that will accept the signal, as shown in Figure 3-19.

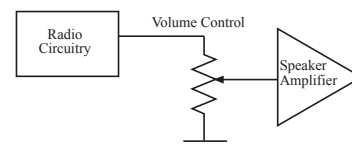


Figure 3-19 Variable Resistor Volume Control

As the volume control knob is turned, the slider (the moving contact) is moved from the bottom of the resistor towards the top. When the slider is near the bottom, it picks off only a tiny part of the complete signal and so the volume is very low. As the slider is moved up, the fraction of the resistance between the slider and ground increases. This makes the fraction of the whole signal that reaches the loudspeaker amplifier also increase and so the volume gets louder.

Info In the early days of electronics variable resistors were used in circuits that compared a variable fraction of some unknown voltage (also called a potential) to a fixed reference voltage. When the two voltages were equal, the operator read the fraction off the variable resistor and so computed the value of the unknown voltage. Because of their common use for measuring potentials in this way variable resistors became known as **potentiometers** (meaning potential measurers) and the name has stuck. It is still very common to hear a variable resistor called a potentiometer or **pot**. You should become familiar with this term.

Variable resistors come in a wide variety of body styles (Figure 3-20). Special linear sliders are used in a few applications, light dimmers and audio mixing consoles for example, but the most common are rotary. Like fixed resistors, these come in different sizes depending on the amount of power that they must dissipate. The most common ones can be turned through 3/4

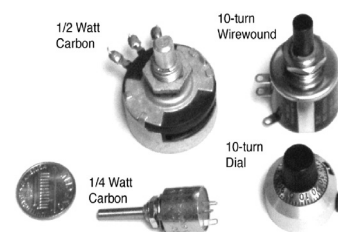


Figure 3-20 Variable Resistors

of a turn but precision ones are available that take 10 full turns to go from lowest to highest setting. These are used in situations where very fine control of a value is required.

Remember Any time that you see a resistor symbol with an arrow sticking out of the side you should look for a component with a knob to turn. It is a variable resistor and not a standard little painted cylinder.

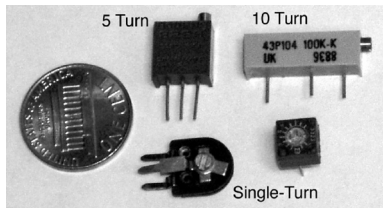


Figure 3-21 Trimmer Resistors

Standard panel mounted variable resistors are used when a value must be easily or frequently changed. There are occasions when you need a variable resistor in order to get exactly the correct value of resistance but the resistance needs to be set only one or only very rarely.

In such a case we can use a Trimmer resistor (Figure 3-21). This is a tiny variable resistor which must be adjusted with a screwdriver or special tool. These are normally hidden away inside the cases of instruments and only adjusted during servicing. Like panel mounted resistors, they come in single turn and multi-turn varieties.

3.7 Measuring Instruments

Although they are not components in the strict, as they are not usually built into circuits, this is also a convenient place to take a first look at some of the tools that we use to examine the internal state of circuit. The most common tools that we have are the Ammeter and the Voltmeter.

3.7.1 Voltmeters

As mentioned in section 2.1.3, a voltmeter is an instrument for measuring electric potential differences. It needs two terminals to connect to two places in a circuit and it has a readout, either a dial or a digital display, to show the potential difference between the two terminals. An ideal voltmeter would not disturb the circuit to which it was attached in any way. Any real voltmeter must draw some current from the circuit in order to operate. This means that any real voltmeter behaves rather like a resistor and thus does alter the currents and voltages in the circuit to which it is attached. A good voltmeter should affect the circuit as little as possible and so should have as high a resistance as possible. A typical little plastic multimeter (see below) has an effective resistance of about $10\text{M}\Omega$. This means that you can safely connect it to circuits that use resistors in the $\text{k}\Omega$ range without any effect but must seek something more exotic if you want to work with circuits that use $\text{M}\Omega$ resistors or operate with currents in the nA range.

3.7.2 Ammeters

An ammeter is an instrument for measuring the flow of electric current. As such it cannot be connected to a circuit; the current would then continue to flow in the circuit and the ammeter would not be able to detect it. Instead it must be inserted into the circuit. Old-style dial ammeters use the force that a current creates when it flows through a wire placed in a magnetic field to move a needle and give a reading proportional to the current. Modern ammeters are basically a small resistor with a sensitive voltmeter connected across the terminals. You break open the circuit that you wish to measure and insert the ammeter. The circuit current flows through the resistor and creates a voltage drop that is measured by the voltmeter. The system is calibrated in terms of the current in the resistor so that you don't have to worry about what is going on inside. When the ammeter is in the circuit there is a small voltage drop across it that alters the behaviour of the circuit slightly. This means that a good ammeter should have as low a resistance as possible. Typical values are in the few Ω range for currents in the mA .

3.7.3 Multi-Meters

A multi-meter is a simply a meter with a switch that allows it to work as an ammeter or a voltmeter depending on the setting. They originated in the analog dial days but all current ones use a digital display, which is much easier to read. Mass production has lead to a reasonable quality battery powered Digital MultiMeter (DMM) costing as little as \$10. Such a meter can

measure DC and AC voltages and currents with an accuracy of about 1% as well as measuring resistances and possibly some other functions.

More expensive battery powered meters offer greater reliability and more ranges. Some may add the ability to measure currents (in the A range) magnetically, requiring no contact between the meter and circuit. Still more expensive, wall powered meters, offer greater precision, with 6-digit displays and accuracies that can be traced to fundamental standards at the National Institutes of Standards and Technologies. Some can measure extremely small voltages and currents and feature exceptional input resistances in voltage mode, up to thousands of $M\Omega$.

Summary

Exercises

1. A resistor has four color bands, red, yellow, orange, gold, from left to right. What can you tell about the resistor from these bands?
2. What markings would you expect to find on a $10k\Omega$, 1%, resistor?
3. I am very careful to make sure that there are no 200Ω in my electronics teaching lab. What markings would you expect to find on such a resistor and why could these lead to confusion?
4. One standard measure of the capacity of a battery is the Ampere-Hour (Amp-Hr). This is the amount of charge drawn from the battery when it supplies a current of 1 Ampere for a period of 1 hour. A typical car battery might be rated at to supply 12V with a capacity of 250Amp-Hr. How many coulombs of charge does the battery hold?
5. How many Joules of energy does the battery in question 3 hold?

Chapter 4: Simple DC Circuits

4.1 Introduction

There are two different ways that we can apply circuit theory. In **circuit analysis** we find the currents and voltages in a given circuit. In **circuit synthesis** we find the component values needed to make a new circuit do what we want. For the moment we will stick to analysis and will start with some simple situations and work to more complex ones.

Later in the book we shall learn how to work with systems in which the voltages and currents vary more or less rapidly in time. For the moment we shall concentrate on the simpler case where the voltages and currents in the circuit are fixed. In such a case the currents always flow in the same direction and we call such uni-directional currents **Direct Currents** and so speak of circuits in which the voltages and currents do not vary in time as **Direct Current Circuits** or **DC circuits**.

4.2 Simple Circuits

The simplest circuit we can build consists of a voltage source and a resistor connected by wires. This is shown in Figure 4-1.

Since the battery is connected directly to the resistor by wires, the voltage across the resistor is equal to the battery voltage, 9V. This means that we can calculate the current flowing in the wires and in the resistor using Ohm's law.

$$I = V/R = 9V/2200 = 0.00409A = 4.1mA$$

Note **Series Connections:** The same current flows in the resistor as in each of the wires because they are connected end-to-end, in series. All the current that flows out of one flows into the next. Any two components that are connected end-to-end, in series, always have the same current flowing in them.

If we add a second resistor to the circuit, in series with the first, we get the slightly more complicated circuit of Figure 4-2. We can analyze this circuit to find the current in each resistor and wire and to find the voltage at each point relative to the negative terminal of the battery—our ground reference. First we notice that all the components are in series so that the same current flows in each wire and resistor. To make this clearer, think of the plumbing analogy of the circuit (Figure 4-3).

All the water that flows in the first wire, pipe ab, has to flow through the first resistor, pipe bc. Similarly, all the water in that resistor has to leave through the next wire, pipe cd, and then to flow through the second resistor, pipe de. Finally, all the water flows back through pipe ef to the pump for recirculation. Thus the current, the number of gallons per second, passing through each pipe in the loop is the same although the water flows much more rapidly in the narrow resistor pipes than it does in the wide wire pipes.

Moreover, we can use our plumbing analogy to see what happens to pressure at each point in the pipe work and thus to see what happens to the voltage at each point in the circuit. The pump pushes the water out at point a and sucks it in at point f. The pressure is highest at point a and lowest at point f, which we will use as our zero of pressure. In the very wide pipes that represent wires there is little or no pressure drop so all the pressure is lost in the thin pipes. Obviously the pressure drop from a to f is equal to sum of the drop from b to c plus that from d to e.

Now we can transfer that information back from the analogy to the circuit (Figure 4-4). In the circuit, we say that there is no voltage drop in any of the wires so the total battery voltage of 9 volts must equal the sum of the voltages across the 2.2k resistor, R1, and across the 4.7k resistor, R2. So we have

$$V_{R1} + V_{R2} = 9V$$

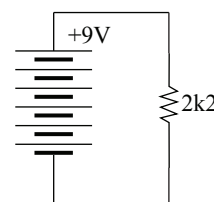


Figure 4-1 A Simple Circuit

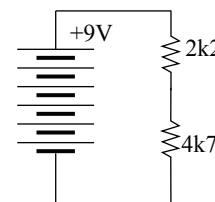


Figure 4-2 Series Circuit

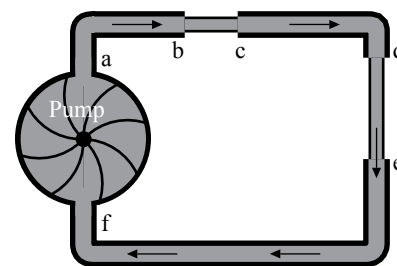


Figure 4-3 Series Plumbing Circuit

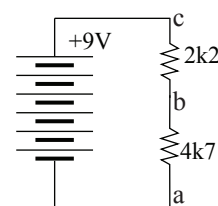


Figure 4-4 Series version 2

where V_{R1} is the voltage across resistor R1 and V_{R2} that across R2. We also know that each resistor obeys Ohm's law so, if we call the current in every part of the circuit I, we have two more equations

$$V_{R1} = I \times R1 = I \times 2200 \quad \text{and} \quad V_{R2} = I \times R2 = I \times 4700$$

We can combine these three equations into one

$$I \times 2200 + I \times 4700 = 9V \quad \text{or} \quad I \times (2200 + 4700) = I \times 6900 = 9V$$

and solve for I to find

$$I = 9/6900 = 0.0013A = 13mA$$

Now that we know the current in each resistor, we can find the voltage at every point in the circuit. If we start from the negative terminal of the battery and call the voltage there zero then the voltage at point a, V_a , is zero since that point is connected by a wire to the negative terminal. Point b is separated from point a by resistor R2 and thus the voltage at point b, V_b , is equal to the voltage across resistor R2.

$$V_b = I \times R2 = 0.0013 \times 4700 = 6.13V$$

The voltage at point c must be 9V since it is connected by a wire to the positive terminal of the battery. We can check that all is well by computing the voltage across R1 and comparing it with the value from Ohm's law.

$$V_{R1} = 9V - 6.13V = 2.87V \quad \text{and} \quad I \times R1 = 0.0013 \times 2200 = 2.87V.$$

The last computation we could do for this circuit would be to calculate the power dissipated in each resistor. We might do this in order to choose the right size resistor for the circuit. Since power is voltage times current, we have

$$P_{R1} = I \times V_{R1} = 0.0013 \times 2.87 = 0.0037W = 3.7mW$$

and

$$P_{R2} = I \times V_{R2} = 0.0013 \times 6.130 = 0.008W = 8mW$$

so we could use 1/4W resistors in both places.

4.2.1 The voltage divider

One of the most common simple circuits is called a **voltage divider**. It is built from two resistors connected together to give a simple 3-terminal device (Figure 4-5). The voltage divider is found in a huge variety of circuits because, as we shall see, it has the property of passing a constant fraction of the input voltage and passing it to the output.

This circuit is used as a four-terminal device with V_{in} and ground forming the input pair and V_{out} and ground forming the output pair. Let us connect the device to a battery and connect a voltmeter to the output (Figure 4-6). Then we can calculate the output voltage as a function of the input voltage.

Now, the magic of a voltmeter is that it draws so little current that we can call it zero current. Thus, the only current flows through the resistors and we know how to deal with that circuit. Following the previous example, the current in each of the resistors is

$$I = \frac{V_{in}}{R1 + R2}$$

then the output voltage, V_{out} is given by

$$V_{out} = I \times R2 = \frac{R2}{R1 + R2} \times V_{in}$$

This is why we call this circuit a voltage divider. The voltage coming out is a fraction of the voltage going in and we can choose the voltage by choosing the resistors. We call this very important equation the **voltage divider equation**.

Note $I \times R$ is voltage ACROSS resistor!

If we wanted to find the voltage at c using the 2.2kΩ resistor, R1, then we would have to be careful. Ohm's law tells us that there is

$$0.0013A \times 2200\Omega = 2.87V$$

ACROSS the resistor but that does NOT mean that the voltage at c is 2.87 volts. One end of the 2200Ω resistor (point b) is connected through a wire to the 9V battery so that the voltage at c is 2.87 volts LOWER than the voltage at point b, that is

$$V_c = V_b - I \times R1 = 9V - 2.87V = 6.13V.$$

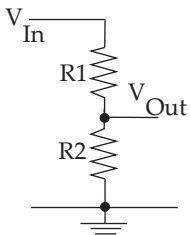


Figure 4-5 Voltage Divider 1

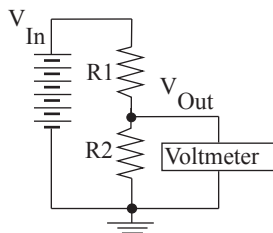


Figure 4-6 Voltage Divider 2

Note We are interested in the behavior of the circuit regardless of the particular values of the resistors so we analyze the circuit in purely abstract terms. We will end up with a formula for the output voltage in terms of the input voltage and the resistors.



Figure 4-7 Resistors in Series

Note Components such as wires and resistors are completely symmetric, either lead can be thought of as the input or the output.

4.3 Resistors in Series

A series connection (e.g. Figure 4-7) is one where the output of one component is connected directly to the input of the next with *nothing else connected to the junction*.

We have already seen an example of this in our second simple circuit (). Now we will use the ideas that we developed there to analyze the general case of two resistors in series. In the example we used the plumbing analogy to see that the current, I , in R_2 is the same as the current in R_1 . This is an example of a very general rule

Remember **Components in Series**
 When two components are connected in series the same current flows in each.

In addition, we see that the total voltage across the pair of resistors is equal to the sum of the individual voltages across each of the resistors so that, if we call the voltage between two points x and y V_{xy} , then we find

$$V_{ac} = V_{ab} + V_{bc}$$

Now we can apply Ohm's law to each resistor in turn giving us

$$V_{ab} = I \times R_1 \quad \text{and} \quad V_{bc} = I \times R_2$$

and then substitute into the first equation to give

$$V_{ac} = I \times R_1 + I \times R_2 = I \times (R_1 + R_2).$$

This means that the final form is the same as Ohm's law

$$V_{ac} = I \times (R_1 + R_2) = I \times R_t$$

where R_t , the total resistance of the pair of resistors is the sum of the individual resistors

$$R_t = R_1 + R_2.$$

So two resistors in series behave EXACTLY the same as a single resistor whose value is the sum of the individual resistors. If we think of our plumbing analogy we can see that this makes perfect sense; a larger value resistor is a longer pipe so two short thin pipes connected by a fat wire pipe behave the same as a single thin pipe of the same total length (Figure 4-8).

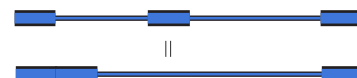


Figure 4-8 Pipes in Series

This extends to as many resistors in series as we want so. Thus we get the general rule for combining resistors in series

Remember **Resistors in Series**
 If we have n resistors, R_1, R_2, \dots, R_n , connected in series then the total resistance, R_t , is given by $R_t = R_1 + R_2 + \dots + R_n$.

Info It is obvious, but useful to remember, that the combined resistance of a set of series resistors is always greater than that of the largest resistor in the set.

Example

If we build the combination in Figure 4-9, then the total resistance is $R = 4,900 + 47000 + 100 = 52000 = 52k\Omega$.

Note that this allows us to build resistors with values that we can't otherwise find by connecting together ones that we can find.

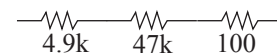


Figure 4-9 Series Example

4.4 Resistors in parallel

The other way that two resistors can be connected together is called a parallel connection, as shown in Figure 4-10.

In this case, the voltage across R_1 is the same as the voltage across R_2 , since the inputs are connected by a wire as are the outputs, so we have

$$V_{R_1} = V_{R_2} = V_{ab} = V.$$

This is an example of the important general rule

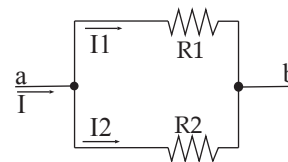


Figure 4-10 Resistors in Parallel

Remember **Components in Parallel**
 When two or more components are connected in parallel the voltage drop across each is the same.

The current, I , that flows in at point a splits into two currents I_1 and I_2 . Because we cannot create or destroy charge (remember rule 1 section 2.1.1), all the current that flows into the junction must flow out again and so we have

$$I = I_1 + I_2$$

Now we can use Ohm's law for each resistor and the fact that there is the same voltage across each to write

$$I = \frac{V}{R_1} + \frac{V}{R_1} = V \times \left\{ \frac{1}{R_1} + \frac{1}{R_1} \right\}$$

Once again the final equation has the current proportional to the voltage so the pair of parallel resistors behaves in the same way as a single resistor. This time, however, the value is more complicated. If we call the total resistance R_t , as before, then we have

$$I = \frac{V}{R_t} = V \times \frac{1}{R_t} = V \times \left\{ \frac{1}{R_1} + \frac{1}{R_1} \right\}$$

so that

$$\frac{1}{R_t} = \frac{1}{R_1} + \frac{1}{R_1}$$

The relationship generalizes to many resistors in the obvious way to give us the general rule for combining resistors in parallel.

Remember Resistors in Parallel

If we have n resistors, R_1, R_2, \dots, R_n , connected in parallel then the total resistance, R_t , is given by

$$\frac{1}{R_t} = \frac{1}{R_1} + \frac{1}{R_1} + \dots + \frac{1}{R_n}$$

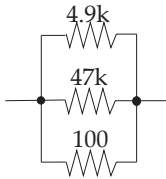


Figure 4-11 3 Resistors in Parallel

Example

If we connect the three resistors of the series example in parallel then we get the circuit of. We can find the new combined resistance with the parallel formula.

$$\frac{1}{R_t} = \frac{1}{4900} + \frac{1}{47000} + \frac{1}{100}$$

$$\frac{1}{R_t} = 0.000204\dots + 0.0000212\dots + 0.01 = 0.01022\dots$$

So that we find

$$R_t = \frac{1}{0.01022} = 97.8\Omega$$

Note This time, the combined resistance is smaller than the smallest resistor in the set. In general, combining resistors in series makes bigger resistors while combining them in parallel makes smaller resistors.

Common Combinations of Resistors

There are some combinations of resistors that are especially useful and common. The series ones are fairly obvious. two equal resistors in series make one of double the value three equal resistors in series make one triple the value etc.

The parallel ones are less obvious but are very useful two equal resistors in parallel make one of half the value three equal resistors in parallel make one of one third the value etc.

Two more combinations that are often useful in a quick analysis of a circuit are:- If you put a large resistor in series with one that is much smaller, then the combined resistance is almost the same as the larger resistor. If you put a large resistor in parallel with one that is much smaller, then the combined resistance is almost the same as the smaller resistor.

4.5 Combinations of Series and Parallel

Once you have a circuit with more than two resistors, the number of possible ways of connecting them gets so large that we can't study each one of them. Instead, we must develop methods to analyze any linear circuit. Most circuits can be broken down into smaller sections that are in series or in parallel and then analyzed a piece at a time. Some circuits cannot be broken up in this way and must use the more complicated methods of chapter 5—methods that are guaranteed to work for any circuit for which you can do the math.

As a rule, the methods of this section are useful for circuits with up to six or seven resistors and can be used on even more complex circuits if it is easy to see how the circuit is made up of simpler pieces. Once the circuit has more than about ten resistors, it is probably time to use a computer to do the analysis for you!

Figure 4-12 shows a moderately complex circuit that we can analyze by breaking it up. Our task is to find the current in each resistor and the voltage at each point in the circuit.

First, we identify a little sub circuit that is a pure series or pure parallel combination and replace it by its equivalent resistance, thus reducing the number of resistors in the circuit. We keep on doing this until there is only one resistor left. At that point, we can easily find the current. Next we reverse the process, replacing each resistor with the series or parallel combination from which it was formed and finding the current in each before going on to the next stage. The process terminates when we reach the original circuit, at which point we know every current and voltage and have solved the circuit.

We start the process by finding one or more pieces that form pure series or parallel combinations. In this case the only one present is the parallel combination of the 2.2k and 3.9k resistors circled in Figure 4-13.

Since these two resistors are in parallel we can replace them by a single resistor if it has the correct value. We find that value using the parallel resistance formula above.

$$\frac{1}{R} = \frac{1}{2200} + \frac{1}{3900} = 0.0004545... + 0.0002564...$$

$$R = \frac{1}{0.0004545 + 0.0002564} = \frac{1}{0.0007109} = 1406\Omega$$

Note We have kept only 4 significant figures in the value of R and we will keep no more later in the process so that there may be very small errors that creep into our calculations because of the rounding. They will be too small to affect the workings of the circuit but if more precision is needed then we can redo the calculations, keeping more digits in our answers as we go.

That leaves us with the circuit of Figure 4-14 in which we recognize that the 1406Ω resistor is in series with the 2.7k resistor, as shown.

Once we replace the two circled resistors by the correct single resistor $R = 1.406k + 2.7k = 4.106k$, we are left with the circuit of Figure 4-15. Now the 4.106k resistor is in parallel with the 10k resistor and we can find their parallel combination

$$\frac{1}{R} = \frac{1}{10000} + \frac{1}{3900} = 0.0001 + 0.000243...$$

$$R = \frac{1}{0.0001 + 0.000243} = \frac{1}{0.000343} = 2911\Omega$$

All that is left of our circuit is a pair of series resistors (Figure 4-16).

The last transformation is obvious. The 1k and 2.911k resistors are in series and their series combination is 3.911k. Thus, we can find the current coming from the power supply.

$$I = \frac{V}{R} = \frac{12}{3911} = 0.00307A$$

Now we can work our way back one step at a time. The two resistors are in series so the current in each resistor is the total battery current, 3.07mA.

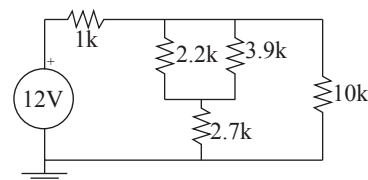


Figure 4-12

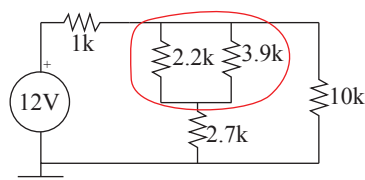


Figure 4-13

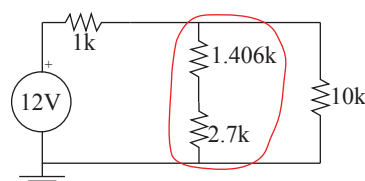


Figure 4-14

Warning When people first start this kind of analysis they sometimes see the 1406Ω resistor in parallel with the 10k but this is not the case. Two resistors must be connected at **both ends** to form a parallel combination.

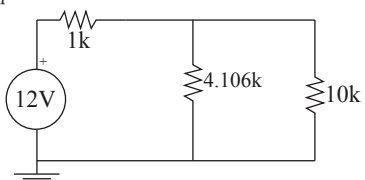


Figure 4-15

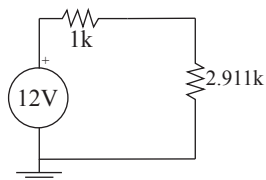


Figure 4-16

Significant Figures

Since electronic instruments have limited precision we are always interested in doing calculations with numbers of limited precision. For example, common resistors are only specified to within 5%. This means that only the first few digits in a number have any meaning. We call those digits the significant digits or significant figures. The most significant digit is the leftmost non-zero digit and then the digits decrease in significance.

In a number the significant digits are the those with the highest value. The safest way to find the significant digits is to write the number in scientific notation, that is to extract the powers of ten leaving the body of the number lying between 1 and 10. Here are some examples.

$$3400 = 3.4 \cdot 10^3 \quad 46.79 = 4.679 \cdot 10^1 \quad 0.0357 = 3.57 \cdot 10^{-2} \quad 0.0003015 = 3.015 \cdot 10^{-4}$$

Once the number is in scientific form, the most significant digit is the one to the left of the decimal point. So, a resistor value that is specified by

3 color bands has 2 significant digits. Now, when we perform calculations the number of actual digits in the numbers tends to increase wildly. For example, if we put a 500Ω resistor in parallel with a 1000Ω resistor we find that the combined resistance is

$$\frac{1}{R} = \frac{1}{500} + \frac{1}{1000} = 0.002 + 0.001 = 0.003 \text{ so that } R = 333.333333 \dots \Omega$$

and it takes an infinite number of digits to represent the answer. However, most of those digits are meaningless since the values from which they were computed only had 2 significant digits. We can express a more sensible answer by keeping as many significant digits in the answer as there significant digits in the original values. So the best answer to give for the total resistance is 330Ω where we have rounded the answer to 2 significant digits.

In general, the result of a calculation has no more significant digits than the least accurate number that went into it. Thus if we multiply a 3 s.f. number (that is a number with 3 significant figures) by a 2 s.f. number we get an answer that is only accurate to 2 significant figures.

Warning! You should only round to the final number of significant figures at the END of a calculation. You must keep more figures in intermediate steps otherwise your values get less and less accurate due to rounding error. As a general rule I would always carry at least 2 extra digits through any calculation. So if I am working with 2 s.f. quantities, I would carry 4 significant figures through all calculations and only round to 2 figures at the very end.

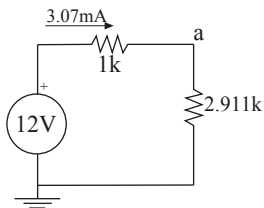


Figure 4-17

Note Point a is not really a point. Instead it is actually a more complex region that incorporates several junctions. Because all the points are connected by wire, they are all at the same potential; in this case 8.93V.

Now we can find the voltage at point a (Figure 4-17) in two ways. The first uses the 2.911k resistor and finds that

$$V_a = I \cdot R = 0.00307 \times 2911 = 8.93V$$

The second uses the 1k resistor and, noting that its left hand end is at 12V, finds that the voltage at point a is

$$V_a = 12V - I \times R = 12 - 0.00307 \times 1000 = 12 - 3.07 = 8.93V$$

Thus, the two methods agree, as of course they must. This kind of check, computing the same result by two independent methods, is very useful in circuit analysis since it helps catch arithmetic errors when they are made. If the errors are not caught, then they propagate into all the later steps and can force us to redo a large amount of work.

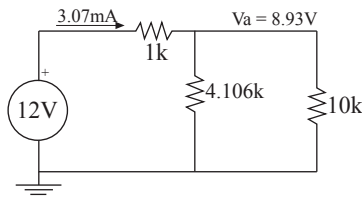


Figure 4-18

At this point we know all the voltages and currents so it is time to undo the next step. The 2.911k resistor came from the parallel combination of a 4.106k and a 10k resistor. When we replace those resistors we get the situation in Figure 4-18.

Since we know the voltage across each of the new resistors, we can find the current in each. For the 10k resistor we have

$$I = \frac{V}{R} = \frac{9.93}{10000} = 0.893mA$$

and for the 4.106k resistor we have

$$I = \frac{V}{R} = \frac{9.93}{4106} = 2.177mA.$$

Once again we can check our work because a total of 3.07mA flows into junction a so that the same total must flow out and, indeed, we have

$$I_{out} = 2.177 + 0.893 = 3.07.$$

In the next step (Figure 4-19) we replace the 4.106k resistor by the series combination from which it was made.

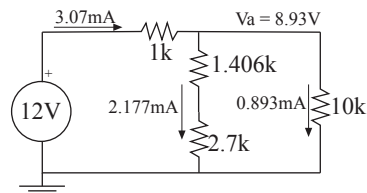


Figure 4-19

This time we know the current in the series pair and we need to find the voltage at point b. Again, we will do it by both methods to check our work.

$$V_b = 0.002177 \times 2700 = 5.88V \quad \text{or} \quad V_b = 8.93 - 0.2177 \times 1406 = 5.87V$$

The two answers do not quite agree because of rounding error, as we foretold near the beginning of the process. They are extremely close—the error is less than 0.2%—so we believe that rounding is the only problem and that our calculations are still on track.

In the last stage, we replace the 1.406k resistor with the parallel combination of 2.2k and 3.9k resistors from which it came (Figure 4-20).

Each of the new resistors has the same voltage across it, $8.93 - 5.88 = 3.05\text{V}$, so we find the current in the 2.2k resistor to be

$$I = \frac{3.05}{2200} = 1.386\text{mA}$$

and that in the 3.9k resistor

$$I = \frac{3.05}{3900} = 0.782\text{mA}$$

Finally, we check is that the currents in this middle chain add up

$$1.386\text{mA} + 0.782\text{mA} = 2.17\text{mA}$$

so the answers agree to within our rounding error.

We have found the voltage and current at every point in the circuit and the circuit analysis is complete. This is a very typical example of this kind of analysis and should serve as a guide for the first few that you do.

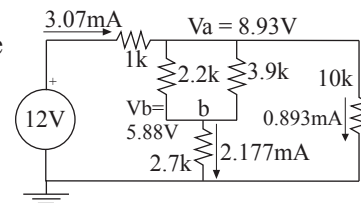


Figure 4-20

4.5.1 The Voltage Divider Again

When we last looked at the voltage divider we found that the ratio of the two resistors determined the relationship between the input and output voltages. Thus we can use a voltage divider to convert any voltage to a lower voltage by a suitable choice of the resistors. As we shall see, we need to add one more piece of information before we can make a unique choice of resistors.

Let us try to design a voltage divider to apply 6V to a 10k resistor when we only have a 9V battery. If we connect a 10k resistor to the output of the divider then we get the circuit of Figure 4-21.

According to what we learned in section 4.2.1, we require that

$$\frac{R_2}{R_1 + R_2} = \frac{6}{9}$$

so that

$$R_2 = 2 \times R_1$$

Let us pick some values and see if the circuit works. If we choose $R_2 = 10\text{k}$ and $R_1 = 5\text{k}$ then we have the circuit of Figure 4-22.

The two 10k resistors are in parallel so that we can replace them by their parallel combination of 5k (remember the common parallel combinations) to get the circuit of Figure 4-23.

Now we can find the total current very easily

$$I = \frac{9}{5000 + 5000} = 0.9\text{mA}$$

and thus find the voltage at point b

$$V = I \times R = 0.0009 \times 5000 = 4.5\text{V}$$

which is NOT 6V. This is no mere rounding error!

So, what has gone wrong? The computations in chapter 2 were made assuming that no current flowed out from the junction of the two resistors into the load. That is not the case here. Instead, if the divider were to work then a current of $6\text{V}/10\text{k} = 0.6\text{mA}$ would flow in the load resistor, a current that is nearly as large as the total current flowing in the rest of the circuit. So what do we do? For the voltage divider to work as we want we must choose the current that flows down the divider chain, through R_1 and R_2 , to be much larger than the current that flows into the load. As a rule we make the chain current at least 10 times the load current and then expect an error in the output voltage of no more than 10%.

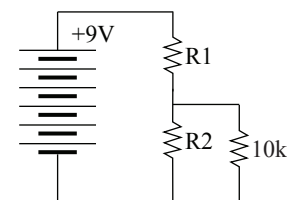


Figure 4-21 Voltage Divider With Load

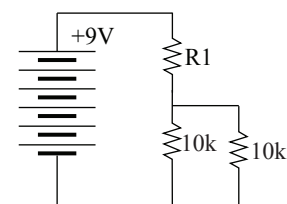


Figure 4-22

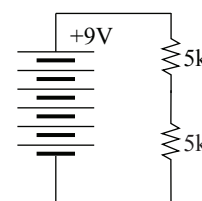


Figure 4-23

Example

Let us reduce the voltage divider resistors by a factor of 10 to get values of 500Ω and 1k. Now the parallel combination of 1k and 10k is

$$R = \frac{1}{\frac{1}{0.001} + \frac{1}{0.0001}} = 909.1\Omega$$

so that the total current flowing is

$$I = \frac{9}{500 + 909.1} = 6.387\text{mA},$$

the voltage at point b is

$$V_b = 909.1 \times 0.006387 = 5.81\text{V},$$

and the circuit makes an error of only 0.19V or about 3%.

Info Rules for choosing resistors in a voltage divider

Choose the chain current, I_c , to be some large multiple of the load current; 10 is the minimum value.

Choose the ratio of the two resistors to give the desired output,

$$V_{\text{Out}} = \frac{R_2}{R_1 + R_2} \times V_{\text{In}}$$

Choose the sum of the two resistors to give the designed chain current.

$$I_c = \frac{V_{\text{In}}}{R_1 + R_2}$$

Solve these two equations for R_1 and R_2 .



Figure 4-24 Series Connection

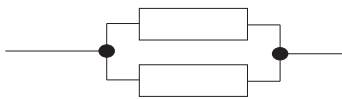


Figure 4-25 Parallel Connection

Summary

Two 2-terminal components are in **series** if the output lead of one connects only to the input lead of the next. In that case the same current flows in each component.

Two 2-terminal components are in **parallel** if the two input wires are connected together and the two output wires are connected together. In that case the voltage drop across each component is the same.

If n resistors R_1, R_2, \dots, R_n are connected in series then the combined component behaves like a resistor of value R_{tot} where

$$R_{\text{tot}} = R_1 + R_2 + \dots + R_n$$

If n resistors are connected in parallel then the combined component behaves like a resistor of value R_{\parallel} where

$$\frac{1}{R_{\parallel}} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_n}$$

Two resistors connected in as shown on the right form a **voltage divider**. So long as the current drawn from the output terminal is very small compared to the current flowing in the resistors the output voltage is given by

$$V_{\text{Out}} = \frac{R_2 \times V_{\text{In}}}{R_1 + R_2}$$

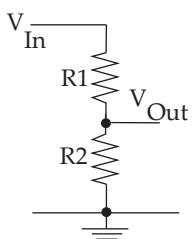
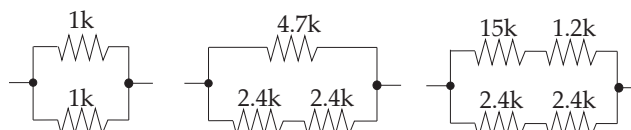


Figure 4-26 Voltage Divider

Exercises

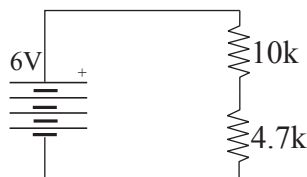
1. A 270Ω, 5% resistor is connected in series with a 1500Ω, 5% resistor. Find the largest and smallest values for the combined resistance.
2. If the resistors of question 1 are connected in parallel instead of in series, what are the largest and smallest possible values of the new combined resistor?
3. Find the combined resistance of each of the following



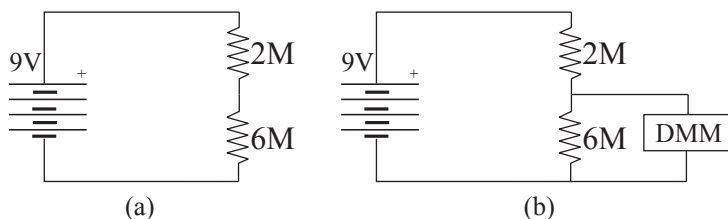
4. A dummy load is a resistor that can be connected to the output of an amplifier while you are testing the amplifier. It has to present the same resistance as the working load (for ex-

ample a loudspeaker) but does nothing with the power except heat up. One common way to build such dummy load is by connecting together a large number of small resistors that share the total power because several low power resistors can cost less than a single high-power resistor. Show how to connect together four 8Ω , $1/4\text{W}$ resistors to make a single 8Ω , 1W resistor for use as a dummy load.

5. A voltage divider made from a $10\text{k}\Omega$ resistor and $4.7\text{k}\Omega$ resistor is connected across a 6V battery as in the figure below. Find the voltage at the junction of the two resistors. Find the power dissipated in each of the two resistors and the power delivered by the battery.



6. A light sensitive switch needs a voltage of 0.45V to compare to the voltage coming from the light sensor but it only has a 5V power supply voltage available. Assuming that the circuit needs to draw less than $1\mu\text{A}$ of current from the 0.45V source, design a voltage divider to generate the 0.45V from the 5V power supply.
7. Consider the problem of delivering power to an electric toaster. If the toaster is plugged into a socket very close to the point where power (115V) enters the house then it sees the full 115V . Normally, however, the toaster is plugged into a socket in the kitchen some distance away. To get to the kitchen, the power has to flow through $100'$ of 12 AWG wire (also called $\#12$ wire) and the voltage seen at the toaster is reduced.
- What is the total resistance of the wires? (Remember that the current has to flow $100'$ to get to the toaster and then $100'$ back to get to ground.)
 - When the toaster is plugged in at the power entry, so that it gets the full 115V , it draws 2kW from the supply. What is the resistance of the toaster?
 - What is total resistance of the toaster plus wires?
 - When the toaster is plugged in the 115V appears across that total resistance. How much current flows in the toaster and in the wires?
 - How much power is now drawn by the toaster and how much power is wasted in the wiring? (Notice that this means that the wires in the walls will get warm. This is why we put fuses or circuit breakers on household circuits. If too much current were allowed to flow in the house wiring then the wires could get so hot that they would start fires!)
8. a) A voltage divider made from a $2\text{M}\Omega$ resistor and a $6\text{M}\Omega$ resistor is connected across a 9V battery as in figure a below. What is the voltage across the $6\text{M}\Omega$ resistor?
- b) A digital multimeter with an input resistance of $10\text{M}\Omega$ is now used to measure the voltage across the $6\text{M}\Omega$ resistor as in figure b below. Explain why the meter does not display the correct voltage and calculate the voltage that it does display.



Chapter 5: Formal Analysis of DC Circuits

5.1 Introduction

There exist circuits that cannot be analyzed by the series/parallel method as well as circuits for which the work just gets too involved to be worthwhile. These circuits can still be solved using a formal analysis method based on Kirchhoff's laws. The formal methods of this chapter can find all the currents and voltages in any circuit made up of only resistors and batteries, no matter how complex. However, the analysis gets mathematically very involved for all but rather simple circuits. Once you have more than six or seven resistors in a circuit, it is often better to use a computer program to do the analysis. It is still important to understand the formal methods, because they are the methods that the computer programs use. The methods are ideal for computers because they don't get bored and don't make silly mistakes!

5.2 Kirchhoff's Laws

Kirchhoff's two laws underlie all formal methods. We have already met the first of these laws, Kirchhoff's current law—

KCL: The sum of all the currents leaving a junction in a circuit must be equal to the sum of all the currents entering that junction.

This is just our old rule about the conservation of current extended to general junctions where many wires meet. If current flows in from several sources then it is the total in-flowing current that matters and if it flows out through several routes the total outflow must match the total inflow. We often use this law in a different form. If we call currents that **flow into** the junction **positive currents** (that is, they are represented by positive numbers) and call those that **flow out** of the junction **negative currents** then Kirchhoff's current law becomes

KCL: The algebraic sum of the currents entering and leaving a junction must be zero.

The only difficulty that we encounter using this rule lies in recognizing a junction. Because points that are connected by a wire are essentially the same point (that is, they are at the same potential and current can flow completely between them), a complicated junction often appears in a circuit diagram as a set of simpler junctions connected together. Look at the examples in Figure 5-1. Each shows two fragments of a circuit diagram and in each case both pairs are electrically identical so the first shows a junction with 4 wires and the second a junction with 6 wires.

Kirchhoff's voltage law is an extension of the observation about the voltage in a circuit, or the pressure in a plumbing system, that we made when analyzing the two-resistor circuit in chapter 3. There we saw that voltage drops in series add together. If we add to that the rather obvious idea that every point in a circuit must have a unique voltage, then we get Kirchhoff's voltage law.

KVL: Around any closed loop in a circuit, the algebraic sum of the voltage drops must equal zero.

Remember Kirchhoff's Laws

- 1) The sum of all the currents leaving a junction in a circuit must be equal to the sum of all the currents entering that junction.
- 2) Around any closed loop in a circuit, the algebraic sum of the voltage drops must equal zero.

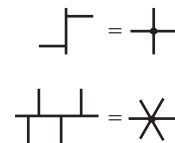


Figure 5-1 Types of Junction

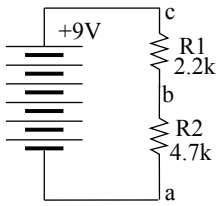


Figure 5-2 Simple Circuit

Note If we get the direction of the current wrong no harm will be done. We will just find that the current has a negative value. After all, a positive current flowing from top-to-bottom is the same thing as a negative current flowing from bottom-to-top.

5.2.1 A Simple Kirchhoff Example

Let us look again at our battery and two-resistor circuit from chapter 4 (Figure 5-2).

If we are going to analyze a circuit the Kirchhoff way then first we have to pick a direction for the current flow to give us a way to know negative from positive. We know that, when current flows in a resistor, the tail of the arrow is at a **higher** potential than the tip. Thus, that if we go along a resistor from tip-to-tail then we add in the voltage $V = I \times R$, if we go from tail-to-tip then we subtract the voltage.

To apply Kirchhoff's voltage law we pick a starting point, in this case point a. Then we work our way around, one component at a time, adding and subtracting voltages as we go. Point a is at 0V so when we go from point a to point c the voltage rises by 9V. From point c to point b we are going from the tail of an arrow to the tip, so we subtract the voltage across the resistor, $I \times R1$. Again, as we go from b to a we are traveling tail-to-tip so we subtract $I \times R2$. Thus we have

$$9V - I \times R1 - I \times R2 = 0V$$

$$I = \frac{9V}{R1 + R2} = \frac{9V}{2200 + 4700} = 1.3mA$$

5.2.2 A More Complex Kirchhoff Example

Figure 5-3 shows one of the simplest circuits that cannot be analyzed by the simple series/parallel method.

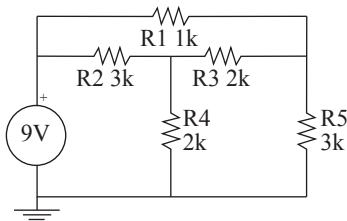


Figure 5-3

At first glance, it is easy to think that this can be broken down into series and parallel combinations. For example, isn't R1 in parallel with R2 and R3? Aren't R2 and R3 in series?

A more careful examination reveals that none of these combinations works. R1 is not in parallel with R2 because they are connected at only one end, R2 and R3 are not in series because R4 is connected at their midpoint, and so on for all other two and three resistor combinations. Instead, this must be analyzed formally. There are two slightly different ways to organize the analysis, although they come to the same answer and do most of the same working in the end.

Info In general, people find the first method, node analysis, a little easier to understand than the second method and so I recommend learning that method and just being aware that the other method exists.

5.3 Method 1: Node analysis

In this method we first label each unique node in the circuit—each point that is at a distinct potential, separated from all other points by voltage sources or resistors. These points are always formed by junctions between components and in this circuit there are four such points, plus the ground junction, labeled a, b, c, and d in Figure 5-4.

Of these four, only two are unknown nodes since the voltage at point a is set to 9V by the battery and point d is our ground reference. We shall solve the circuit by finding voltages at the two remaining nodes, b and c.

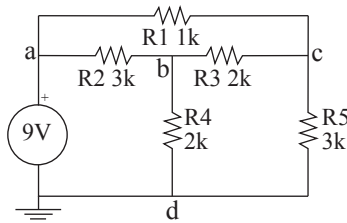


Figure 5-4

We start by assigning a current to each of the resistors, giving it a name and an assumed direction. The direction is quite arbitrary at this point and if we get the direction wrong the value of the current will simply be negative to show that it flows the other way. This gives us the situation in Figure 5-5.

Once the circuit has been labeled, with the nodes and currents identified, we can start the analysis. In this case there are 6 unknown currents and 2 unknown voltages. We are first going to find the voltages and then use those to find the currents.

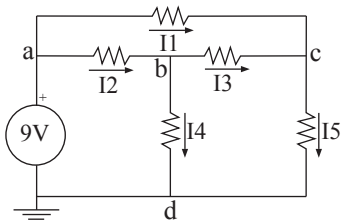


Figure 5-5

In order to find two unknown voltages we shall need two equations. We get those equations by applying Kirchhoff's current law at each of the unknown nodes. At node b we see that current I2 flows into the junction and currents I3 and I4 flow out of the junction. That means that

$$I2 = I3 + I4 \quad \text{or} \quad I2 - I3 - I4 = 0$$

At node c, currents I1 and I3 flow into the junction and current I5 flows out. Thus we have

$$I1 + I3 = I5 \quad \text{or} \quad I1 + I3 - I5 = 0$$

Next we use Ohm's law to write each of the resistor currents in terms of the node voltages, $Va (= 9V)$, Vb , and Vc .

Remember The current flows from the positive end of the resistor to the negative end. The current is always given by

$$I = \frac{V_{\text{Tail}} - V_{\text{Tip}}}{R}$$

where V_{Tail} is the voltage at the tail end of the arrow and V_{Tip} the voltage at the pointed end.

$$I_1 = \frac{V_a - V_c}{R_1} = \frac{9 - V_c}{1000} \quad I_2 = \frac{V_a - V_b}{R_2} = \frac{9 - V_c}{3000}$$

$$I_3 = \frac{V_b - V_c}{R_3} = \frac{V_b - V_c}{2000} \quad I_4 = \frac{V_b - 0}{R_4} = \frac{V_b}{2000} \quad I_5 = \frac{V_c - 0}{R_5} = \frac{V_c}{3000}$$

Note that I_4 and I_5 are particularly simple in form because one end of each is at zero potential.

Now we substitute these values for the currents back into the node equations to get

$$\text{At b: } \frac{9 - V_c}{3000} - \frac{V_b - V_c}{2000} - \frac{V_b}{2000} = 0$$

$$\text{At c: } \frac{9 - V_c}{3000} + \frac{V_b - V_c}{2000} - \frac{V_c}{3000} = 0$$

Multiplying both equations by 6000 we get

$$\text{At b: } 18 - 2 \times V_b - 3 \times V_b + 3 \times V_c - 3 \times V_b = 0$$

$$\text{At c: } 54 - 6 \times V_c + 3 \times V_b - 3 \times V_b - 2 \times V_c = 0$$

which simplify to

$$\text{At b: } 8 \times V_b - 3 \times V_c = 18$$

$$\text{At c: } -3 \times V_b + 11 \times V_c = 54$$

Now we have two simultaneous linear equations in two unknowns that we can easily solve. We will multiply the first equation by three and the second by eight and then add the two equations

$$\begin{aligned} 24 V_b - 9 V_c &= 54 \\ + -24 V_b + 88 V_c &= 432 \\ 0 V_b + 79 V_c &= 486 \end{aligned}$$

So that

$$V_c = \frac{486}{79} = 6.152\text{V (to 4 significant figures)}$$

Now we can use either the a or the b equation to find V_b . We will use the b equation

$$8 V_b = 18 + 3 \times V_c = 36.456\text{ V}$$

so that

$$V_b = 4.557\text{ V (to 4 s.f.)}$$

Once we know V_b and V_c , we can go back and find the resistor currents. Note that we round them all to 3 s.f. now that we are finished.

$$I_1 = \frac{9 - V_c}{1000} = 2.848\text{mA} = 2.85\text{mA} \quad I_2 = \frac{9 - V_c}{3000} = 1.481\text{mA} = 1.48\text{mA}$$

$$I_3 = \frac{V_b - V_c}{2000} = -0.80\text{mA} \quad I_4 = \frac{V_b}{2000} = 2.28\text{mA} \quad I_5 = \frac{V_c}{3000} = 2.05\text{mA}$$

and finally we can apply Kirchoff's current law to node a to find the current drawn from the battery

$$I_{\text{Battery}} = I_1 + I_2 = 4.33\text{mA}$$

So the complete circuit is solved as shown in Figure 5-6. Notice that the arrow on the current through R_3 has been reversed to conform with the negative answer that we found for I_3 .

Remember Node Analysis

Here are the rules for nodal analysis.

- 1) Label all the nodes in the circuit and identify the unknown nodes.
- 2) Assign currents to each battery and resistor in the circuit.
- 3) Write down Kirchoff's Current Law for each unknown node.
- 4) Write down Ohm's law for each resistor in the circuit.
- 5) Substitute the currents from step 4 into the equations of step 3.
- 6) Solve the resulting equations for the unknown node voltages.
- 7) Substitute the node voltages into the equations from step 4 to find the current in each resistor.

Note When we do calculations by hand we have to choose a suitable accuracy before we begin the calculation (unless we want to write down all the numbers on our calculators, most of which are meaningless). For the purposes of electronics it is usually adequate to get answers accurate to about 1% which means keeping 3 significant figures. Because of the rounding errors that will accumulate in the calculation, it is a good idea to carry 4 significant figures and round to 3 s.f. only at the end.

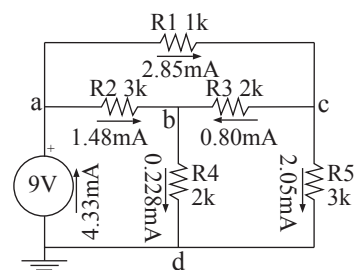


Figure 5-6 Solved Circuit

5.4 *Method 2: Loop Analysis

Loop Analysis provides a different method for performing essentially the same calculations. It is a little more complicated to set up than node analysis though it can offer some mathematical savings for complicated circuits. I include it for completeness rather than to suggest that you use this method.

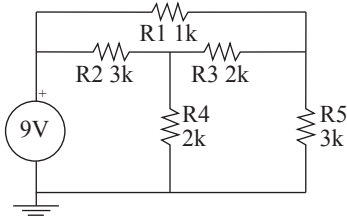


Figure 5-7

In this method, rather than finding the voltages at several unknown nodes we will find the currents flowing in the circuit by applying Kirchhoff's voltage law round circuit loops. Note that the node method finds node voltages using Kirchhoff's current law while the loop method finds loop currents using Kirchhoff's voltage law! Figure 5-7 shows our circuit again.

This time our preparation step is to identify all the circular sub circuits around which to apply Kirchhoff's voltage law. In this case we find

1. the loop running from a through R1 to c and back through R3 and R2 back to a
2. the loop running from a through R2 to b and then through R4 and the battery back to a
3. the loop running from b through R3 to c and then back through R5 and R4.

There are several other loops possible, e.g. one from a through R1 to c and back through R5 and the battery, but they are all mixtures of the three we have already found and contain no new information. Only the obvious loops need to be identified.

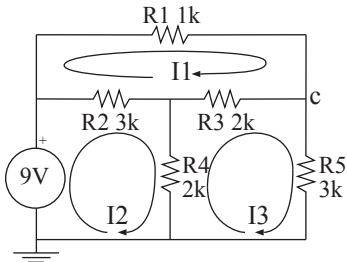


Figure 5-8 Loops Labelled

Once the loops are identified we associate each with a **loop current** traveling in the clockwise direction round the loop (Figure 5-8).

These loop currents are convenient mathematical tools not real currents. The real currents that flow in the individual resistors are mixtures of these loop currents. For example R4 has current I₂ flowing in it from top to bottom and current I₃ flowing in it from bottom to top so that the total current is $I_{R4} = I_2 - I_3$ in the downwards direction. The important property of the loop currents is that they automatically satisfy Kirchhoff's current law since they flow out of each node that they flow into.

The next step is to write the current in each resistor in terms of the loop currents. Here we have, assuming all currents flow left-to-right or top-to-bottom,

$$\begin{aligned} I_{R1} &= I_1 \\ I_{R2} &= I_2 - I_1 \\ I_{R3} &= I_3 - I_1 \\ I_{R4} &= I_2 - I_3 \\ I_{R5} &= I_3 \end{aligned}$$

Next we apply Kirchhoff's voltage law to each of the loops in the circuit, using Ohm's law and the resistor currents above to find the voltage across each resistor. This gives us

$$\begin{aligned} \text{Loop 1: } & -I_{R1} + R_3 I_{R3} + R_2 I_{R2} = 0 \\ & -1000 \cdot I_1 + 2000 \cdot (I_3 - I_1) + 3000 \cdot (I_2 - I_1) = 0 \\ \text{Loop 2: } & +9V - R_2 \cdot I_{R2} - R_4 \cdot I_{R4} = 0 \\ & +9 - 3000 \cdot (I_2 - I_1) - 2000 \cdot (I_2 - I_3) = 0 \\ \text{Loop 3: } & +R_4 \cdot I_{R4} - R_3 \cdot I_{R3} - R_5 \cdot I_{R5} = 0 \\ & 2000 \cdot (I_2 - I_3) - 2000 \cdot (I_3 - I_1) - 3000 \cdot I_3 = 0 \end{aligned}$$

Now we have three equations in three unknowns, I₁, I₂, and I₃. First we will rearrange them to collect all the I's on one side and the constants on the other side.

$$\begin{aligned} \text{Loop 1: } & 6000 I_1 - 3000 I_2 - 2000 I_3 = 0 \\ \text{Loop 2: } & -3000 I_1 + 5000 I_2 - 2000 I_3 = 9 \\ \text{Loop 3: } & 2000 I_1 + 2000 I_2 - 7000 I_3 = 0 \end{aligned}$$

There are several different ways to solve such a system, ranging from the simple elimination method that we used in the node method to matrix methods capable of handling thousands of simultaneous equations. By the time the problem is complicated enough to be worth using matrix methods on, it is probably time to solve the whole problem with a computer! We will

use the elimination method again, starting with I1. If we add twice the loop 2 equation to the loop 1 equation we will have

$$\begin{aligned} 6000 \times I_1 - 3000 \times I_2 - 2000 \times I_3 &= 0 \\ + -6000 \times I_1 + 10000 \times I_2 - 4000 \times I_3 &= 18 \\ 0 \times I_1 + 7000 \times I_2 - 6000 \times I_3 &= 18 \end{aligned}$$

Similarly we can eliminate I1 from equations 1 and 3 by subtracting 3 times equation 3 from equation 1

$$\begin{aligned} 6000 \times I_1 - 3000 \times I_2 - 2000 \times I_3 &= 0 \\ - 6000 \times I_1 + 6000 \times I_2 - 21000 \times I_3 &= 0 \\ 0 \times I_1 - 9000 \times I_2 + 19000 \times I_3 &= 0 \end{aligned}$$

Now we have only two equations in two unknowns. We repeat the process to eliminate I2 by adding 7 times the second of the new equations to 9 times the first

$$\begin{aligned} 63000 \times I_2 - 54000 \times I_3 &= 162 \\ - 63000 \times I_2 + 133000 \times I_3 &= 0 \\ 0 \times I_2 + 79000 \times I_3 &= 162 \end{aligned}$$

so that we can say

$$I_3 = 162/79000 = 2.051 \text{ mA}$$

Now that we know I3 we can work our way back up to find I2 and thence I1

$$\begin{aligned} I_2 &= \frac{19000 \times I_3}{9000} = \frac{19 \times 2.051}{9} = 4.329 \text{ mA} \\ I_1 &= \frac{-2000 \times I_2 + 7000 \times I_3}{2000} = \frac{7 \times 2.051 - 2 \times 4.329}{2} = 2.848 \text{ mA} \end{aligned}$$

Of course, the loop currents are not the final answer. We need to convert them back to the individual resistor currents and round them to 3 s.f.

$$\begin{aligned} I_{R1} &= I_1 = 2.85 \text{ mA} \\ I_{R2} &= I_2 - I_1 = 4.329 - 2.848 = 1.48 \text{ mA} \\ I_{R3} &= I_3 - I_1 = 2.051 - 2.848 = -0.80 \text{ mA} \\ I_{R4} &= I_2 - I_3 = 4.329 - 2.051 = 2.28 \text{ mA} \\ I_{R5} &= I_3 = 2.05 \text{ mA} \end{aligned}$$

These are, of course, the same answers that we found by the node method. The last step is to find the two unknown voltages at the top ends of R4 and R5 (points b and c in the node method).

$$\begin{aligned} V_b &= I_{R4} \times R_4 = 2.278 \times 2000 = 4.56 \text{ mA} \\ V_c &= I_{R5} \times R_5 = 2.051 \times 3000 = 6.15 \text{ mA} \end{aligned}$$

Loop Analysis

Here are the rules for Loop Analysis.

- 1) Identify the non-degenerate loops in the circuit.
- 2) Associate a named current with each loop.
- 3) Express the current in each component in terms of the loop currents.
- 4) Write down Kirchhoff's Voltage Law round each of the loops.
- 5) Write down Ohm's law for each resistor in the circuit.
- 6) Substitute the voltages from step 4 into the equations from step 3.
- 7) Solve the equations for the loop currents.
- 8) Substitute the loop currents into the equations from step 3 and find the component currents.
- 9) Use Ohm's law to find the unknown voltages.

5.5 Thévenin's Theory

One powerful way of simplifying the analysis of electronic circuits is to break them up into simpler pieces, analyze the pieces separately, and then understand the overall behavior in terms of the pieces. We can do this because of Thévenin's theory, which states that

Any two-terminal linear circuit can be replaced by a unique series combination of a voltage source and a resistor.

We call that unique combination the **Thévenin Equivalent** of the circuit.

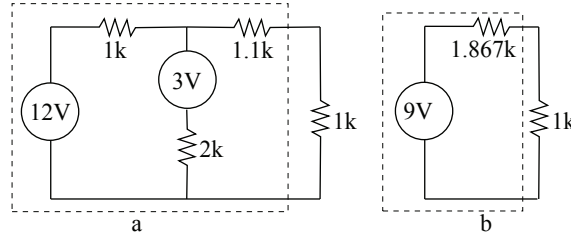


Figure 5-9 Thévenin Equivalents

Thévenin's theory means that we can replace some very complicated mess of batteries and resistors by a single battery and resistor without the circuit to which it is connected being able to tell the difference. So, for example, everything in the dotted box in Figure 5-9a can be replaced by the single battery and resistor of Figure 5-9b and the 1k load resistor will not be able to tell the difference. That means that the voltage across the load resistor and the current through that resistor will be just the same for the Thévenin circuit as it is for the real circuit no matter what load resistor is used.

How did I find the battery and resistor that replaced this circuit? Here are the rules

Remember Steps for finding the Thévenin equivalent of a 2-terminal device.

- 1) Draw the circuit with the terminals of the device connected by a wire and calculate the current that flows in the wire. This is called the short circuit output current, I_{sc} .
- 2) Redraw the circuit with nothing, including the load resistor, connected to the terminals. Calculate the voltage between the output terminals. This is the open circuit output voltage, V_{oc} .
- 3) If both values are non-zero then the Thévenin equivalent consists of a voltage source $V = V_{oc}$ in series with a resistance $R_{th} = V_{oc} / I_{sc}$.
- 4) If both values are zero then there is no Thévenin voltage source and you must apply a voltage to the terminals, measure the current, and compute the single resistance value from Ohm's law.

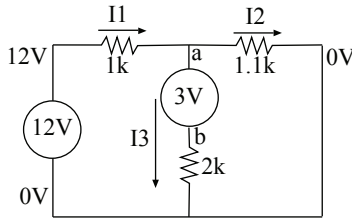


Figure 5-10 Circuit for I_{sc}

Let's apply this to the circuit of Figure 5-9. The first step tells us to draw the circuit with the output terminals connected by a wire. That gives us the circuit of Figure 5-10.

We need to analyse this circuit to find the current in the wire, which is clearly the same as I_2 , the current in the 1.1k resistor. We can find that current using the node method. I have already labelled all the nodes in the circuit with their voltages, known and unknown, and labelled all the resistors with their currents. Since there is no ground point marked on the circuit I have had to decide where to put 0V and selected the negative terminal of the main power supply. This is the most common choice.

It appears that there are two unknown nodes, a and b. However, these two nodes lie either side of the 3V power source and so, while we do not know what V_a or what V_b is, we do know that

$$V_b = V_a - 3V.$$

This means that we have only one real unknown, V_a and so we need write the current equation only for node a. Current I_1 flows into node a and currents I_2 and I_3 flow out so that we have

$$I_1 = I_2 + I_3.$$

So, when we substitute in the resistor currents using Ohm's law, we get

$$\frac{12 - V_a}{1000} = \frac{V_a - 0}{1200} + \frac{V_a - 3}{2000} = 0.$$

We can simplify this by first multiplying by the common denominator, 6000, to get

$$72 - 6 \times V_a = (5 \times V_a - 0) + (3 \times V_a - 9)$$

and then collecting terms together to find

$$81 = 14 \times V_a \quad \text{or} \quad V_a = \frac{81}{14} = 5.786V$$

Finally, we can use this voltage to find the short-circuit current, the current that flows in the output wire. Clearly, that current is I_2 and so we have

$$I_{sc} = I_2 = \frac{5.786}{1100} = 5.260mA$$

The second step in finding the Thévenin equivalent requires us to redraw the circuit and recalculate all of the voltages and currents. This time we remove the load completely to get the circuit of Figure 5-11

This time we have to find the open circuit voltage, the voltage between the two open wires. Since the lower wire is at 0V, the open circuit voltage is equal to the voltage at point c.

This time we seem to have three unknowns, V_a , V_b , and V_c . However, once again we know that

$$V_b = V_a - 3$$

and we also know that $I_2 = 0$ since there is nowhere for the current to flow from the end of the output wire. This means that

$$V_c = V_a.$$

Thus, once again we have only a single unknown V_a and need write down only 1 current equation. Since $I_2 = 0$, at node a we have $I_1 = I_3$ or, putting in Ohm's law,

$$\frac{12 - V_a}{1000} = \frac{V_a - 3}{2000}.$$

It is easy to see that the solution of this is

$$V_a = 9V \quad \text{so that} \quad V_{oc} = 9V.$$

Since neither the short-circuit current nor the open circuit voltage is zero we can use the third step to find that

$$V_{th} = V_{oc} = 9V \quad \text{and} \quad R_{th} = \frac{V_{oc}}{I_{sc}} = \frac{9V}{5.26mA} = 1.710k\Omega$$

We could now check that the procedure worked by calculating the current that flows in the 1k load resistor in Figure 5-9a and by calculating the current that flows in the equivalent resistor in figure Figure 5-9b. This is left as an exercise.

Note Thevenin in the world

The same procedure works just as well for real circuits measured with real voltmeters and ammeters as it does for theoretical circuits. You can find the Thévenin circuit for a physical box with two terminals by connecting a voltmeter to it to measure the open circuit voltage and an ammeter to measure the short circuit current. The only time this does not work is for a circuit that contains no voltage source. In that case both values will be zero and the answer is indeterminate. Then you have a circuit that is equivalent to a single resistor and can measure its resistance with a standard ohmmeter!

In practice you must be very careful when trying to measure the short circuit current of a device or you may easily destroy both the device under test and the test meter. It is better to measure the current through a small resistor and calculate the short circuit current from that.

5.5.1 The Thévenin Equivalent of a Power Supply

In Chapter 3 we met the idea of a power supply, a device that can provide a fixed current or voltage to a circuit. In the real world there are no ideal power supplies. In a real power supply the output voltage and current are not independent. Let us consider first the most common kind of power supply, a voltage source such as a battery.

Thévenin Equivalent of a Voltage Source

If we plot the output voltage of a real voltage source as we increase the amount of current drawn from the supply then we get a V-I curve that looks like Figure 5-12.

The ideal voltage supply produces a constant 5V output regardless of the amount of current drawn. The output from the real device falls as the current drawn rises. Over a fairly wide range, from 0A to about 0.3A, the output slope is fairly constant and we can represent the real power supply by its Thévenin equivalent.

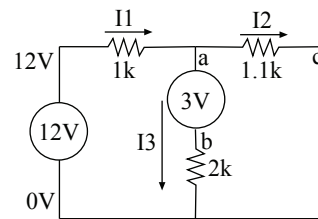


Figure 5-11 Circuit for V_{oc}

Note The voltage at point a, and all of the currents in the circuit, are completely different in the V_{oc} calculation from the values that we found in the I_{sc} calculation. By removing the short-circuit wire from the output we have altered the circuit and changed every voltage and current in it. This is why we have to start from scratch when we calculate the V_{oc} even though we already found V_a when we were computing I_{sc} .

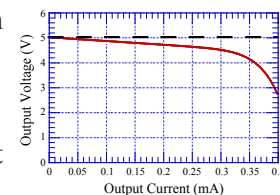


Figure 5-12 Power Supply V-I Curve

Note Thévenin's Theory relies on the linearity of the I-V plot and so we can only find a Thévenin equivalent for the straight line portion of the behavior of a real battery.

This is a situation where we cannot find the Thévenin equivalent using the short-circuit current because that would either destroy the power supply or, if it was well designed, cause it to shut down and cease operation. In either case we would get no useful information. However, we can use any two points and a little more math to get the same information. The Thévenin voltage is easy because that is the open circuit voltage, 5V. Then the voltage at any output current I is just

$$V_{Out} = V_{Th} - I \times R_{Th}$$

which can easily be rearranged to give

$$R_{Th} = \frac{V_{oc} - V_{oc}}{I}$$

At a current of 0.3A the voltage is 4.497V so that

$$R_{Th} = \frac{5 - 4.997}{0.3} = \frac{0.503}{0.3} = 1.68\Omega$$

Thus our power supply acts like an ideal voltage source in series with a 1.68Ω resistance. We call this resistance the **internal resistance** of the power supply. An ideal voltage source has zero internal resistance.

Info A high quality, wall powered supply will have an internal resistance of a tiny fraction of an Ohm. Batteries have internal resistances that vary with the size and design of the battery. A 9V carbon-zinc (“Heavy Duty”) battery may have an internal resistance of more than 30Ω while alkaline AA, C, and D batteries have resistances of 0.4Ω, 0.2Ω, and 0.1Ω respectively. At the other extreme a car battery, which must be capable of supplying several hundred Amps of current to start a car, has an internal resistance of 0.001Ω or less.

Thévenin Equivalent of a Current Source

A constant current source should adjust the voltage between its terminals as needed to keep the current flowing at the design value regardless of the resistance of the load. These components are much less common than voltage sources but can be made in much the same way and find considerable use inside amplifiers. It is easiest to understand their behaviour if we look at a conventional I-V plot (Figure 5-13).

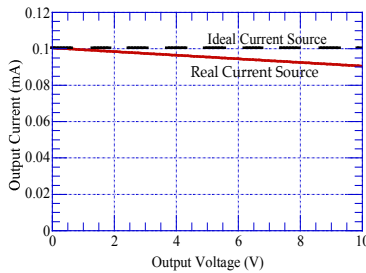


Figure 5-13 Current Source I-V Curve

Again, the real source is imperfect. The current available drops as the voltage needed rises. This time the behavior is linear over the range plotted and so we can find a Thévenin equivalent for the complete 0-10V range. A current source must have a path for current to flow and so we cannot measure the open-circuit voltage of a current source. We can however measure the short-circuit current and can use that and one other point to compute the Thévenin equivalent.

When $V_{out} = 0$, we have $I = 0.1A$. When $V_{out} = 10V$, we have $I = 0.09A$ so that

$$0 = V_{Th} - 0.1 \times R_{Th} \text{ and } 10V = V_{Th} - 0.09 \times R_{Th}$$

We can solve these two simultaneous equations to find

$$V_{Th} = 100V \text{ and } R_{Th} = 1000\Omega$$

In general a perfect current source has an infinite internal resistance and realistic constant current sources used in circuits might have an internal resistance of 100kΩ while a mains operated power supply in constant current mode could have an internal resistance of thousands of MΩ

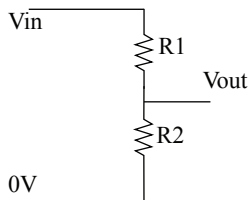


Figure 5-14 Voltage Divider V_{OC}

5.5.2 Thévenin and the voltage divider

Let us apply Thévenin theory to the voltage divider. It will help us to understand the behavior that we saw earlier in this chapter and to see why the rules we developed there work.

Figure 5-14 shows our old friend all ready for measuring the open circuit output voltage. In fact, this is exactly the voltage that we calculated in chapter 2, when we put a voltmeter directly across the output

$$V_{OC} = V_{out} = \frac{R2}{R1 + R2} \times V_{in}$$

We calculate the short circuit output current by connecting a wire across the output.

Now all the current flows in the wire and none in the resistor, since the two ends of the wire are at the same voltage. This makes the current that flows in the wire

$$I_{sc} = \frac{V_{in}}{R1}$$

So the Thévenin voltage $V_{Th} = V_{OC}$ and the Thévenin resistance, R_{Th} , is

$$R_{Th} = \frac{V_{Th}}{I_{sc}} = \frac{R2 \times V_{in}}{R1 + R2} \times \frac{R1}{V_{in}} = \frac{R1 \times R2}{R1 + R2}$$

That is, the Thévenin resistance is equal to the parallel combination of R1 and R2.

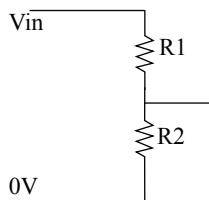


Figure 5-15 Voltage Divider I_{sc}

So, what does that tell about the behavior of the divider circuit when connected to a load? Well, for the output voltage to be essentially equal to the Thévenin voltage, we require that the load resistance be very large compared to the Thévenin resistance. Our rule for choosing the resistances of the divider to be small compared to the load resistance is an example of a general rule for connecting two devices together:

Remember A rule for connecting two circuits.

To transfer a voltage from one circuit to another, the input resistance of the destination should be at least ten times the output resistance of the source.

5.5.3 The importance of Thévenin's theorem

As we shall see, even a straightforward electronic component such as a power supply has a fairly complicated internal circuit, while a measuring instrument such as a digital multimeter or an oscilloscope may have hundreds of components inside it. If we had to take those components into account every time we used a power supply or a voltmeter then we would never be able to understand any but the simplest of circuits. Thévenin's theorem makes understanding electronics possible because it allows us to break a complex piece of equipment up into simpler sections and understand each piece at a time.

This idea of the Thévenin equivalent is so prevalent that we have special terms for the Thévenin resistance of devices used as voltage sources and as loads. We call the Thévenin resistance of a device that is a source of a voltage or signal the **output impedance** of the device. We call the Thévenin resistance of a load such as a voltmeter, the input of an amplifier, or the input of an oscilloscope, the **input impedance**. Then when we connect two pieces of apparatus we need only compare the output impedance of the source with the input impedance of the load to know what will happen.

Example

Consider trying to connect a loudspeaker directly to the output of a signal source such as a CD player. The output of a CD player might typically provide a voltage that varies up to 2V but with a resistance of 1k or more. A typical loudspeaker has a Thévenin resistance of only 8 Ohms so when we connect them we get the circuit of Figure 5-16.

Now we can find out how well the signal is passed to the loudspeaker by computing the voltage at a. This is easy if we note that the two resistors form a voltage divider so that

$$V_a = \frac{8}{1000 + 8} \times 2 = 0.0159\text{V}$$

Thus, almost none of the voltage appears across the loudspeaker. Practically no current flows in the loudspeaker, and so no sound is heard. This is why we have to put an amplifier between the CD-player and the loudspeaker. It does not make the voltage much bigger but an amplifier has a very low output impedance, a few mΩ, and so can deliver all of the voltage to the loudspeaker.

5.6 Solving Resistor Problems with PSpice

Spice is a computer program which can apply a variety of mathematical techniques to model electronic circuits. It can apply a variety of different analyses to predict the behaviour of the circuit under various conditions.

The basic Spice program is driven by a text file which contains a description of the circuit and a set of commands to control the analysis and format the output. It is quite straightforward to program Spice in this manner but it is quite tedious and involves familiarity with a mass of details. Fortunately, we have a version of Spice which uses extra programs to hide these details and let us work directly with circuits and graphs.

We shall use Cadence PCB's Student Version of PSpice (formerly produced by OrCAD). This is a freely distributed, limited functionality version of a professional design suite based on Spice. It consists of a graphical editor, which allows us to draw circuits in conventional notation and to select the analyses to perform, the underlying Spice engine, an application to examine the outputs of the simulations, and some hidden applications to control the whole process.

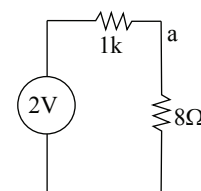


Figure 5-16 Connecting a Loudspeaker

Fully functional commercial versions of Spice are available from a number of vendors but they tend to be quite expensive (thousands of dollars). They are usually produced as part of a complete suite of program for designing and simulating circuits and then producing printed circuit board designs. These commercial versions are capable of remarkable feats, such as simulating complete computer processors at the transistor level. There are also a number of other free versions available for various platforms. These tend to be less elegant than Student PSpice (some require you to write the low-level text descriptions of the circuits) and vary in their functionality. Student PSpice is limited to simulating circuits with no more than 64 nodes and does not have the fancy features for building models of components that the full version has, but it is unlikely that you will run up against any of these limitations in this course.

5.6.1 A Tutorial Introduction using Schematic

Let us follow through the process of finding the unknown voltages and currents in a simple resistor/battery circuit. We shall use a circuit that is simple enough that we can solve it without any help from PSpice. That way we can make sure that the program gets the answer that we expect. For the tutorial, we shall use the circuit of Figure 5-17; a simple resistive divider. Obviously, we expect to find 5V at the junction of the resistors R1 and R2.

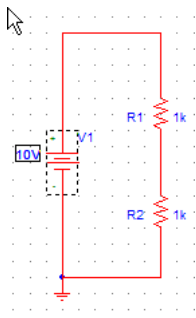


Figure 5-17 Divider in PSpice

1) Starting PSpice.

There are two ways to do this. The simplest is to click on the desktop shortcut, if it is there. Otherwise, you need to go to the Start menu (at the bottom left of the Windows screen) and bring up the complete All Programs list. You should then see an entry for PSpice Student. Resting the cursor on this will bring up a list of programs from which you should choose the Schematic application. Which ever method you use, you should end up with a screen that looks something like this.

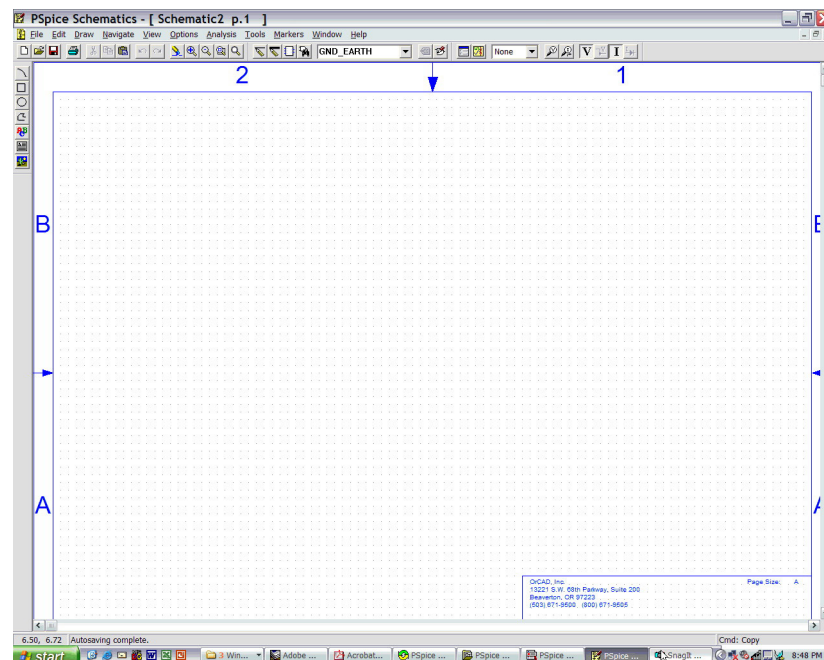


Figure 5-18 Bare Schematic in PSpice

Most of the screen is now taken up by the Schematic window with its menu and toolbars. Across the bottom of the screen the Windows Taskbar shows four entries for PSpice rather than the one that you might expect. This is because PSpice runs as several interacting applications but we normally only need to look at two of them, the Schematic program that is front-most and the Probe program that we shall use to look at the output from advanced simulations. We can ignore the extra programs.

2) Getting the Components

Start the process either by selecting Get New Part from the Draw menu or by pressing the  button on the toolbar. This will bring up the Part Browser window as shown on the right.

The main item of interest is the large scrolling list on the left-hand side. In here are all the components that Student PSpice knows about. Some of these are quite obvious, like +5V, some will become familiar during the course, like the7400, and some are extremely obscure, like DBreakCR. We shall need three components, ground, a DC power source, and a resistor.

Scroll down into the G's and you will find GND_EARTH and GND_ANALOG. These are both 0V terminals but it can be useful in some more advanced circuits to separate power supply grounds from those carrying only tiny signal currents. We only need the common GND_EARTH. Select the item and click on the Place button. Now when you move the cursor off the Part Browser you will see it turn into an arrow with a little ground symbol hanging from it. Click the left mouse button to drop the ground onto the sheet somewhere. Don't worry much where it goes as we can move it round later. You should now see something like this (Figure 5-20).

Click the Right mouse button to tell Schematic that you are done putting ground symbols on the page. If you accidentally click the left mouse button then you will leave extra grounds on the page. Again, don't worry because it will be easy to remove them later.

Now go back to the Part Browser and scroll down to the bottom to find the VDC component and add it to the schematic. Remember to use the right mouse button to tell Schematic that you are done putting batteries on the page. Finally, go and get the r component and add two resistors to the diagram then go back and Close the Part Browser. You should end up with something that looks rather like Figure 5-21. At this point we have a 0V battery rather than the 10V battery that we want. We shall soon fix that.

Note The most recently used components are also stored in a list in the toolbar. If you look back at Figure 5-18 you will see GND_EARTH in the box. If you click on the arrow to open up the list you will see all the components that you have just used ready for re-use. If you wanted to add another resistor you could get it from this list rather than going back to the Part Browser.

3) Arranging the Parts

Now that the parts are on the sheet you can move them, rotate them, delete them, and alter them to suit your purpose. We need at least to rotate the resistors and to set the component values. First select a part by clicking on (or very near) it with the mouse. The item(s) that you have selected will be shown in red. Once an item has been selected the cursor will change to have a four headed arrow on it when it is near the selected item. When this is showing you can left click on the item and drag it round the schematic. The part will move to the new position when you release the left mouse button. Try this. Drag several parts around to get them into place for the final circuit.

When an item is selected you can go to the Edit menu and select Rotate (or press Ctrl+R) to rotate the item 90 degrees. This can be repeated until you have the part in the orientation that you want. Do this to the two resistors to place them upright. You should end up with a schematic that looks something like Figure 5-23.

Note that if you accidentally left any extra parts on the schematic then you can delete them by first selecting them (as above) and then pressing the Delete key.

4) Setting Part Values

As you have seen, any time you add a resistor to a sheet you get a 1k resistor and any time you add a VDC you get one for 0V. This is not usually what you want so you next have to alter the values. There are actually quite a lot of values associated with a part. You can see and edit them all by double-clicking on the part symbol. For example, here is the list for a resistor.

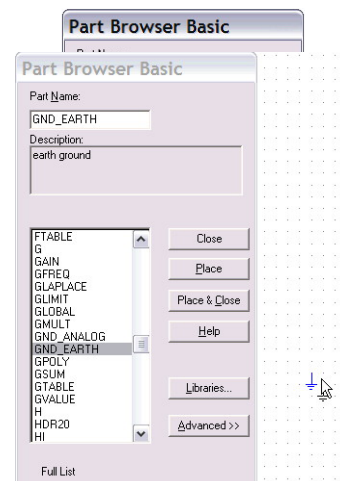


Figure 5-20

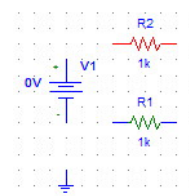


Figure 5-21 Raw Parts

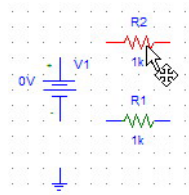


Figure 5-22 Moving a Part

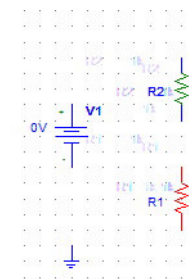


Figure 5-23 Positioned Parts

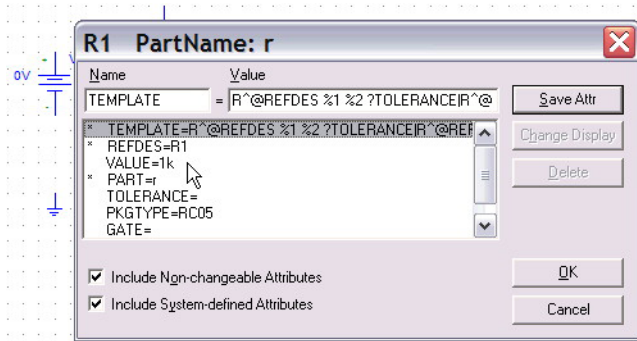


Figure 5-24 Resistor Attributes Dialog

The only parameter we might usually change is the Value, which we can alter by selecting that line and altering the text of the value.

With resistors and power supplies we usually only want to alter the value so PSpice make this easy. Instead of double-clicking on the part symbol, double click on the text of the value. If you have troubling clicking in the right place then slow the process down. First click on the value that you want to change and you will see both the value and symbol outlined (Figure 5-25). Now you can double-click in the box round the value and be pretty sure to get the alter value dialog, which looks like Figure 5-26.

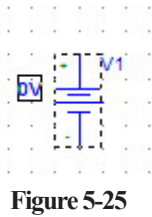


Figure 5-25

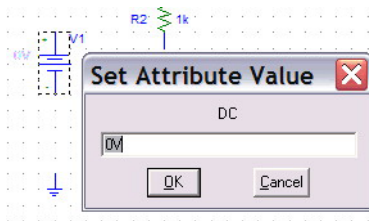



Figure 5-26 Set Value Dialog

Type in a new value and press OK. This will set the new value and the schematic will update.

5) Wiring the Circuit

Next we have to wire the circuit together. You do this with the Wiring Tool  from the toolbar. BEWARE, there are two similar icons side-by-side. You want the leftmost one. Leaving the cursor over it for a few seconds will cause the tool-tip to pop up under the button saying *Draw Wire*.

Note PSpice knows the correct units for each parameter and you may either include them or leave them off, so most people leave them off.

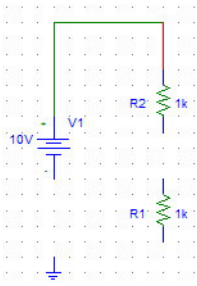


Figure 5-27

Once you have clicked on this tool the cursor will change into a pencil and you can now draw wires on the sheet. Every component has terminal points that automatically attach themselves to wires when the wiring tool is clicked near them. For the power supply or a resistor these are the ends of the little wire stubs that stick out of the top and bottom of the symbol. Click on the top end of the battery and drag a wire to the top of the upper resistor. You can control where the bends appear in the wire by clicking on the bare sheet at any time and then dragging the wire off in a different direction. Finish the wire off by clicking on the top terminal of the resistor. Try to make your wire look something like Figure 5-27.

Repeat this process to add wires from the lower battery terminal to the ground symbol, from ground to the bottom of the second resistor, and the wire between the resistors. When you are finished, right-click the mouse to put away the wiring tool and return to the pointer cursor.

Note It is possible to move components and even wires around once the circuit has been wired. You do the same as before. Click in the item that you want to move and then click-drag it with the cursor.

At the end you should have a circuit that looks like Figure 5-28.

6) Solving the Circuit

Now that the circuit is complete, we want to run Spice to tell us the voltages and currents. Before we can do this we must save the circuit. We do this in the usual way by going to the File menu and selecting Save. Give the circuit a name that you will recognize, such as Tutorial1.sch or VoltageDivider.sch.

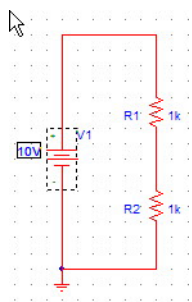


Figure 5-28

Spice is capable of performing a wide range of different kinds of analysis, most of which are not really applicable to a circuit this simple. We want only the most basic, called an Operating Point analysis. This analysis is required to be run before any of the more elaborate options and so if we don't do anything else PSpice will just run this for us. To start the analysis, go to the Analysis menu and select Simulate. This will cause Schematic to run the Spice program in the background and when it is done to pop up an output window like this.

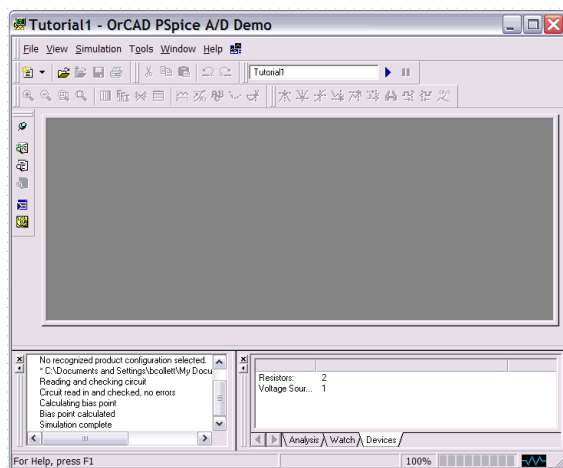


Figure 5-28 PSpice Output Window

There is not much to see here because our analysis did not produce any graphical output so we can just close this box.

To see our results we need to tell schematic that we want to see the analysis output. Go to the Analysis menu again and this time put the cursor on Display Results On Schematic. This will pop up a secondary menu where you need to select Enable, Enable Voltage Display, and Enable Current Display. You will have to do this in several operations because every time you click on a menu item the menu goes away. When you have finished your schematic should have the results written on top of it (Figure 5-29).

Conclusion

You have now learnt how to draw circuits using the Schematic part of PSpice and how to simulate them so that Spice performs the node analysis rather than making you do it. For simple circuits such as this one the task is a lot more work than doing the job yourself. However, it is little more work to set up a much more complicated circuit and then solving the circuit is trivial instead of a LOT of algebra. For example, Figure 5-30 is the output of the analysis of a far more complex problem, a cage of twelve 1k resistors arranged along the sides of a cube and fed from a battery connected at opposite corners.

This can be done by hand but it is very fiddly, with 6 unknown nodes.

Summary

Formal analysis methods turn the task of solving for the voltages and currents in a circuit into a straightforward mathematical problem. These methods are based on Kirchhoff’s circuit laws.

Remember Kirchhoff’s Laws

- 1) The sum of all the currents leaving a junction in a circuit must be equal to the sum of all the currents entering that junction.
- 2) Around any closed loop in a circuit, the algebraic sum of the voltage drops must equal zero.

We have explored two methods, the Node method and the Loop method. In general the loop method is more straightforward and works well for all circuits that are simple enough to be solved by hand.

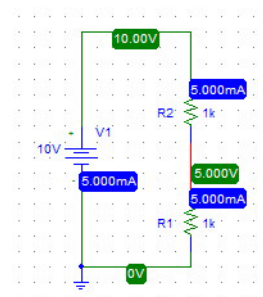


Figure 5-29 Solved Divider

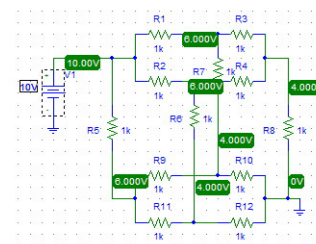


Figure 5-30 9 Resistor Cube Solved

Note I turned off the current display because it was making the picture a bit too confusing!

Remember Nodal Analysis

Here are the rules for nodal analysis.

- 1) Label all the nodes in the circuit and identify the unknown nodes.
- 2) Assign currents to each battery and resistor in the circuit.
- 3) Write down Kirchhoff's Current Law for each unknown node.
- 4) Write down Ohm's law for each resistor in the circuit.
- 5) Substitute the currents from step 4 into the equations of step 3.
- 6) Solve the resulting equations for the unknown node voltages.
- 7) Substitute the node voltages into the equations from step 4 to find the current in each resistor.

Thévenin's Theory states that

Any combination of batteries and resistors inside a two-terminal circuit can be replaced by the Thévenin equivalent circuit of a voltage source V_{Th} series with a resistance R_{Th} .

We can calculate these values by:-

1. Calculate (or measure) the open circuit voltage, V_{OC}
2. Calculate (or measure) the short circuit current, I_{SC}
3. Compute V_{Th} and R_{Th} from

$$V_{Th} = V_{OC} \text{ and } R_{Th} = \frac{V_{Th}}{I_{SC}}$$

One common application of Thévenin's Theory is the rule for connecting two circuit together

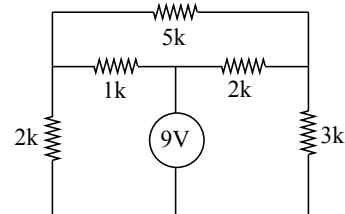
Remember A rule for connecting two circuits.

To transfer a voltage from one circuit to another, the input resistance of the destination should be at least ten times the output resistance of the source.

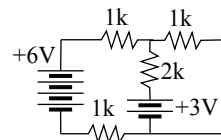
Once a circuit has more than three or four unknown nodes it is time to switch to a circuit simulator such as PSpice.

Exercises

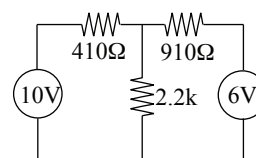
1. Label each distinct node in the circuit on the right.
2. Find all unknown currents and voltages in the circuit on the right.



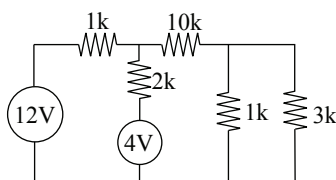
3. Label each distinct node in the circuit on the right.
4. Find all unknown currents and voltages in the circuit on the right.



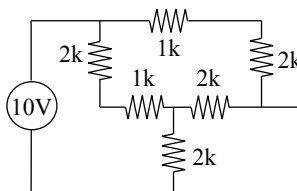
5. Use one of the formal analysis methods to find the unknown currents and voltage in this circuit. (Note that this is one where you will have to use a calculator and can't do simple tricks like finding common denominators.)



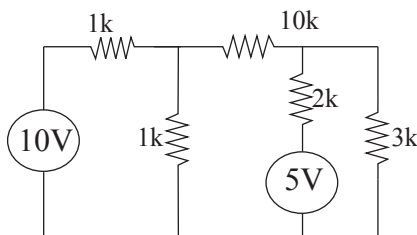
6. Calculate the voltages and currents in this circuit. Please give your answers to 3 significant figures.



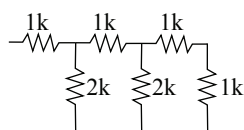
7. Find all the currents and voltages in the circuit on the right.



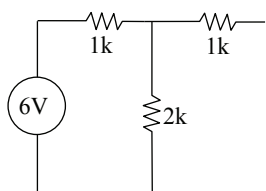
8. Calculate the voltages and currents in the circuit on the right. Please give your answers to 3 significant figures.



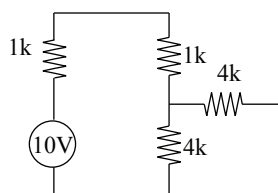
9. The circuit below is equivalent to a single resistor. Calculate the value of that resistor.



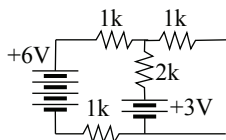
10. Find the Thévenin equivalent of the following 2-terminal circuit



11. Find the Thévenin equivalent of the circuit on the right.



12. Find the Thévenin equivalent of the circuit on the right. Note that you will have to be careful with the 3V battery. Its negative terminal is NOT at the same voltage as the negative terminal of the 6V battery. You will have to treat it in the same way as the Thévenin example in section Thévenin's Theory



13. An AA sized alkaline-manganese battery has a Thévenin voltage of 1.5V and an internal resistance of 0.4Ω. What is the largest current that can be drawn from it if the terminal voltage is not to drop below 1V?

14. How much power can the battery in question 12 deliver to a load of resistance R? What value of resistance R gives the largest power delivered to the load? What is that power?

Chapter 6: Time Varying Voltages

6.1 Introduction

Fixed voltages are all very well for shining lights or powering motors but most of electronics is about voltages and currents that change in time. We need to use electronic voltages to represent and process real world quantities that vary in time. Some examples are the small variations in air pressure that make up the sound waves that we hear and the variations in brightness and color that make up the television pictures that we see. Thus, we spend most of our time working with time varying voltages of one kind or another.

We write a time varying voltage or current as a function of the time thus

$$V = V(t) \quad \text{or} \quad I = I(t)$$

and we draw pictures of the voltage as a function of time to help us understand what is going on inside a circuit. For example, Figure 6-1 is a graph of the voltage in an American wall socket.

Time varying voltages come in two basic kinds, periodic voltages and non-periodic voltages. A periodic voltage is one that has a shape that repeats after a fixed interval of time and which goes on repeating that same shape

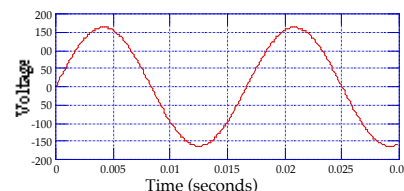


Figure 6-1 American Line Voltage

6.2 Periodic Voltages

Because we shall work so much with periodic voltages, we have a large body of terms that we use to describe them as shown in Figure 6-2

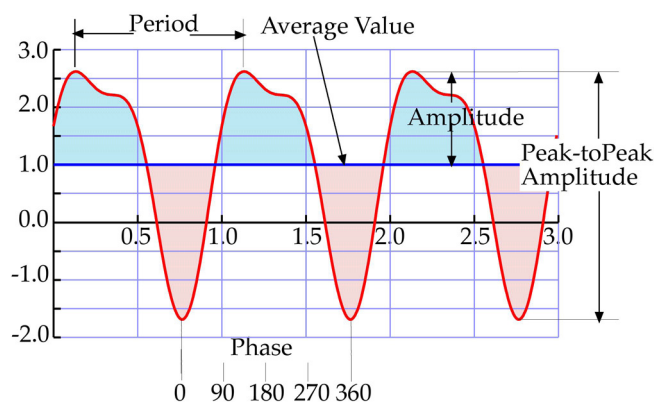


Figure 6-2 Terminology for Periodic Voltages

Period

The **Period** is the time between two points at which the voltage is behaving in exactly the same way, usually the two points at the same voltage and with the same slope. The units of period are seconds and we use symbols such as T and τ to represent period.

Frequency

The **Frequency** is the number of cycles through the which the voltage passes in 1 second. Thus frequency is the inverse of period so that we have the relations

$$f = \frac{1}{T} \quad \text{and} \quad T = \frac{1}{f}$$

where f is the usual symbol for frequency. The units of frequency are cycles-per-second or **Hertz**.

$$1 \text{ Hertz} = 1 \text{ Hz} = 1 \text{ cycle-per-second}$$

Example

If the period of a voltage is 0.1s then in 1 second $1/0.1=10$ periods will have passed so the frequency is 10 Hz or 10 cycles per second.

Average Value

The **Average Value** is roughly that voltage above which the signal spends half its time and below which it spends half its time. Strictly, the average voltage is a level chosen so that the **area** on the voltage-time graph that lies above the average voltage line and below the signal line is the same as the area that lies below the average voltage line and above the signal line. This is shown in Figure 6-2 by the shaded regions above and below the average voltage line—each positive blue area is the same size as each negative pink area.

Amplitude

Note The symbols for amplitude are the same as those for the Volt and Amp. This can occasionally cause confusion and we sometimes use subscripts such as A_0 to distinguish the amplitude from the unit.

The **Amplitude** is the maximum amount by which the signal deviates from the average value. For many signals the amplitude above and below the average voltage lines is the same. The units of amplitude are Volts (or Amps if we are describing a time varying current) and the common symbols are V or A.

Peak-to-Peak Amplitude

Note We often abbreviate peak-to-peak as pk-pk or p-p.

The **Peak-to-peak amplitude** is the total range over which the signal varies, measured from the lowest or most negative point on the signal to the highest or most positive point on the signal. For many common signals this is 2x the amplitude. Again, the units are the same as the quantity being measured so we speak of Peak-to-Peak Voltage (symbol V_{pp}) or Peak-to-Peak Current (symbol I_{pp}).

Root Mean Square Amplitude

The **RMS (Root Mean Square) amplitude** is a separate measure of the magnitude of a time varying voltage. It is that DC voltage which would produce the same average heating of a resistor as the alternating voltage. Each different shape of voltage has a different relationship between the RMS amplitude and the peak-to-peak amplitude.

Phase

Phase is a convenient way to refer to the progress of the voltage through a single cycle. A single period of the wave is equated to a complete circle, 360° or 2π radians. Fractions of a period are then given as smaller angles. For example, a quarter of a wave is often referred to as 90° or $\pi/2$ radians.

Signals with different frequencies have different periods so that a given time interval, e.g. 0.01s, will take through different amounts of different waves. However, phase is a universal quantity: 180° of phase takes you through 1/2 cycle of any wave, whatever its period or shape.

Example

Two waves have the same period, 1.2mS, but they are offset one from another by 0.26mS. We can find the phase difference between the waves by noting that $360^\circ \sim 1.2\text{mS}$ so that the phase difference corresponding to 0.26mS must be

$$\phi = 0.26\text{mS} \times \frac{360^\circ}{1.2\text{mS}} = 78^\circ$$

6.2.1 Some Common Periodic Shapes

Here are several common shapes of periodic function that we use to stimulate or investigate electrical circuits. They are shown with their usual names—more words to learn! Note that a particular shape of periodic signal is often called a wave so that a sinusoidal voltage shape is often called a sine wave, and a sawtooth function shape called a sawtooth wave.

Figure 6-3 shows a **sawtooth** wave with an amplitude of 2V peak-to-peak, an average voltage or **offset** of 1V, a period of 0.1mS, and a frequency of $1/0.0001 = 10\text{kHz}$.

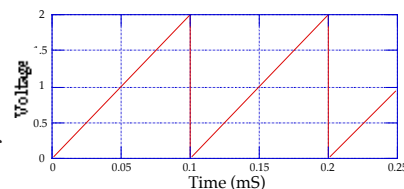


Figure 6-3 Sawtooth Wave

Figure 6-4 shows a **triangle** wave with an amplitude of 2V, an offset of 0V, a period of 3.2mS, and a frequency of 312.5Hz. Note that a triangle wave has, normally, symmetric regions of positive and negative slope while a sawtooth has one gentle slope and one very steep slope.

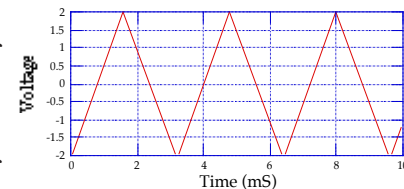


Figure 6-4 Triangle Wave

Figure 6-5 shows a **square** wave, called a 5V square wave, running from 0 to 5V and so having a peak-to-peak amplitude of 5V. The conventional amplitude, rarely quoted for the square wave, is 2.5V, as is the offset or average voltage. The period is $0.38\mu\text{s}$ giving it a frequency of 2,631,579Hz or 2.63MHz (to 3 s. f.).

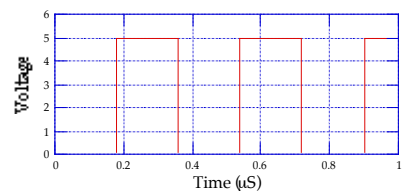


Figure 6-5 Square Wave

The most useful periodic waveform for analyzing the behavior of most electronic circuits is the **sinusoid** or **sine wave**. Figure 6-6 shows one with an amplitude of 1.2V and a period of 0.333mS (corresponding to a frequency of 3kHz). The offset, or average, voltage is 0.

The reason that sinusoids are so popular is twofold. First, the world abounds in sinusoidal waves; speech, music, and radio waves are all full of sinusoids. Second, mathematics tells us that if we know how a circuit responds to sinewaves then we can extrapolate how the circuit will respond to any input.

Info Sinusoid

The term **sinusoid** is used for all the variations of a sinewave. A real sine wave passes through 0 at phase = 0° while a sinusoid may pass through zero at any phase angle. So $\cos(t)$ and $\sin(\phi + t)$ are both sinusoids.

6.3 Sinewaves

We will use sinewaves in almost all of our analog circuits so we shall study them in a little more detail. Here is the mathematical description of a sine wave

$$V(t) = V_0 \times \sin(2\pi f \times t + \phi)$$

where V_0 is the amplitude, f is the frequency, and ϕ is the phase of the wave. We often use an abbreviation ω for the whole term $2\pi f$ that occurs in all sine wave systems. This is called the **angular frequency**.

$$\omega = 2 \times \pi \times f$$

The phase ϕ tells us where the wave is in the cycle at the moment that we call time $t = 0$. The phase is an angle that must be given in radians if we are doing math with the expression but which we often speak of in degrees. Let's look at some different sinewaves at different phases to get a feel for the phase

Figure 6-7 shows two sinewaves; the solid one is at phase 0 and the dashed one at phase 180° (π radians).

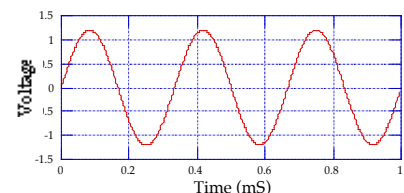


Figure 6-6 Sine Wave

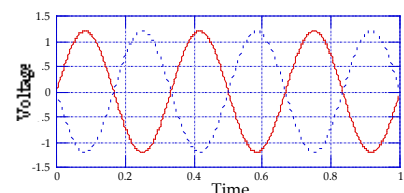


Figure 6-7 180° Phase Difference

Next Figure 6-8 shows two sinewaves where the dashed one has a phase of +90° ($\pi/2$ radians). We call a positive phase difference a leading phase so that here the dashed line **leads** the solid one by 90°.

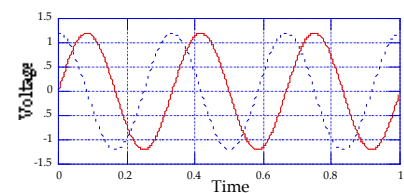


Figure 6-8 +90° Phase Difference

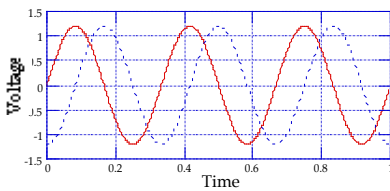


Figure 6-9 -90° Phase Difference

To make sure that we know what the sign does, Figure 6-9 shows a pair in which the dashed one has a phase of -90° . We say that the dashed curve **lags** by 90° . People often find this relationship confusing because the lagging wave appears to the right of the main wave and so looks as though it is ahead. The trick is to remember that time increases **to the right**. If you think of the curves as waves then they are traveling to the **left** and so the solid one is in the lead and the dotted one lags behind.

6.3.1 RMS Measurements and Sine Waves

It is a very common practice to specify the sinewave output of power sources in terms of the RMS amplitude rather than specifying the amplitude because the RMS voltage makes it easy to calculate the power delivered by the source as

$$P = \frac{V_{RMS}^2}{R}$$

The RMS amplitude is a kind of average and, as thus is lower than the peak voltage. For a sine wave, the RMS voltage is given by

$$V_{RMS} = \frac{\text{Amplitude}}{\sqrt{2}} \quad \text{OR} \quad \text{Amplitude} = V_{RMS} \times \sqrt{2}$$

Example

The voltage that comes out of a standard American wall socket is $115V_{RMS}$ so the amplitude = $115 \times \sqrt{2} = 163V$. Thus the peak voltage leaving a wall socket is 163V not 115V.

Note The relationship between the amplitude of a signal and its RMS voltage is complex one. It depends on the shape of the wave and upon both the maximum and minimum voltages. It can be found by advanced mathematical methods (integration) for any waveform but the process is too complex to cover here. The key thing to remember is that the factor of $\sqrt{2}$ only works for pure sine waves with no DC component, so that the average voltage is 0.

Note Although the names come from descriptions of currents, they are actually used for all sorts of other quantities. For example, it is perfectly possible to refer to the AC resistance of a component and mean the resistance to time varying currents, which may not be the same as its DC resistance.

6.4 A little helpful notation

We call a steady, time invariant quantity a **DC quantity** and a time varying quantity, particularly one that has both a positive and a negative part, an **AC quantity**. These terms stand for

DC *Direct Current* *current that flows in one direction*
 AC *Alternating Current* *current that flows in both directions*

When we have signals that are basically time varying but which have an average value that is not near to zero, it is often useful to split the voltage up into a time varying part and a time independent offset. The offset is equal to the average voltage. For example, the sinewave signal in Figure 6-10 results from squaring a 3kHz sinewave (we shall meet circuits to do this later).

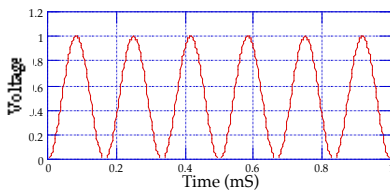


Figure 6-10 3kHz Sine Wave, Squared

This has an average voltage of 0.5V so we can split the voltage into two parts, a steady offset of 0.5V and a 2kHz sinusoid with an amplitude of 0.5V. That is we can write

$$\sin^2(2000\pi \cdot t) = 0.5 - \cos(4000\pi \cdot t)$$

where we use a negative cosine to get the phase of the wave correct. There is a common notation for doing this sort of split. We write

$$V(t) = V + v(t)$$

where the steady voltage is written in capitals and the time varying part is written in small letters. By convention, any quantity written in small letters is the time varying part of a quantity that may or may not also have a steady DC part. It is called a **small signal** quantity. We call the full quantity, DC part and AC part together, a **large signal** quantity and write it with capital letters.

6.5 Time varying voltages and resistor circuits

Because resistors have a straight line I-V curves, resistive circuits have the property that all the voltages and currents in the circuit are proportional to the driving voltage and thus have the same shape. For example, if we replace the battery in this divider circuit, taken from chapter 2, with a signal generator then we get the circuit of Figure 6-11.

Now, we found that, so long as no current is drawn from the output, we have

$$V_{bc} = \frac{R2}{R1 + R2} \times V_{ac}$$

That means that if we put a sinewave in, as shown, then we get a sinewave out. Moreover, the sinewave coming out is exactly in phase with the sinewave going in. If we plot the input and output voltages on the same graph then we get something like Figure 6-12.

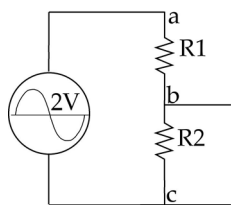


Figure 6-11 Voltage Divider

Notice that the dashed output wave has a 0° phase difference from the solid input wave. We say that the input and output are **in phase**. No matter what shape the input voltage is, the output voltage is exactly the same shape as the input; it differs only in size. This is a very important property of a circuit called **LINEARITY**. It means, roughly, that the circuit does not alter or distort a signal as it passes through it. For example, linearity is one of the most important properties of an audio amplifier. A sound that passes through an audio amplifier must do so without altering its shape or the sound that comes out will be different from the sound that went in—obviously not a desirable property

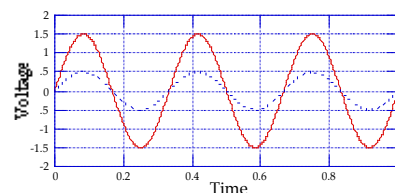


Figure 6-12

Note Linearity

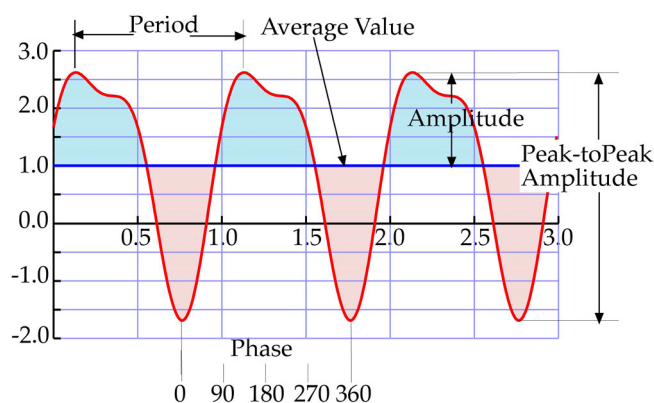
A more formal definition of linearity states that the output from the device is a linear function of the input. This allows more general relationships because the operations of differentiation and integration are also linear functions.

In general if you put a sine wave into a linear circuit then output will be a sine wave at the same frequency. The size and phase may be changed but the shape and frequency will still be the same. This means that non-sinusoidal signals may have their shapes changed.

6.6 Non-Sinusoidal Voltages

Summary

We describe periodic time varying voltages by their period, amplitude, frequency, average value, and phase.



The frequency f and the period T are related by

$$f = \frac{1}{T}$$

To find the phase difference between two waves we treat a single period as a complete circle of 360° and then express the time difference between two waves as a fraction of that period using the formula

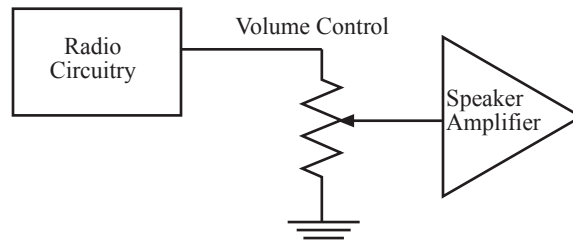
$$\text{Phase} = \text{delay} \times \frac{360^\circ}{\text{Period}}$$

The Root Mean Square Amplitude (RMS amplitude) of a sinewave is that DC voltage which dissipates the same power in a resistor as the sinewave does. The RMS amplitude, V_{RMS} , is related to the amplitude by

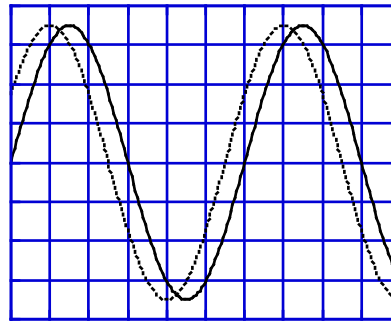
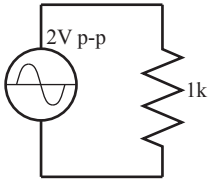
$$V_{RMS} = \frac{\text{Amplitude}}{\sqrt{2}}$$

Exercises

1. Find the duration of one cycle of an FM radio signal at 89.5MHz.
2. The power in an American wall socket alternates at a frequency of 60Hz. What is its angular frequency?
3. What is the largest peak-to-peak voltage swing that you could encounter from an English wall socket that delivers 240V R.M.S. power?
4. A variable resistor is often used as a volume control using a circuit similar to the figure below. Explain how this circuit works and state what property of a circuit made up of resistors allows the volume control to alter the loudness of the signal without altering what it sounds like.



5. A 1kHz sinewave is connected to an 8Ω loudspeaker. If the R.M.S. average power delivered is 20W, find the peak voltage (the amplitude) of the wave and the peak power delivered to the loudspeaker.
6. 3-Phase AC power is delivered with a 3-wire system rather than the 4-wire system of standard single-phase AC. The signals in the three wires are 120° apart in phase. Draw a diagram showing the three signals in a 115V R.M.S. 3-Phase AC supply.
7. Find the peak and average power dissipated in the 1k resistor in the figure on the right. (Hint: read the part about R.M.S. voltage).
8. The figure below shows a trace as seen on an oscilloscope with a vertical scale of 0.1V/div and a horizontal scale of 2mS/div. Using that figure find the amplitude, RMS voltage, period, and frequency of the solid wave. Estimate the phase difference between the solid wave and the dotted wave. Which wave leads?



9. A 1200W hair dryer is connected to the $115\text{ V}_{\text{RMS}}$ 60Hz AC supply coming from a standard wall socket. Draw accurate figures showing the voltage across the hair dryer's resistance and the current flowing through that resistance. Assume that at time $t=0$ the wall voltage is zero.

Chapter 7: The Capacitor

7.1 Introduction

The next component that we meet is one that is only useful in circuits with time-varying voltages. In a circuit driven by a fixed voltage source, a capacitor would have absolutely no effect on the circuit, as we shall see.

A capacitor consists of two sheets of metal that are electrically isolated from each other. In its simplest form, a capacitor is made from two flat metal plates with a gap between them and a wire leading to each plate (Figure 7-1). More elaborate versions put an insulating material, such as paper or plastic, between the plates and may roll the plates up but the principle is the same.

Since there is no electrical connection between the plates, current cannot flow from one plate to the other and the component seems to be useless. However, if we put some charge on one of the plates (connect it to a battery) then the charges in that plate exert forces on the charges in the other plate, pulling them up into the plate, and we get a voltage difference between the plates. Experimentally, we find that if we put a positive charge on one plate then we get an equal and opposite negative charge on the other plate and that the magnitude of the charges, Q , is proportional to the voltage, V , across the capacitor

$$Q \propto V \text{ or } Q = C \times V$$

where C is a quantity called the **Capacitance** of the component. The bigger the capacitance, the more charge it takes to get a certain voltage drop across the component. The unit of capacitance is the **Farad**. If you have 1 Coulomb of charge on each plate with 1 Volt across the capacitor then you have a capacitance of 1 Farad.

Now, although a steady current cannot flow through the capacitor, a temporary or transient current can flow. Current can flow into one plate of the capacitor and out of the other for a short while, increasing the charge on the plates and the voltage drop across them, so long as it later flows out of that plate and back into the other, restoring the voltage to zero. Very roughly, a capacitor blocks steady currents but allows varying currents to pass through it.

7.1.1 The water model of a capacitor

There is a simple plumbing analogy for this component—a chamber with a rubber membrane across the middle (Figure 7-2). As with the electrical component, no steady current can flow from one pipe to the other because of the membrane but we can push water into the left hand end and it will push water out the other side. A positive amount of water in on the left will produce an amount out, a negative amount in, on the right. The amount of water in on the left is the same as the amount pushed out on the right and the pressure drop in the rubber membrane is proportional to the amount of water pushed in and out. The analogy of the capacitance is the volume of the chamber on each side of the membrane—the bigger the chamber, the more water it takes to push the membrane a given amount.

7.1.2 Driving a capacitor with alternating current

When we apply an alternating current to a capacitor, the frequency of the alternation as well as the size of the current affects the voltage that is generated. A small current at a low frequency will produce a much larger voltage swing than the same current flowing at a higher frequency. The low frequency current spends more time flowing into the capacitor and so deposits more charge on the plates, yielding a higher voltage. In terms of our analogy, a slowly alternating current flow has time to push much more water in to the chamber and so move the membrane much further, producing a much larger pressure change. A rapid alternation of the current will

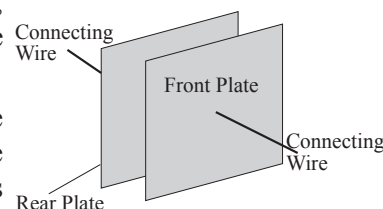


Figure 7-1 Capacitor

Info We meet a wide range of capacitor values in various kinds of circuits, but most of them are very much less than 1 Farad. In high frequency circuits, circuits operating at tens of MHz or higher, most of the capacitors we meet are only a few picofarads ($1\text{pF} = 10^{-12}\text{F}$). Lower frequency circuits and digital logic circuits are full of capacitors in the μF range while power supplies use the largest values of capacitor, anything from about $10\mu\text{F}$ up to tens of thousands of μF . The values up to about $1\mu\text{F}$ tend to be quite small, only a little larger than a standard $1/4\text{W}$ resistor, while the large values used in power supplies come all the way up to fist sized.

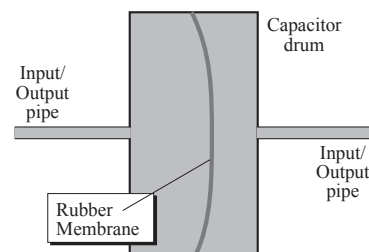


Figure 7-2 Plumbing Model of a Capacitor

not allow time for much water to flow into and out of the chamber and so the pressure fluctuations will be small.

7.2 Resistor-capacitor circuits

The simplest interesting capacitor circuit has a resistor in series with a capacitor driven by a battery and switch (Figure 7-3). Initially, there is no charge on the capacitor so there is no voltage across it. We will follow what happens when the switch is flipped to connect the battery.

The moment after the switch is flipped, the voltage at the top of the resistor, at point a, is at the full battery voltage. At the same moment, the voltage at the bottom of the resistor is 0V because there is no voltage across the capacitor. That means that a current $I = V_0/R$ flows into the top resistor and out the bottom into the capacitor. As the current flows, the charge on the top plate of the capacitor builds up and the voltage at point b rises (Figure 7-4a). The voltage rises at a rate proportional to the current flowing in the resistor.

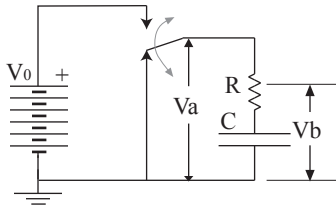


Figure 7-3 R-C Charging

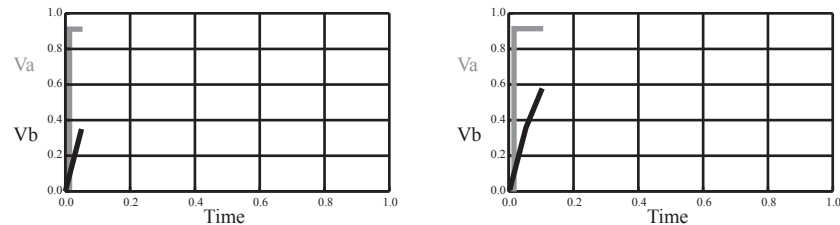


Figure 7-4 a

b

Now, as the voltage at point b rises and the voltage at point a stays constant, the voltage across the resistor falls and the current flowing falls. With less charge arriving in each second, the rate at which charge builds up on the capacitor is less. Thus the voltage at point b builds up more slowly and the slope of the voltage/time curve gets lower (Figure 7-4b).

Time passes and the voltage goes on increasing and the current goes on falling so that the slope of the V/t curve keeps getting lower and lower. The voltage at point b gets closer and closer to the voltage at point a but it never quite equals it (Figure 7-5a).

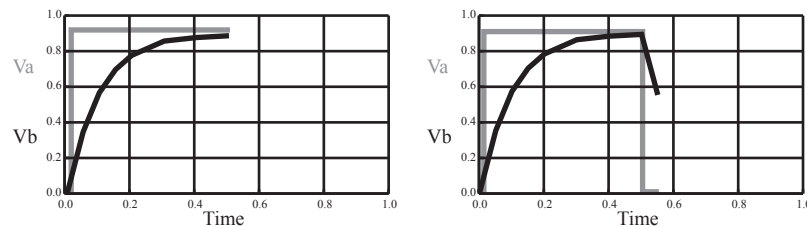


Figure 7-5 a

b

Next we flip the switch again so the voltage at point a drops to zero. Now point b is still up at nearly the battery voltage so that current flows in the resistor in the opposite direction, draining charge out of the resistor (Figure 7-5b). As the charge drains out, the voltage across the capacitor falls. Since the current at this time is practically the same as the starting current when we first flipped the switch, the rate at which the voltage falls is just the same as the rate at which it first rose. Thus the downwards slope of the curve at the start of the discharge phase just matches the upwards slope at the beginning. Again, as time passes and the voltage across the capacitor falls, so the current in the resistor falls and the slope falls.

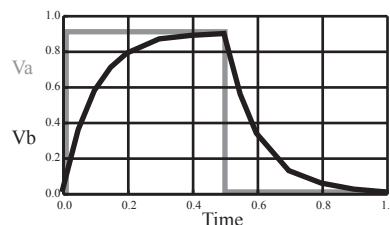


Figure 7-6

The V/t curve steadily flattens out as the voltage falls toward zero but never quite gets there (Figure 7-6).

We can look at the process in our plumbing system to get further insight into what is happening. Here is the plumbing equivalent of the circuit. Initially the valve is set so that the capacitor is connected to itself in the circle through the resistor pipe. Thus there is no pressure difference across the membrane and the membrane is relaxed (Figure 7-7).

Now we turn the valve and watch. We have made the resistor pipe very thin so that the process will take a while and we can watch it happen. The pump starts to force water into the resistor pipe and so water flows out the other end. This water flows into the upper chamber of the capacitor and starts to increase the volume of water there so that the membrane starts to bend and the pressure across the capacitor starts to rise (Figure 7-8a).

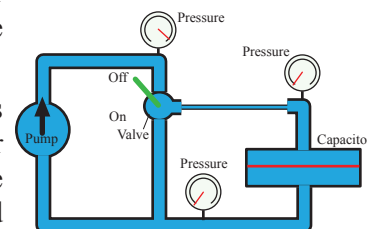


Figure 7-7

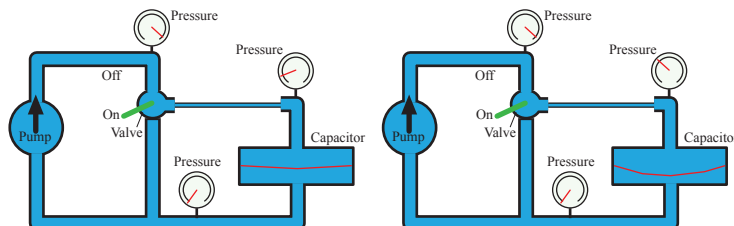


Figure 7-8 a

b

As the pressure across the capacitor rises, it pushes back on the water flowing out of the resistor, decreasing the pressure drop and thus decreasing the rate at which water flows through the resistor. The rate at which the pressure across the capacitor rises is now a little smaller and so the slope of the pressure/time curve falls (Figure 7-8b).

This process continues. The pressure across the capacitor gets larger and larger (Figure 7-9a) until there is little or no pressure drop across the resistor and the pump can no longer force water through. As in the electrical case, the current theoretically takes forever to fall to zero but it gets immeasurably small after a short while.

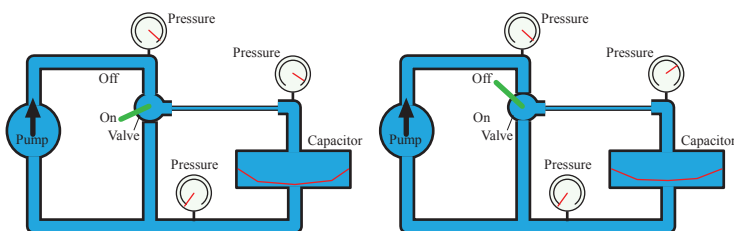


Figure 7-9 a

b

Once the current has fallen to nearly zero, we switch the valve back to the off position. Now the pump is not involved and the pressure on the valve side of the resistor falls to zero while the membrane is still pushing very hard. Water now flows out of the top chamber of the capacitor and back round to the bottom chamber, reducing the pressure drop across the capacitor (Figure 7-9b).

Time passes, water flows out of the top chamber and the pressure falls so the flow of water out and through the resistor falls. The rate at which the pressure falls gets lower and lower and the pressure/time curve gets flatter and flatter again as the pressure across the capacitor falls to zero (Figure 7-10).

From a detailed mathematical analysis (see box on the next page) we find that the actual shapes of the rising and falling voltage curves are exponential with the same time constant, $\tau = RC$. That means that it takes a time $t = RC$ for the voltage to rise to 63% of its final value or to fall by 63% of its initial value as shown in Figure 7-11.

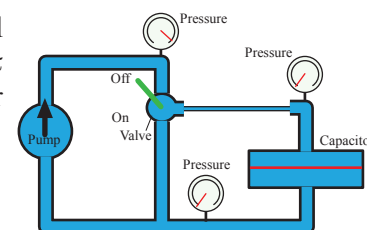


Figure 7-10

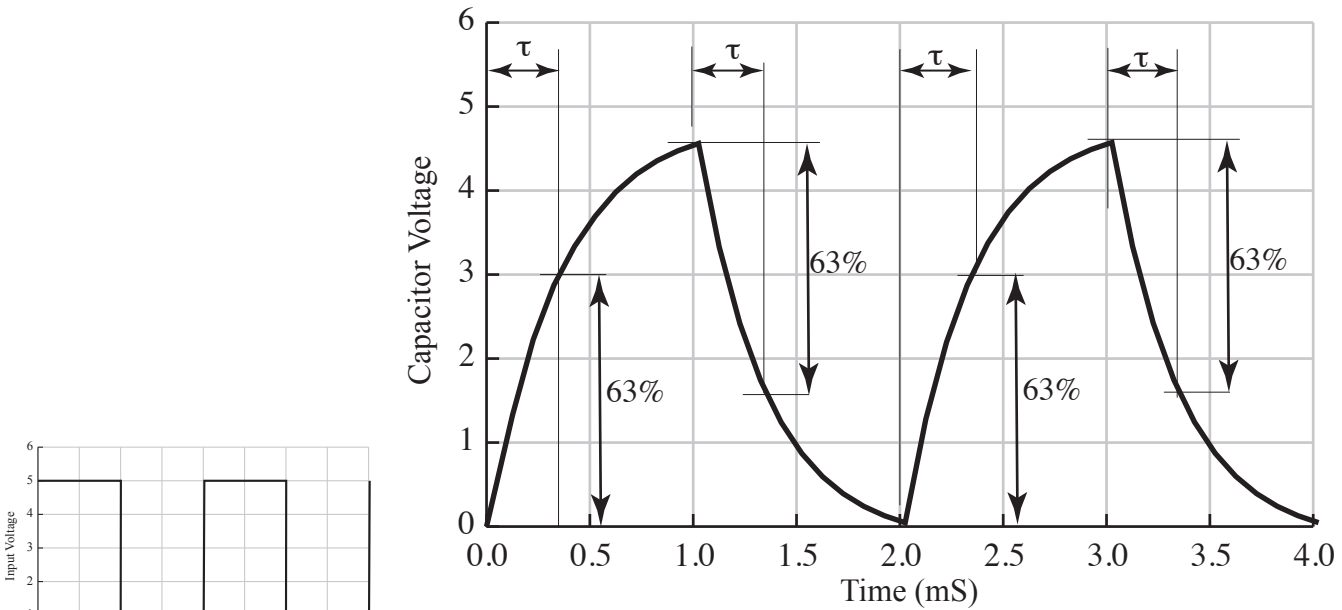


Figure 7-11 Charging and Discharging a Capacitor

If we look at the voltage at point a Figure 7-3 in we see that it executes a square wave with full amplitude V_0 as in Figure 7-12. This means that we have found out what happens to a resistor/capacitor circuit when it is driven by a square wave. Unlike the purely resistive circuits that we have studied before, the RC circuit severely distorts the square wave so that the waveform coming out of the circuit is very different from the waveform that went in.

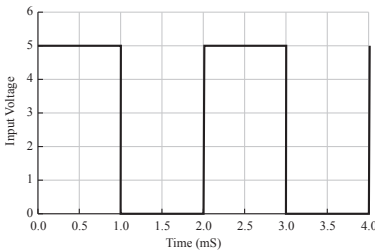


Figure 7-12 Input For RC

7.3 The Capacitor and the Sinewave

We know that when we apply a sinusoidal voltage to a resistor then the current that flows is also sinusoidal and is in phase with the driving voltage. If we have

$$V(t) = V_0 \sin(\omega \cdot t)$$

then

$$I(t) = (V_0/R) \sin(\omega \cdot t)$$

The capacitor behaves somewhat differently. The voltage across the capacitor determines the charge on the capacitor, $Q = CV$, instead of the current through the capacitor. The current is determined by the rate of change of the charge on the capacitor and so depends on the rate of change of the voltage across the capacitor.

If we look at a graph of the voltage across the capacitor driven by a sinewave then we can make a rough plot of the current through the capacitor, since the rate of change of the voltage is the **slope** of the voltage versus time plot.

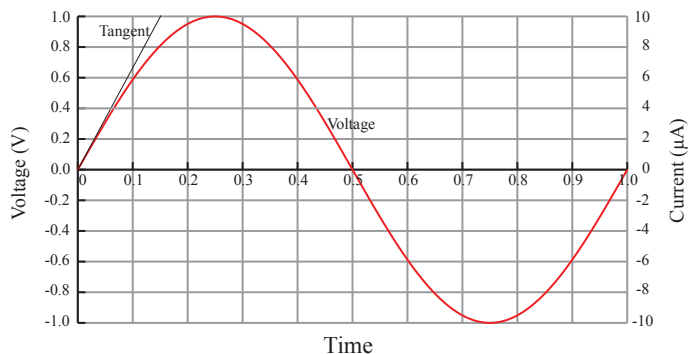


Figure 7-13

Theory of the RC circuit driven by a squarewave

Using calculus we can analyze the behavior of this system in complete detail.

First we note that, as current flows into the capacitor, the rate at which the charge changes is exactly equal to the current flowing so that charge is conserved.

$$I = \frac{dQ}{dt}$$

where I is the current flowing into the capacitor. If the current flows out then the input current is negative, so that dQ/dt is negative and the charge on the capacitor falls rather than rising.

Now, when the resistor and capacitor are connected to the battery, the voltage across the resistor is $V_a - V_b$. Then Ohm's law tells us that the current in the resistor is

$$I = \frac{V_a - V_b}{R}$$

This current flows into the capacitor so that we have

$$\frac{dQ}{dt} = \frac{V_a - V_b}{R}$$

Now we can use the capacitor rule, $Q = C \times V$, to express this in terms of voltages

$$C \frac{dV_b}{dt} = \frac{V_a - V_b}{R}$$

We can re-write this in the usual form of a first order linear differential equation, replacing V_a by V_i since the switch is in the upper position,

$$\frac{dV_b}{dt} + \frac{1}{RC} V_b = \frac{V_i}{RC}$$

which we solve in the usual fashion with the integrating factor $e^{t/RC}$.

$$d(V_b \times e^{t/RC}) = \frac{V_i}{RC} \times e^{t/RC} dt$$

The initial conditions are $t = 0$, $V_b = 0$, so that we can integrate

$$\int_0^{V_b} d(V_b \times e^{t/RC}) = \int_0^t \frac{V_i}{RC} \times e^{t/RC} dt$$

to obtain

$$V_b = V_i (1 - e^{-t/RC})$$

This is the equation that is plotted for the rising curves of Figure 7-11, where we see that the voltage takes a time $t = RC$ to rise 63% of its final value. We call this time the **time constant** of the RC circuit.

Once we turn off the switch, the current flows the other way in the resistor, driven only by the capacitor voltage so that

$$I = \frac{dQ}{dt} = C \times \frac{dV_b}{dt} = -\frac{V_b}{R}$$

We can manipulate that to get

$$\frac{dV_b}{V_b} = -\frac{dt}{RC}$$

Now, if we call the time t at which we turned off the switch $t = 0$ and then call the voltage at that instant V_i , then we find

$$V_b = V_i \times e^{-t/RC}$$

and the voltage decays exponentially with the same time constant, $\tau = RC$, at which it rose. That means that after time $t = RC$ the output voltage has fallen back to 37% of its maximum value, as shown in the falling curves of Figure 7-11.

Figure 7-13 shows the voltage across a $1\mu\text{F}$ capacitor that is driven by a 1Hz sinewave. I have drawn a line that is tangent to the sinewave at the origin. The slope of that line is about $1.2\text{V}/0.18\text{seconds} = 6.7\text{V/S}$ so the current is $I = C dV/dT = 6.7\mu\text{A}$.

Now we know that the current at $t=0$ is $I=6.7\mu\text{A}$ and I have added a cross at that value to make the plot of Figure 7-14. Note that the current scale is quite different from the voltage and is given on the right-hand side of the graph. Next we will advance the time to $t=0.1$. The new slope line, tangent at $t = 0.1$, rises only about 0.9 units in about 0.175 time units giving it a slope of $0.9/0.175=5.1$ for a current of $5.1\mu\text{A}$. Again, we can add the new cross to the current line and move the tangent line over.

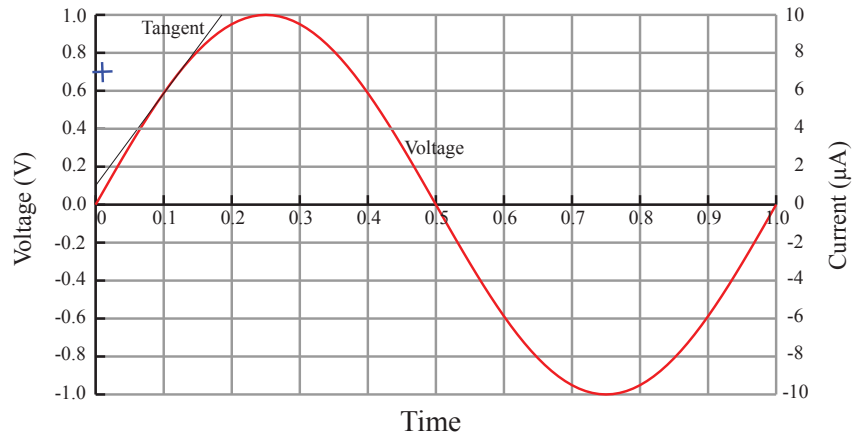


Figure 7-14

If we keep doing this for lots of different points along the voltage curve then we build up a plot of the current as a function of time, as in Figure 7-15. Now we can see that the current follows a path that is also sinusoidal and has the same frequency as the voltage, but is 90° out of phase. If the amplitude of the voltage were increased then the slope at each point would increase and so the current would increase.

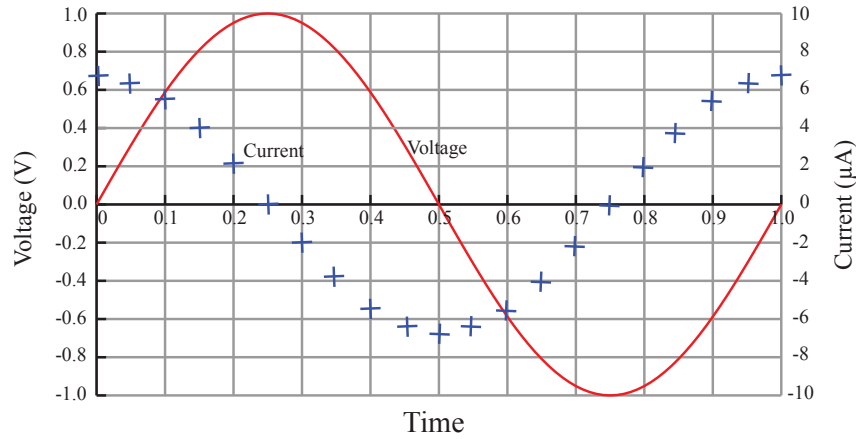


Figure 7-15

Similarly, if the frequency of the sinewave were increased then the voltage would have to rise the same amount in a shorter time, so the slope at each point would increase. In fact, with the aid of a little calculus, we can see that if

$$V(t) = V_0 \times \sin(\omega t)$$

then

$$I(t) = C \times \frac{dV}{dt} = \omega C \times V_0 \cos(\omega t) = \omega C V_0 \sin\left(\omega t + \frac{1}{2}\right)$$

which is in complete agreement with our schematic version in Figure 7-15.

So, the current in a capacitor leads the voltage by 90° and the magnitude of the current depends on both the magnitude of the driving voltage and on its frequency. Apart from the phase shift, the capacitor behaves like a resistor whose magnitude depends on the frequency. At low frequencies the current is very small; the effective resistance is very high. As we go to higher and higher frequencies, the current gets larger and larger as the effective resistance gets smaller and smaller. We often call this effective resistance the **impedance** of the capacitor, symbol Z . Thus we say that the impedance of a capacitor to sinewaves of angular frequency ω is

$$Z = \frac{1}{\omega C}$$

Note We have now derived rigorously the behavior that we described intuitively when we first looked at the capacitor. The higher the frequency, the smaller the voltage that a given current produces—the smaller the effective resistance of the capacitor.

Info Impedance

There is a subtlety that we are ignoring at this point. If we put a sinewave across a real resistor, the current will be in phase with the voltage. As we shall see below, the current in a capacitor driven by a sinewave voltage will be 90° out of phase with the voltage—a sine voltage will result in a cosine current. A more advanced study of the behaviour of capacitors accounts for that phase shift using complex numbers and so the correct definition of the impedance is

$$Z = \frac{1}{i\omega C}$$

where i is square root of -1

7.4 RC Circuits and Sinewaves

We have seen that even a very simple circuit containing a resistor and a capacitor behaves very differently from a purely resistive circuit when it is driven by a square wave. The output voltage is a very different shape from the input. By contrast, as we shall explore below, when the same circuit is driven by a sinusoidal voltage we find that the output voltage is also sinusoidal and that it has the same frequency as the input. This observation can be extended to all circuits made up from resistors and capacitors.

Remember If a circuit that is made up only of resistors and capacitors is driven by a sinewave, then **all** currents and voltages in the circuit will be sinusoids with the **same frequency** as the driver but different amplitudes and phases.

This property is called **linearity**. Strictly, a linear component or circuit is one where the currents and voltages are all related by linear functions. Linear functions include addition and subtraction, multiplication or division by a constant, and integration and differentiation with respect to time.

Info In electronics a **constant** is a quantity which does not vary with time so that, for example, ω and ϕ are constants even though they may take different values under different driving conditions.

7.4.1 The series RC-circuit with sinewave drive

Figure 7-16 shows essentially the same circuit that we studied earlier (section 7.2).

Back then we drove it with a square wave and found that a very different shape emerged. Now we drive it with a sinewave and we find that a sinewave emerges but its amplitude and phase are altered. If the driving voltage, V_i , is given by

$$V_i(t) = V_i \times \sin(\omega t)$$

then the output voltage, $V_o(t)$, is found to be (see text box overleaf)

$$V_o(t) = \frac{V_i}{\sqrt{1 + (\omega RC)^2}} \times \sin(\omega t + \phi)$$

where the phase angle ϕ is given by

$$\tan \phi = -\omega RC$$

The details of the behavior depend on the frequency through the terms in ωRC . At very low frequencies, this quantity is much less than 1 but it increases as ω increases. At some special frequency, which we shall call ω_0 , the quantity is equal to 1. This happens at

$$\omega_0 = \frac{1}{RC}$$

So we can write this quantity ωRC , upon which the behavior depends, as ω/ω_0 and find that

$$V_o(t) = \frac{V_i}{\sqrt{1 + (\omega/\omega_0)^2}} \times \sin(\omega t - \tan^{-1}(\frac{\omega}{\omega_0}))$$

The relationship is made somewhat clearer if we look at the input and output for several different drive frequencies. At low drive frequencies, we have the behavior of Figure 7-17.

Here the output wave is very nearly in phase with the input and has essentially the same amplitude. This makes sense since at low frequencies the capacitor has a very high reactance. If we think of the circuit as a voltage divider with the capacitor forming the lower leg then it is clear that, when the reactance of the capacitor is very much greater than that of the resistor, almost all the voltage will appear across the capacitor. Then we have the situation shown here where the output voltage is essentially equal to the input voltage.

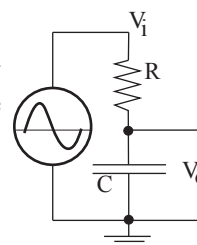


Figure 7-16

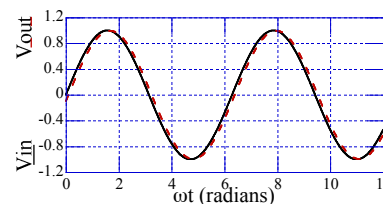


Figure 7-17 RC at $\omega = 0.1\omega_0$

Theory of the RC circuit driven by a sinewave

As we saw in the last section, the current in a capacitor driven by a sinewave is also sinusoidal but it is 90° out of phase with the driving voltage. This suggests that the output from the RC circuit will also be a sinusoid that is out of phase with the driving voltage and that may have a different amplitude. So if we let the driving voltages be

$$V_i = V_0 \sin(\omega t)$$

then we will assume that we can write the output voltage as

$$V_o = V \sin(\omega t + \phi)$$

where V and ϕ are quantities that we need to find. Our assumption will be proved if we can find values of V and ϕ that are independent of time.

We start from the fact that the current in the resistor is given by Ohm's law, so that

$$I_R = \frac{V_i - V_o}{R} = \frac{V_0 \sin(\omega t) - V \sin(\omega t + \phi)}{R}$$

Then we know from section 7.3 that the current in the capacitor is related to the voltage V_o across the capacitor by

$$I_C = \omega C V \cos(\omega t + \phi)$$

Assuming that no current is drawn from the output, Kirchoff's current law tells us that $I_C = I_R$ so that

$$V_0 \sin(\omega t) - V \sin(\omega t + \phi) = \omega R C V \cos(\omega t + \phi)$$

This equation is a little tricky to solve because it requires some of the less remembered trig. identities. First we use the sum-of-angles formulae to split the $\omega t + \phi$ terms into their pure sin and cosine components:

$$V_0 \sin(\omega t) - V \sin(\omega t) \cos(\phi) - V \cos(\omega t) \sin(\phi) = \omega R C V \cos(\omega t) \cos(\phi) - \omega R C V \sin(\omega t) \sin(\phi)$$

Then we note that the only way that the two sides can be equal at all times is if the sin components are equal and, separately, the cosine components are equal. That means that this one equation is equivalent to the two simultaneous equations

$$V_0 \sin(\omega t) - V \sin(\omega t) \cos(\phi) = -\omega R C V \sin(\omega t) \sin(\phi)$$

and

$$-V \cos(\omega t) \sin(\phi) = \omega R C V \cos(\omega t) \cos(\phi)$$

From the second equation we find

$$-\sin(\phi) = \omega R C \cos(\phi) \rightarrow \tan(\phi) = -\omega R C$$

which we can substitute into the first equation to find

$$V_0 - V \cos(\phi) = \omega^2 R^2 C^2 V \cos(\phi)$$

so that

$$V = \frac{V_0}{(1 + \omega^2 R^2 C^2) \cos(\phi)} = \frac{V_0}{\sqrt{1 + \omega^2 R^2 C^2}}$$

Our assumption has worked out; the final values of amplitude and phase are independent of time. We find that if you drive a circuit consisting of a series combination of a resistor and a capacitor with a sinewave then the output is a sinewave with the same frequency but a different amplitude and phase!

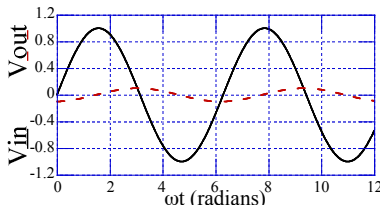


Figure 7-18 RC at $\omega=10\omega_0$

At high frequencies ($\omega = 10\omega_0$, Figure 7-18), the output amplitude is much smaller than the input and the output is about 90° out of phase with the input. Again this makes sense. At high frequencies the reactance of the capacitor is much lower than that of the resistor and so the voltage dropped across the capacitor (the output voltage) is much smaller than that dropped across the resistor.

The last case to examine is the special one where $\omega = \omega_0$. At this frequency the reactance of the resistor is exactly equal to that of the capacitor and so we expect that the output voltage will be just half of the input voltage, Figure 7-19.

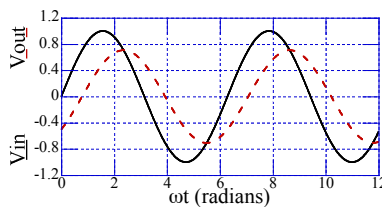


Figure 7-19 RC at $\omega=\omega_0$

As we see that is not quite right. The actual output voltage is more than half the input voltage, in fact, from the equation, we see that $V_o = V_i \times \sqrt{2}$. In making our guess we ignored the fact that the current and voltage in the capacitor are out of phase. This means that at the time at which the voltage across the capacitor peaks, the voltage across the resistor has already fallen below its maximum value. In fact the resistor voltage was a maximum when the capacitor current peaked. Thus the maximum capacitor voltage is bigger than half the input voltage and the output voltage is 45° out of phase with the input.

7.5 The Bode Plot

We have just found a complicated relationship between the input and output voltages for a circuit containing both a resistor and a capacitor. The usual way to study such a relationship is with a **Bode Plot**. This is a plot that shows the relationship between the input to a circuit and

the output from it as a function of frequency. Because we are dealing with sinewaves we have to look at both the amplitude and the phase of the signals so that we either use one plot with two lines on it or two separate plots. I will start with two separate plots for clarity.

7.5.1 Amplitude

The first plot shows the relationship between the amplitude of the input sinewave and that of the output sinewave. On the vertical axis we put the ratio of the output to the input while we put the frequency of the driving sinewave on the horizontal axis. Figure 7-20 shows the plot for our circuit in the case where $RC = 0.0001\text{sec}^{-1}$. The first thing to notice about this plot is that the axes are logarithmic. The range of frequencies and of ratios is too large to show up sensibly on a linear graph.

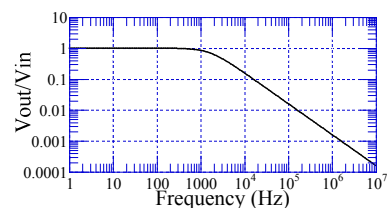


Figure 7-20

For example, Figure 7-21 is the same graph plotted on the best set of linear axes that I could find. On this scale you really can't see very much of the behavior because the entire flat portion at low frequencies is lost in the first millimeter or so of the curve and the long tail at high frequencies appears almost flat on a linear scale. So, a Bode plot always has its frequency and amplitude axes plotted logarithmically. In fact, the most common way to plot the amplitude is on a scale in decibels as in Figure 7-22.

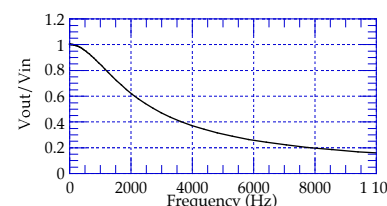


Figure 7-21

Once we put the graph on a linear decibel scale, we see that it is made up of three portions. At low frequencies, below about 600Hz in this example, the graph is a straight horizontal line at 0dB indicating that no power is lost in the circuit; the output voltage is the same as the input. At high frequencies, above about 6000Hz in this case, the graph is again a straight line but this time the line falls at a steady rate of 20dB for each factor of ten increase in frequency. We say that the output falls at the rate of **20db per decade**. Another way to express this is to note that the line falls by 6dB for each factor of 2 increase in frequency. So we can also say the output falls at a rate of **6dB per octave**.

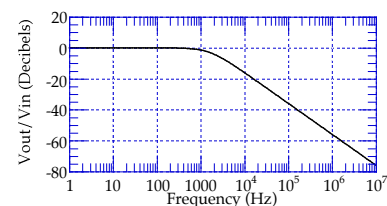


Figure 7-22

Finally there is a smooth curved portion that connects the two straight lines. This portion has the distinction that it passes through the 3dB loss point when the angular frequency of the drive is exactly equal to ω_0 . At this characteristic **cut-off frequency** we have

$$f = f_0 = \frac{1}{2\pi RC} \quad \text{at } V_{out} = \frac{V_{in}}{\sqrt{2}}$$

Note Octaves & Decades

We usually plot frequencies on a logarithmic scale so that we use measurement words that describe ratios of frequencies rather than additive frequencies.

We get our notation for the case $f_2 = 2 f_1$ from music, where doubling the frequency raises the pitch of a note by 1 octave, so that we say that if $f_2 = 2 f_1$ then f_2 is 1 octave higher than f_1 .

In another common case we say that if $f_2 = 10 f_1$ then f_2 is one decade higher in frequency. Similarly we might talk about the interval from 100Hz to 10,000Hz as two decades in frequency.

The Decibel

Ratios, which may take values that range over many orders of magnitude, are often described using a scale based on logarithms. We call these units **decibels**. These are defined in terms of power by the equation

$$\text{dB} = 10 \times \log_{10} \frac{P_o}{P_i}$$

For example, an amplifier that produces 1W of output from 50mW of input has a gain of 20 ($=1/0.05$) so that we would specify the gain as $10 \log_{10} 20 = 13\text{dB}$.

Because the power in a resistor, R, is related to the voltage across the resistor, V, by $P = V^2/R$, and because $\log(x^2) = 2 \log(x)$, we can write the ratio in terms of voltages as

$$\text{dB} = 20 \times \log_{10} \frac{V_o}{V_i}$$

Here are some common ratios and their decibel equivalents.

$P_o = 100P_i$	+20dB	$V_o = 10V_i$	+40dB
$P_o = 10P_i$	+10dB	$V_o = 3.16V_i$	+20dB
$P_o = 4P_i$	+6dB	$V_o = 2V_i$	+6dB
$P_o = 2P_i$	+3dB	$V_o = \sqrt{2}V_i$	+3dB
$P_o = P_i$	0dB	$V_o = V_i$	0dB
$P_o = 0.25P_i$	-3dB	$V_o = 0.5V_i$	-3dB
$P_o = 0.5P_i$	-6dB	$V_o = 0.7V_i$	-6dB
$P_o = 0.1P_i$	-10dB	$V_o = 0.316V_i$	-10dB
$P_o = 0.01P_i$	-20dB	$V_o = 0.1V_i$	-20dB
$P_o = 0.001P_i$	-30dB	$V_o = 0.0316V_i$	-30dB

Note The decibel is formally defined only for ratios. It can be extended to an absolute measure by specifying a standard as a reference.

One common standard is a power of 1mW. Thus absolute power of sometime specified in dBm units defined by

$$\text{dBm} = 10 \times \log_{10} \frac{P}{1\text{mW}}$$

For example, a power level of 0.3W could be written as $10 \times \log_{10} 300 = 25\text{dBm}$.

This circuit is known as a **low-pass filter** because it cuts off high frequencies but lets low frequencies pass through without change.

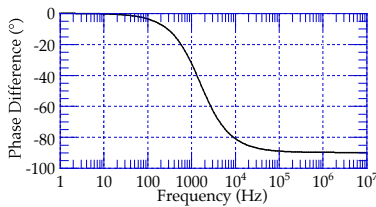


Figure 7-23 Phase Bode Plot for Low-Pass Filter

Remember Ordinary frequency is given in Hertz or cycles-per-second while angular frequency is given in radians-per-second using the equation $\omega = 2\pi f$.

7.5.2 Phase

The phase part of the Bode plot is usually presented in degrees. Figure 7-23 shows the phase half of our example plot using the same value of RC.

Combined Bode Plot

Lastly, we have the combined plot of Figure 7-24. Here the amplitude and phase are shown on the same plot with the amplitude axis on the left and the phase axis on the right.

With the two graphs superimposed we can see that at low frequencies ($f \ll 1/(2\pi RC)$, $\omega \ll 1/(RC)$) the signal passes through untouched in amplitude and in phase. As the frequency approaches the critical value $f_0 = 1/(2\pi RC)$ the output starts to fall behind the input and the amplitude starts to fall. At the critical frequency, the output is 45° behind the input and the amplitude has fallen by 3dB compared to the input.

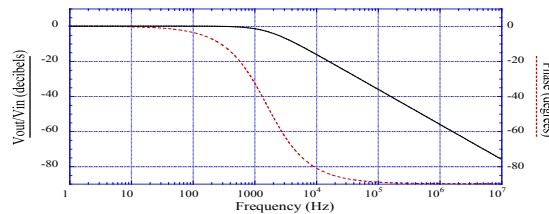


Figure 7-24 Amp/Phase Bode Plot for Low-Pass Filter

From then on the phase shift quickly rises to -90° and the amplitude falls at a rock steady rate of -20dB/decade of frequency. This circuit lets low frequencies pass through unchanged, but blocks high frequencies very effectively. We shall see this again in chapter 9.

7.6 Thévenin and capacitors

Once we add capacitors to our repertoire of components, we have to modify slightly our understanding of Thévenin's theorem. For a full restatement of Thévenin's theorem we require complex mathematics that is beyond the scope of this book, but we can get a useful version by saying that we have to add a Thévenin capacitance in parallel with the Thévenin resistance. Unfortunately, we need more advanced methods to calculate this capacitance, but it does mean we can keep our idea of replacing complex circuits by simple Thévenin equivalents. The main effect is to make the Thévenin resistance frequency dependent. In practice, the capacitive portion of the Thévenin resistance causes so many problems that we try to keep it as small as possible so that it plays as little role as possible in the behavior of circuits. We can normally ignore the capacitance of well-designed equipment at all but the highest frequencies, tens to hundreds of megahertz.

Info A **passive component** is one that can only decrease or leave unaltered the power of a signal applied to it. By contrast, an active device can increase the power of a signal so long as it has the use of an external power source. Passive components include resistors, wires, switches, capacitors, inductors, transformers, diodes (except for very rare diodes such as tunnel diodes and Gunn diodes), light bulbs, and transducers. Active devices include transistors, integrated circuits, and vacuum tubes. Almost all interesting electronic circuits need active devices to function. Active devices lie at the heart of all the amplifiers, oscillators, and logic gates that make up the building blocks of our electronic world.

7.7 Common uses of capacitors

After resistors, capacitors are the second most common passive component found in most electronic circuits. Most of those uses fall into one of three categories: frequency selective circuits, AC coupling/DC blocking, and power supply smoothing or **decoupling**.

7.7.1 Frequency selective circuits.

Because the effective resistance of a capacitor to sinewave signals varies with the frequency of the signal, becoming smaller as the frequency becomes larger, we can use capacitors to build circuits that respond to different frequencies of signal in different ways. One common example of such a circuit is called a filter. A filter is a circuit that allows some frequencies to

pass through unaltered while preventing other frequencies from passing through. We shall explore such filters in Chapter 8.

Another example is an oscillator, a circuit that is designed to generate a signal of some desired frequency. A capacitor is very often found as one of the elements that determines the operating frequency of an oscillator. The most familiar example is the tuning control of a radio. When you turn a knob to adjust the tuning of a radio you are altering the capacitance of a capacitor in an oscillator that tells the radio what frequency of signal it is to look for. We shall see examples of this use of capacitors in later chapters.

7.7.2 AC coupling/DC blocking capacitors

Because a capacitor does not allow DC current to flow through it but offers only a small resistance to high frequency AC currents, it can be used to carry a time varying signal from one circuit to another regardless of the average DC levels in the two circuits.

For example, the output of a simple transistor amplifier (chapter 19) is an amplified copy of the input signal added to a large DC offset. For example, if we start with a 0.1 V sinewave and feed it to a gain of 10 amplifier then the output voltage might be a sinewave going between 7 V and 8V. Obviously, we are only interested in the 1 V sinewave part of this signal and not in the 7.5 V DC offset that is simply there to make the amplifier work. We can remove this offset by adding a capacitor to the output of the circuit. The capacitor lets through the time varying part of the voltage but not the DC offset and so the signal seen by the load is the desired 1 V sinewave. We call this capacitor a **coupling** capacitor or a DC blocking capacitor and we say that the signal is **AC coupled**.

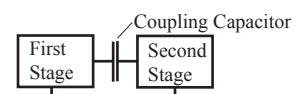


Figure 7-25 Capacitive Coupling

7.7.3 Decoupling/Smoothing capacitors

The most fundamental property of a capacitor is that it stores charge. This property leads to its use in power supply and other similar circuits where a capacitor is used to remove ripples or noise from a DC power supply.

A power supply is usually intended to deliver current at a steady constant voltage regardless of variations in the load current or in the underlying power source. It is common practice to put one or more large capacitors on the output of a power supply to smooth out any variations in the output voltage. Sudden changes in the amount of current coming from the power supply are absorbed by the capacitor. For example, a sudden increase in the current drawn from the power supply might normally cause the output voltage from the supply to drop temporarily. With a large value smoothing capacitor, anywhere from $10\mu\text{F}$ to $>10,000\mu\text{F}$, on the supply output the extra current is drawn from the capacitor without making much change in the output voltage. We often call such capacitors **smoothing** capacitors.

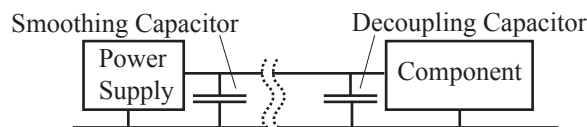


Figure 7-26 Smoothing and Decoupling Capacitors

At the other end of the power delivery chain, we often put a capacitor across the power and ground terminals of an individual integrated circuit, particularly a fast logic circuit. This serves the same purpose as the smoothing capacitor on the output of the power supply but it just acts on a single component, holding its power supply steady. If the power supply current drawn by the chip changes suddenly it will be some time (nanoseconds or even microseconds) before the power supply notices and increase the output current. In that interval the power supply voltage will drop. This has two bad effects. First it means that the chip is no longer getting as much voltage as it would like and so it may not operate properly. Second, it means that the power supply voltage for other components sharing the same power supply will be affected and the other components, instead of seeing a steady voltage, will see a noisy one. We say that the power supply noise is **coupled** from the source chip to the rest to the circuit. If we put a

Info Decouple is a fairly common term in electronics. It means, roughly, to separate a signal into separate pieces and is most often used in connection with separating a signal from noise. In particular, a decoupling capacitor is often used on power supply lines to absorb noise and so stop the noise from being coupled into a signal.

small (usually $0.01\mu\text{F}$) capacitor on the power supply terminal of the chip then the extra current will be taken from the capacitor and the voltage won't have to drop as much. This means that the operating chip will be happier and will no longer be feeding noise into the power supply for the rest of the circuit. The capacitor breaks this connection and **decouples** the chip from the rest of the circuit.

7.8 AC PSpice Simulations

Now that we have a component whose properties vary with frequency we need to learn to use some of the other forms of simulation. The first kind that we shall study is the AC Sweep analysis which creates a Bode plot of the response of a circuit. It does this without actually letting you see the sinewaves going either in or out. For that, we need the second kind of analysis called the Transient analysis. This curious name actually means the sort of analysis that one would naturally expect, one that traces the voltages sources and outputs as a function of time to produce an oscilloscope style display.

As before, we shall first explore the new analysis with a familiar example, the Low-Pass Filter. We have already studied the Bode Plot for this system and so will be able to evaluate the results of the simulation.

1) Creating the Circuit.

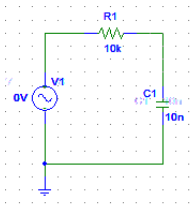


Figure 7-27 Low-Pass Filter

For this circuit we shall need two new kinds of component, the capacitor and the AC voltage source. As before, we get these from the Part Browser. The capacitor is simply called c and the voltage source VAC. Using these new parts construct the circuit shown on the left.

We have to set the magnitude of the AC voltage in the usual fashion, by double-clicking on the value label. We usually set the AC voltage to 1V so that the output gives the gain of the circuit directly. Note that we do not get to say anything about the frequency of the signal at this stage. That will come next.

2) Specifying the Analysis

Since we want more than the default Bias Point analysis we have to select the analysis type. Go to the analysis menu and click on the Setup... item. This will bring up the dialog box below.

The new analysis is called an AC Sweep. Click on the check-box to set the flag as shown in the figure. This tells PSpice that we wish to perform the analysis but does not tell it how. For that we have to click on the AC Sweep button.

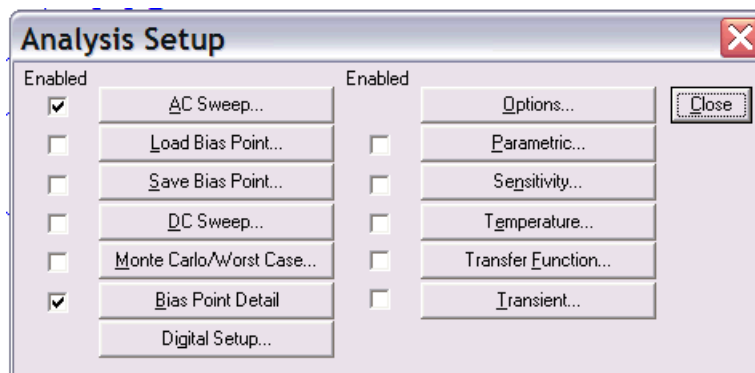


Figure 7-28

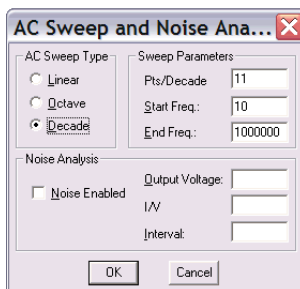


Figure 7-29 AC Sweep

Clicking on that button will bring up the AC Sweep parameter dialog shown on the right. There are a lot of boxes to fill in here. For the moment we are concerned only with the upper half of the dialog since we do not need to perform a noise analysis.

First we must select the kind of Frequency axis that we wish to use; we must select the **AC Sweep Type**. The most common form is the Decade analysis. This gives us a logarithmic frequency axis upon which we choose the number of measurement points per power of ten in frequency. For example, with one measurement per decade you might get readings at exactly

10Hz, 100Hz, 1kHz, and 10kHz. The Octave setting allows you to choose the number of readings per octave of frequency rather than per decade and tends to result in rather odd frequency axes. The linear axis is only useful for studying the behaviour of a circuit over a very narrow range of frequencies, 1 decade or less.


Next we need to specify the range of frequencies to study. First we specify the number of points per decade of frequency at which the analysis will be performed. The more points we choose, the smoother will be the resulting graph but the longer the analysis will take. I usually choose 11 points for a first analysis, though I might use more than a hundred when trying to study some small region of the curve in more detail.

Lastly we specify the range of the sweep. This can be extremely broad. Since the circuit is only simulated and there are none of the weird parasitic components and instrumental limitations that constrain us in real life this can be as wide a range as we like. In the case of a totally unfamiliar circuit you might make this a very large range, 0.001Hz to 1,000,000,000Hz for example, in order to get an idea of where the interesting behavior took place (though in that case a reduction to 3-5 points per decade might be prudent). In this case, however, we expect the interesting behavior to occur around the range $f=1/2\pi RC=1591\text{Hz}$. Thus we will study the range from 10Hz to 1MHz, much as we would in lab.

Once you have entered the values, so your dialog box looks like the figure, click on OK or press return to dismiss the dialog and then press the close button of the Analysis Setup dialog.

3) Specifying the Measurement Points

If you ran the analysis now you would see the bias point labeled on the schematic but see no output from the analysis. All the voltages and currents at all the frequencies would be computed but we not selected any for display. It is as though we ran the circuit without connecting up the oscilloscope. We must now select the points to display, much as we must connect up the oscilloscope probes in the lab. The only differences are that these 'probes' cannot have any effect upon the circuit and that we can place as many of them as we like.

We select analysis points by placing voltage (or current) markers at the desired points in the circuit. Go to the toolbar and select the Place Voltage Marker button, . The cursor will turn into an arrow dragging a voltage marker, as shown in Figure 7-30 on the right.

Drag the cursor around and left click on the upper terminal of the battery and on the junction between the resistor and the capacitor to get a schematic that looks something like Figure 7-31.

4) Run the Analysis

As before, we set the analysis in motion either by selecting Simulate from the Analysis menu or by pressing the F11 function key. If this is the first analysis since you started schematic then the Probe window will automatically come to the front when the simulation is over. Otherwise, it may stay hidden but the task-bar icon will be highlighted to show that the window wants attention and you must click on the task bar icon to bring the window to front. Either way you should end up with a window that looks something like this.

5) Interpreting the Results

The main part of the window shows a graph of the simulation results with a line for each voltage marker placed on the schematic. The legend in the lower left of the pane shows us that the first line is the voltage at the + terminal of component V1, the AC voltage source, while the second line is the voltage at terminal 2 of resistor R1.

The graph shows that the input voltage is a constant 1V regardless of frequency while the output is flat at first and then decreases as the frequency rises above about 1kHz, getting very small by about 100kHz.

It is more usual to view a Bode plot with a logarithmic Voltage axis and we can easily alter the display in that fashion. The straightforward way is to go to Axis Settings... item of the Plot menu and to click on the Y Axis tab and select Logarithmic. The easy way is to click on the

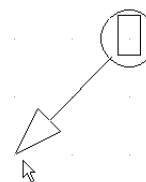


Figure 7-30

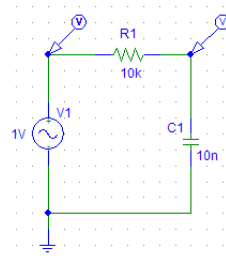


Figure 7-31

I have altered the colours, in particular changed the background from black to white, to make the output more visible when printed. On the screen you will have coloured lines on a black background.

toolbar button for this purpose, which looks like this . Either of these actions will change the graph to look like this.

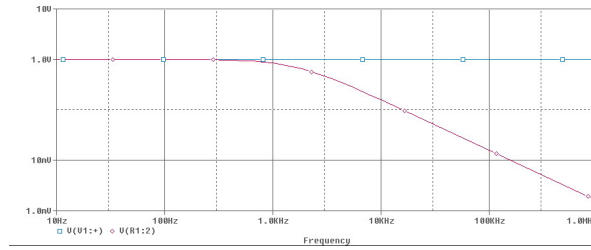


Figure 7-32 Low-Pass Output on Log Axes

Here we see the familiar turn over from a horizontal straight line below the cut-off frequency to a straight line falling at 3dB per octave (10dB per decade).

6) Plotting the Phase

The default output graph shows only the magnitude of the output voltage. We are often also interested in the phase of the output relative to the input. First, use either the Log Y Axis button or the Axis Setting dialog to set the Y axis back to Linear mode.

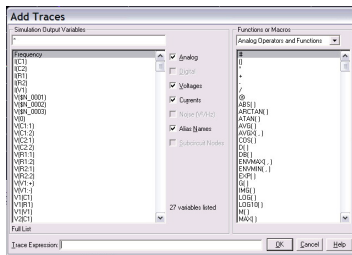


Figure 7-33

Now the Add Trace menu item from the Trace menu will bring up the dialog shown in Figure 7-33 on the left. The left-hand list shows all the quantities computed during the simulation. The phase function is called P(). You can either build the expression by clicking first on P() and then on V(R1:2) or by typing the expression into the Trace Expression box.

Once you dismiss the dialog box the graph will be redrawn with the new variable added (Figure 7-38).

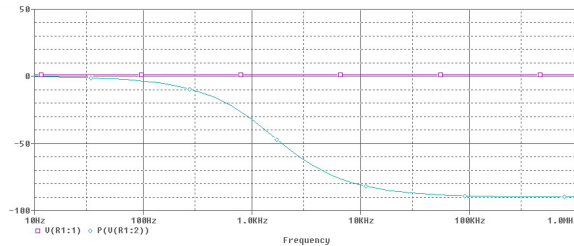


Figure 7-34 Low-Pass Phase Plot

7.9 Real Capacitors

Real capacitors come in a bewildering variety of shapes and sizes and in values spanning 12 decades of capacitance. There are several different major types with different purposes. Here are the most common ones.

7.9.1 Silver-Mica

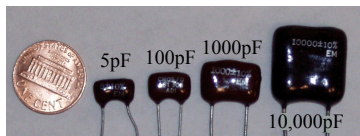


Figure 7-35 Silver-Mica Capacitors

These are small value capacitors, typically 1pF-500pF. They are quite stable and they work well up to extremely high frequencies, in the Giga-Hz range. They are made from flat plates of mica with a thin silver coating. Wires are soldered to the coating and the whole thing is dipped in a plastic coating (usually dark red). The value is usually marked quite plainly on the coating. For example, a 10pF 100V capacitor will be marked 10pF, 100V. These are expensive (\$1-2 each) and are chiefly used in very high frequency radio circuits.

7.9.2 Ceramic Disks

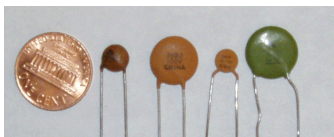


Figure 7-36 Ceramic Disk Capacitors

These are the commonest small and medium value capacitors. They are available in values from 10pF up to 1μF and come in voltage ratings of 50V, 100V, and 500V. They are made from a flat disk of ceramic material with metal plates evaporated onto the surface and wires soldered on either side of the disk. The disk is dipped in a thin coating, which is usually buff or brown but can be green or other colors. The value is marked in one of several confusing codes

and it is always wise to test the value with a meter. Very small capacitors are usually marked with the value in pico-farads using little numbers and the resistor coding scheme so that a small circular capacitor marked 101 is probably a 100pF capacitor (10 and 1 more zero = 100), a similar capacitor marked 470 is a 47pF capacitor, and one marked 103 is 10,000pF or 0.01 μ F. Larger values are normally marked in microfarads so that a capacitor marked 0.01 is almost certainly 0.01 μ F. Ceramic capacitors are not particularly stable and so are chiefly when the precise value is not important, often on power supply lines to decouple them from noise.

7.9.3 Polystyrene/Polyester/Mylar Capacitors

These are the kings of stability. They are made of thin strips of tightly wound metal coated plastic (several different types) with metal wires attached at each end. These are then coated with epoxy resin and stamped with a value using the same sort of messy scheme used for ceramic disks. For example, the 0.47 μ F = 470,000pF capacitor in the figure is marked 474. These are quite expensive (\$0.50-\$2 each) and are found mostly in precision circuits such as oscillators and filters.

7.9.4 Electrolytic capacitors

These strange beasts dominate the large value world, coming in values from about 1 μ F all the way up to 1F. They are made from thin sheets of aluminum separated by paper dampened with a chemical. The insulation is an ultra-thin layer of aluminum oxide formed on the surface of the metal when a DC voltage is applied to the capacitor. This oxide layer will dissolve if the voltage is reversed and so electrolytic capacitors have a positive end and a negative and must be used carefully to make sure that they are never used backwards. If you do reverse bias an electrolytic then it will usually be destroyed, sometimes explosively! Electrolytic capacitors are not terribly stable or well specified—it is not unusual to find that the actual value is 20% low or 50% high—and they perform very poorly at high frequencies. They are chiefly used in power supplies as filters and energy stores. Electrolytic capacitors are big enough that they have their value and working voltage stamped on them in clearly readable, sensibly phrased, letters. They range in price from about \$0.20 for a 100 μ F, 6.3V capacitor to \$5 for a 10,000 μ F, 10V capacitor, to tens of dollars for high value, high voltage units.

7.9.5 Tantalum capacitors

Tantalum capacitors are actually a kind of electrolytic but they are much better performers than their aluminum cousins. They look like little blobs of colored epoxy with wires sticking out. While they usually have their values stamped on them as numbers, some of the smallest ones use color bands like resistors. Tantalum capacitors are only useful at low voltages and are quite expensive (the 47 μ F, 10V unit in the figure cost about \$2) but they operate well at high frequencies and are the preferred capacitor for high value decoupling tasks.

7.9.6 Surface mount capacitors

Many of the above kinds of capacitor can now be purchased as little blocks, without leads but with contact pads at each end. These are designed for use in surface mount situations where the components all sit on the same side of the board as the wiring and are soldered directly to little pads instead of having leads that pass through the board. Some of the largest, such as the tantalum capacitor in the figure, have values printed in clear letters. However, the smaller ones have no markings on them at all. You just have to keep them straight when you buy them!

7.9.7 Variable capacitors

Just as there are variable resistors, so there are variable capacitors. These come in several styles but only in small values, usually less than 100pF. They are mainly used to tune high frequency circuits. They are used where a non-standard value is needed or where a variable value is needed. The best known application is in the tuner of a radio. When you turn a dial to tune a radio you are turning a variable capacitor and selecting a different frequency by selecting a different value for the capacitor

Info Ceramic capacitors tend to undergo significant changes in value as the temperature changes. This is usually a problem and is the reason that these are not normally used in precision circuits. Occasionally it can be useful. There are special ceramic capacitors with well defined temperature properties which are occasionally found in precision circuits. The idea is that you choose a capacitor (or mixture of capacitors) so that their variation in value with temperature compensates for some other property of the circuit which changes with temperature. These tend to be significantly more expensive than the standard ones, \$0.20 each instead of \$0.02 each.

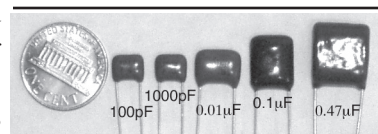


Figure 7-37 Polystyrene/Polyester/Mylar

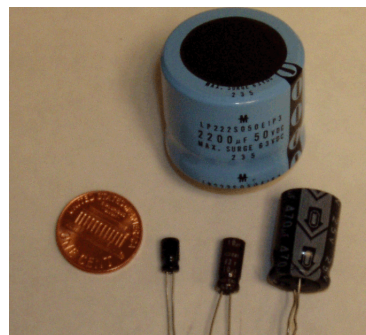


Figure 7-38 Electrolytic Capacitors

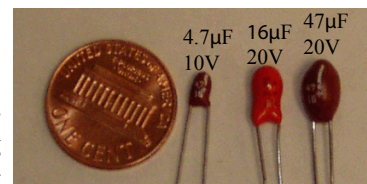


Figure 7-39 Tantalum Capacitors

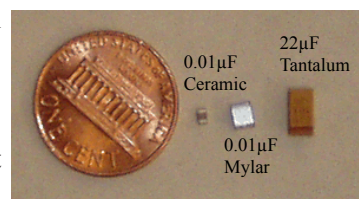


Figure 7-40 Surface Mount Capacitors

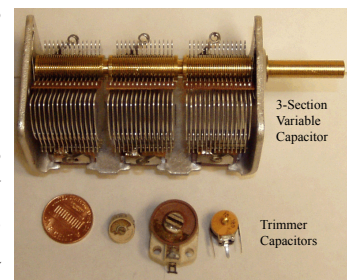


Figure 7-41 Variable Capacitors

Summary

A capacitor is a component that stores charge on two metal plates separated by an insulator. The amount of charge is related to the potential difference (voltage) between the plates by the capacitor equation

$$Q = C \times V$$

where Q is the charge in Coulombs, V the potential difference in Volts, and C the **capacitance** in **Farads**.

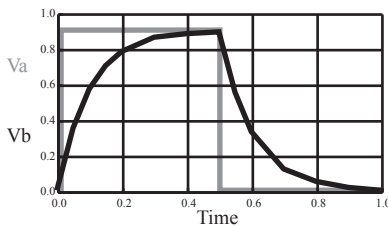
When a steady voltage, V_p , is applied to a capacitor through a resistance, R , the voltage across the capacitor changes from its initial value, V_0 , to the final value, V_p according to the equation

$$V(t) = V_0 + (V_p - V_0) \times \left\{ 1 - e^{-\frac{t}{\tau}} \right\}$$

where τ is a constant, called the **time constant**, that is given by the equation

$$\tau = R \times C$$

τ is the time that it takes for the voltage across a capacitor to reach 63% of a total change following an abrupt change in the voltage driving an RC circuit.



In the figure shown left, the input voltage (gray) changes from 0V to 0.9V at time $t=0$. The capacitor voltage (black) has reached 63% of its final value ($0.63 \times 9V = 5.7V$) after about 0.1 seconds so the time constant $T = 0.1$ sec.

In a circuit driven by a sinewave of frequency f , angular frequency $\omega = 2\pi f$, the capacitor acts like a frequency dependant resistor of value

$$Z_c = \frac{1}{\omega C}$$

The current in the capacitor leads the voltage across the capacitor by 90° .

Decibels.

Ratios are often described using decibel units. These are defined in terms of power according to the equation

$$\text{dB} = 10 \times \log_{10} \frac{P_o}{P_i}$$

So that eg. an amplifier which produces 1W of output from 50mW of input has a gain of 20 or a gain of $10 \log_{10} 20 = 13\text{dB}$.

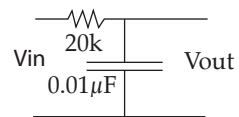
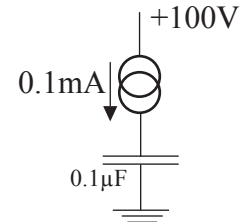
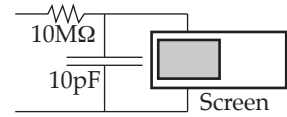
Because power is related to voltage by $P = V^2/R$, the equation for voltage ratios is

$$\text{dB} = 20 \times \log_{10} \frac{V_o}{V_i}$$

Exercises

1. A person, wearing insulating shoes, forms a small capacitor with the earth as its second plate. For an average sized person a typical value would be 100pF. A typical person also has an average resistance of about 100k Ω . When the person walks across a nylon carpet, they pick up a static charge of about 2 μC . What is the voltage on the person?
2. When the person in question 1 touches a grounded object, that charge flows to ground through the person's own resistance. Find the time constant for a normal RC discharge process for a 100pF capacitor and a 100k Ω resistor. This gives you an estimate of the duration of the spark!
3. As a capacitor discharges, the slope of the V, T curve gets flatter and flatter. Explain why this is so.
4. A typical computer fan produces a sound level of about 40dB. The sound level near an operating jet engine is about 120dB. How many times more powerful is the sound near the jet engine than the sound near the computer?

5. An amplifier increases the amplitude of a signal by a factor of 30. How many dB does this correspond to?
6. By what factor does the amplifier of question 5 increase the power of the signal?
7. The figure on the right shows the equivalent circuit of an oscilloscope input when using a scope probe. Calculate the frequency at which the impedance of the capacitor equals the resistance of the resistor. Assuming that the circuit is connected to a signal generator that produces a 2V p-p sinewave at various different frequencies, describe how the apparent size of a sinewave on the screen varies with the frequency.
8. You know that a capacitor charged through a resistor produces an interesting exponential voltage. What happens to the output voltage if a capacitor is fed from a constant current source? In particular draw and discuss the voltage across a $0.1\mu\text{F}$ capacitor driven by a constant 0.1mA current as shown on the right. You may assume that at the start of the experiment the voltage across the capacitor is zero.
9. In the adjacent figure, V_{in} is a 0-5V square wave at 1kHz. Draw the output voltage, V_{out} , as accurately as you can. Pay particular attention to the time scale over which the voltage changes.



Chapter 8:R-C Frequency Selective Circuits

8.1 Introduction

A circuit that is built entirely from resistors is, in theory, completely insensitive to the frequency of the signals upon which it operates. It can decrease the size of a signal by a fixed amount but cannot alter the shape of the shape of the signal. By contrast, a capacitor responds differently to signal with different frequencies; it presents a high impedance to low frequency signals and a low impedance to high frequency signals.

We can make circuits that respond in a wide variety of different ways to different frequencies by mixing capacitors and resistors in the same circuit. We call circuits that transmit some frequencies unchanged while preventing the passage of others **filters**. This chapter examines some simple R-C filter circuits, showing how to design them and what to do with them. More sophisticated filters can be made with larger numbers of components and especially by adding a third kind of component, an inductor, that is not described in here.

8.2 Low-pass Filter

A low-pass filter is one which allows low frequency signals to pass through but which blocks high frequency signals. The ideal low-pass filter would allow all frequencies below some limit to pass through unaltered in amplitude and phase but would completely block all higher frequencies. The Bode plot for such a filter would look like Figure 8-1.

A filter with a characteristic like this is known as a **Brick Wall** filter because any frequency outside the pass-band is stopped as if it had hit a brick wall. Frequencies within the pass band go through the filter without any attenuation.

The **Pass-band** of a filter is the range of frequencies which pass through the filter with little alteration.

The **Stop-band** is the range of frequencies which are more or less prevented from passing through the filter.

For the brick wall filter the difference is obvious, frequencies either pass through unchanged or are stopped completely. For less abrupt filters we usually say that the stop band is the range of frequencies which are attenuated by 3dB or more.

The filters that we shall study have rather less abrupt characteristics. Instead of a flat line followed by a vertical line they look like a flat line followed by a sloping line (Figure 8-2). The steeper the slope of the line, the more the filter looks like a brick wall filter. We specify the steepness of the line by the number of decibels that the output falls for each factor of two (octave) increase in frequency. A better filter has a steeper slope, often called a sharper slope.

8.2.1 The Passive Low-Pass Filter

A passive filter is one made up only of passive components; resistors, capacitors and inductors. We met the simplest such filter back in chapter 6 when we looked at a circuit containing both a resistor and a capacitor. The circuit consists of a voltage divider whose lower leg is a capacitor rather than a resistor (Figure 8-3).

At very low frequencies the capacitor acts like an open circuit, acts as if it were not there. Thus the output voltage is equal to the input voltage and in phase with it. As the frequency rises the impedance of the capacitor decreases and the impedance of the capacitor becomes an ever smaller fraction of the total resistance. Thus V_o makes up an ever smaller fraction of V_{in} .

Note In practice, a circuit built from resistors will behave differently once the signal frequency gets very high. This is because a real resistor is only an approximation to the ideal component. In reality, the leads add tiny amounts of capacitance and inductance that are negligible at lower frequencies but which must be considered once the frequency gets high enough. For circuits built using laboratory prototype techniques these effects may start to show themselves at a few MHz. Circuits built with standard components and good construction techniques can reach about 100-200MHz but above that special leadless components must be used.

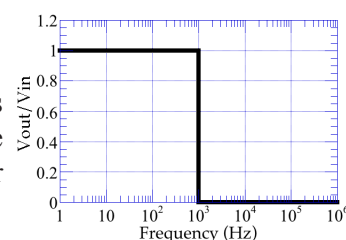


Figure 8-1 A Brick Wall filter

Note To **attenuate** something is to make it smaller. A signal that is attenuated by 3dB has its amplitude reduced by a factor of $1/\sqrt{2}$.

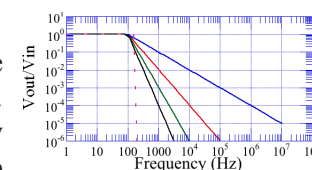


Figure 8-2 Ideal Low Pass Filters

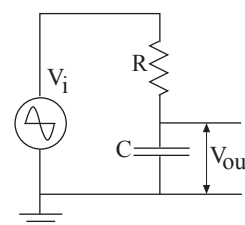


Figure 8-3 RC Low-Pass Filter

Simultaneously, the phase shift increases towards the 90° phase shift of a capacitor. Thus the Bode plot for this circuit in the case where $RC = 0.001$ looks like Figure 8-4

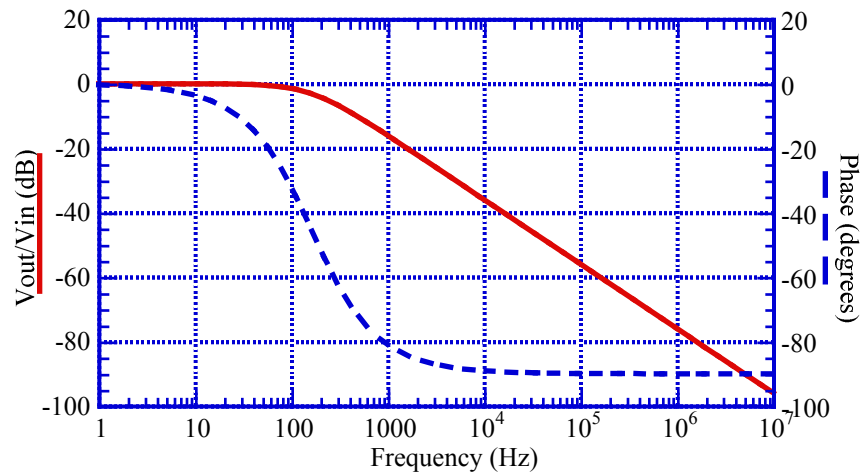


Figure 8-4 Low-Pass Filter Response

Remember Frequency ω is the angular frequency in radians/s, which is 2π times the standard frequency in Hz.

In chapter 7 we found that, mathematically,

$$\frac{V_{out}}{V_{in}} = \frac{1}{\sqrt{1+(\omega RC)^2}} \text{ and } \tan\phi = -\omega RC.$$

where ϕ is the phase shift of the output referred to the input.

Such a characteristic is quite **soft**. That is, there is a very gradual transition between the frequencies which are transmitted and those which are blocked. As I said above, the standard practice is to say that the filter passes all those frequencies that are attenuated by less than 3dB and to say that it stops those frequencies that are attenuated by more than 3dB. A 3dB attenuation corresponds to a voltage ratio of $1/\sqrt{2}$ so that the transition occurs at the frequency for which

$$\omega = \frac{1}{RC} \quad \text{or} \quad f = \frac{1}{2\pi RC}.$$

We call this transition frequency the **cut-off frequency**. It separates the pass band from the stop band.

So there are three regions to the characteristic of a simple low pass filter, a passband in which the output is equal to the input, a short transition region round the cut-off frequency, and then a stop band in which the transmission falls at the rate of 20dB for every decade of frequency increase. As noted above, we usually specify the fall-off rate in the stop band in terms of the fall in an octave, a factor of two in frequency. In those terms the simple RC filter attenuates at the rate of 6dB per octave.

8.2.2 Designing a filter for a given cut-off frequency

We can design an RC low-pass filter to have any cut-off frequency we like by choosing the values of R and C using the rule that

$$RC = \frac{1}{\omega} = \frac{1}{2\pi f}.$$

This rule only determines the product of R and C; we need a second piece of information to help us to pick the individual values. We get that piece of information from the output impedance of the circuit. For frequencies in the pass band, the output impedance of the filter is essentially equal to R.

Remember the rule that says that the output impedance of a circuit should be $<1/10$ th of the input impedance of the following circuit. We can use that rule to choose R with the following algorithm

Remember Low-Pass Filter Design Rules

- 1) Find the input impedance of the following circuit.
- 2) Choose $R < 1/10$ th of that impedance.
- 3) Select $C = \frac{1}{\omega R} = \frac{1}{2\pi Rf}$.

Example

Design a filter with a cut-off at 70Hz to deliver a signal to a circuit with a 1 M Ω input impedance.

For an input impedance of 1 M Ω we need to choose $R < 100$ k Ω . I will pick $R = 10$ k Ω , a standard value nicely smaller than the limit.

$$\text{Then I set } C = \frac{1}{2\pi \times 10000 \times 70} = 227 \text{ nF.}$$

That is not a standard value and so we have a choice to make. In all but high precision situations we can use the nearest standard value, in this case 220 nF (0.22 μ F) and get a filter that is close enough for most practical needs. Otherwise I will make a part of either the resistance or the capacitance variable and will adjust it to exactly tune the correct cut-off frequency.

8.3 RC High-pass Filter

If we reverse the resistor and the capacitor, then we get a high-pass filter (Figure 8-5). This time the impedance of the lower leg, the resistor, remains constant but the impedance of the upper leg, the capacitor, decreases as the frequency rises. At very low frequencies little or no current flows into or out of the capacitor and so there is no voltage across the resistor. As the frequency rises and the impedance of the capacitor falls, more and more current flows in the resistor. The output voltage rises until the impedance of the capacitor is much smaller than that of the resistor. At that point, the output voltage is equal to the input voltage.

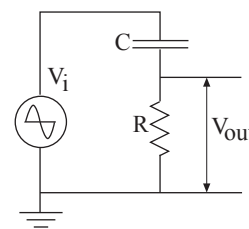


Figure 8-5
RC High-Pass Filter

Using mathematical methods similar to those of chapter 7, we find

$$\frac{V_{out}}{V_{in}} = \frac{\omega RC}{\sqrt{1+(\omega RC)^2}} \text{ and } \tan \omega = \frac{1}{\omega RC}.$$

Figure 8-6 shows the Bode plot for the high-pass filter drawn, as before, for the case where the time constant, RC , is 0.001s.

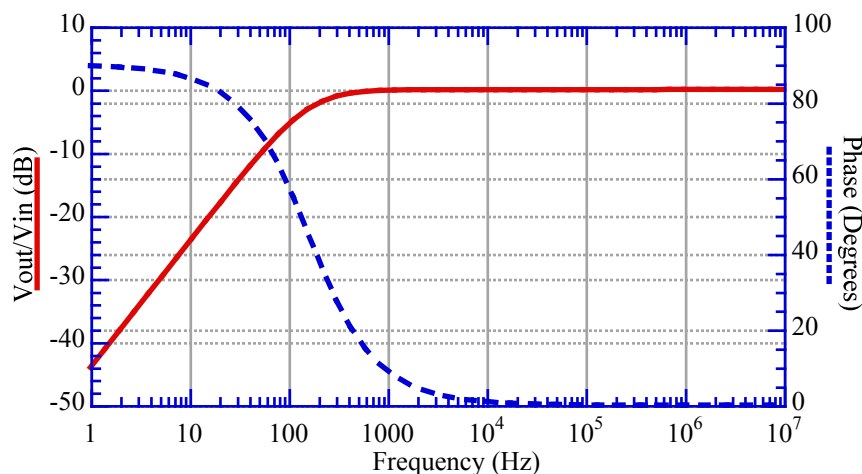


Figure 8-6 High-pass filter response

As we would expect, this filter is very similar in shape to the low-pass filter but reversed in orientation. This time the filter passes all frequencies above $\omega = 1/(R \cdot C)$ and attenuates those below at a rate of 20dB/decade or 6dB per octave.

While the details of the high-pass filter are a little different from the low-pass filter, we can choose the component values using the same rules as for the low-pass filter.

Remember High-Pass Filter Design Rules

- 1) Find the input impedance of the following circuit.
- 2) Choose $R < 1/10$ th of that impedance.
- 3) Select $C = \frac{1}{\omega R} = \frac{1}{2\pi Rf}$.

In other words, if you have a particular set of R and C values that work as a low-pass filter, then if you swap them you will have a high-pass filter with the same cut-off frequency.

8.4 Tone Controls

We have all encountered the effects of circuits like these in radios. Most radios have a pair of knobs marked treble and bass. These knobs are attached to variable resistors in circuits like those above. As you turn the treble knob you are altering the cut-off frequency of a low-pass filter. When it is all the way at the high end, the cut-off is set at about 20kHz and the filter passes all audible frequencies. As you turn the knob, the resistance increases and the cut-off frequency decreases so that the filter removes more and more of the audible, high frequency sound. Eventually you remove so much of the high frequency sound that the sound seems to be coming from under a pile of blankets. The bass knob adjusts the cut-off frequency of a high-pass filter. When it is turned all the way up then the cut-off frequency is set below about 20Hz and all the audible frequencies pass through. When the knob is turned down, it raises the cut-off frequency and starts to remove the low frequencies from the signal until the sound is thin and tinny.

8.5 Bandpass Filter

Fancier stereos, even some pricey car stereos, have a little row of slider knobs making up a graphic equalizer; a circuit that allows you more control than the usual treble and bass controls. The graphic equalizer splits the audio frequency band from 20Hz to 20,000Hz into several pieces and allows you to control the attenuation of each frequency band separately. To build such a control we need a set of **bandpass** filters. A bandpass filter allows through a band of frequencies. It has two cut-off points. Everything below the lower cut-off, ω_{bottom} is attenuated as is everything above the upper cut-off, ω_{top} . Of course, $\omega_{\text{bottom}} < \omega_{\text{top}}$.

We can make a simple bandpass filter by joining a high-pass filter and a low-pass filter together (Figure 8-7).

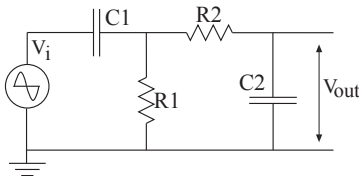


Figure 8-7 RC Band-Pass Filter

Remember Connecting two filters together

In this circuit the low-pass filter, R2 and C2, form the load for the high-pass filter, R1 and C1. This means that we have to follow the basic rule for connecting one circuit to another. This will only work as a bandpass filter so long as we choose R1, which sets the output resistance of the first filter, to be $< R2/10$, since R2 sets the input resistance of the second filter. Of course, R2 itself will usually have been chosen to be $< 1/10$ of some other resistance, whatever forms the load to which the whole bandpass filter is connected.

This produces, as you might expect, a rather soft-sided bandpass filter. For example, Figure 8-8 on the facing page shows the response of a bandpass filter with cut-off frequencies at 300Hz and 3000Hz.

So long as the top and bottom cut-off frequencies are not too close to each other, then the cut-off frequencies are given by

$$\omega_{\text{bottom}} = \frac{1}{R_1 C_1} \quad \text{and} \quad \omega_{\text{top}} = \frac{1}{R_2 C_2}$$

However, if these frequencies get too close then things get more complicated. In that case the pass-band of one section of the filter interferes with the stop band of the other section and so

there are no frequencies that pass through the filter unattenuated. Narrow bandwidth bandpass filters usually use coils of wire, called inductors, as well as capacitors and resistors. These filters involve mathematics that puts them beyond the scope of this text. For more information see e.g. Horowitz and Hill or The Radio Amateurs Handbook.

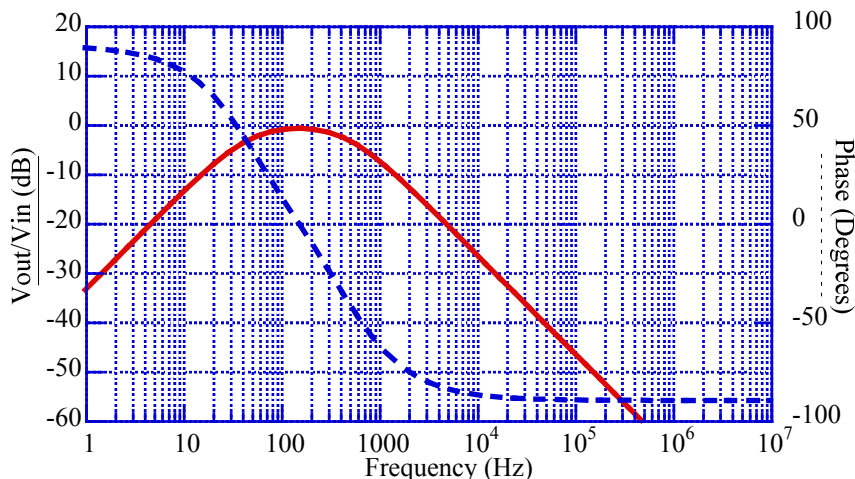


Figure 8-8 Band-Pass Response

8.6 Multi-section Filters

A single resistor-capacitor pair gives us a filter that has a slope of 6dB/octave in its stop band. Quite often this is not enough. Consider the problem of making a CD. As we shall see when we study Analog-to-Digital conversion (Chapter 28), we need to record frequencies up to 20kHz using a method that suffers from spurious noise problems if there are frequencies of 40kHz or above present. Our 6dB per octave filter will reduce the 40kHz signal by at most 6dB, a factor of 2 in amplitude. Since typical sound signals contain information with a 60-70dB range this 6dB decrease is useless. We need a filter that is MUCH better at rejecting unwanted signals. We want a filter with a sharper edge.

The simplest way to get a sharper edge is to connect the output one filter to the input of another, rather like building a bandpass filter. We call this **concatenating** the filters and we would refer to each of the individual filters as **sections** of the complete filter. Hence the name **multi-section** filter. Let us see if we can improve the 20kHz low-pass filter described above.

We will start with a single filter section. To design it we will need to know the input impedance of the load to which it will be connected. Let us take 1MΩ as a reasonably easy value to achieve. In that case the connection rule tells us that we can use a resistor of no more than 100kΩ and I will pick 51kΩ (the closest E24 value to 50k) to give us a little extra precision. Now we can pick the capacitor using

$$C = \frac{1}{\omega R} = \frac{1}{2\pi Rf} = \frac{1}{2\pi 51,000 \times 20,000} = 156\text{pF}$$

They don't make 156pF capacitors so we will settle for a 150pF capacitor (Figure 8-9), which will give us an actual cut-off frequency of 20.8kHz, quite close enough for our purposes. Now, we know that that filter will only give us a x2 reduction in voltage for a 40kHz signal so we will design another 20kHz filter to put between it and the source.

Since the filter in Figure 8-9 will become the load for our new filter we will have to choose a new resistor value. We used 1M/20 for the first filter so we will need to use 51k/20 for this filter. The closest E24 value to 2.55k is 2.4k so we will choose a 2.4k resistor. That will give us a capacitor value of

$$C = \frac{1}{\omega R} = \frac{1}{2\pi Rf} = \frac{1}{2\pi 2,400 \times 20,000} = 3316\text{pF}$$

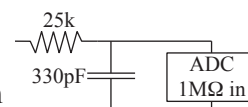


Figure 8-9 Single-section 20kHz low-pass filter.

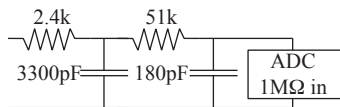


Figure 8-10 2-section low-pass filter

Again, they don't make 3316pF capacitors but we can get a 3300pF capacitor that will give us a cut-off frequency for this section of 20.1kHz, again close enough. Putting this together we get the circuit of Figure 8-10.

If we build this circuit and measure its frequency response we find two things. First, the slope of the stop band is now 12dB/octave, as we hoped. However, the cut-off frequency is NOT almost exactly 20kHz, instead it is only 10kHz!

What happened? Well, first remember that the cut-off frequency of a filter is the frequency at which it decreases the signal amplitude by 3dB. But this means that if we feed a 20kHz signal into the filter of Figure 8-10 then the signal at the input to the second section (the junction of the two resistors and the 3300pF capacitor) sees a signal that is 3dB smaller than that. When this passes through the second section it suffers another 3dB decrease so that the total loss at 20kHz is 6dB. In order to get a 3dB loss at the input to the ADC we can only lose 1.5dB in each of the two filter sections. The frequency at which a single filter cuts the output by 1.5dB is 10kHz, which is exactly what we get from the real filter. So to build a 2-section 20kHz filter we would need to use two 40kHz filters, not two 20kHz filters.

There are two lessons to learn from this:

1. When you concatenate filters, you change the cut-off frequency. You would need new rules to design multi-section filters.
2. You cannot easily concatenate more than about 2 sections because the resistor values quickly get out of hand, since each section has to have ten times the resistance of the preceding section.

There are two practical ways to make multi-section filters, with their sharper cut-offs. The old way was to replace the resistors in a filter design with components called inductors (basically coils of wire). Thanks to some mathematics that are beyond the scope of this book, inductors allow us to build multi-section filters with a variety of responses. You can find more about such filters in texts such as Horowitz and Hill or the ARRL Handbook for Radio Communication.

The only way to make large multi-section from resistors and capacitors is to find a way to isolate one section from the next. We can do this with circuits called amplifiers that we shall study starting in Chapter 18. We call circuits that use amplifiers **active** circuits and so we call filters that are built round amplifiers **active filters**. We will study such filters in Chapter 24.

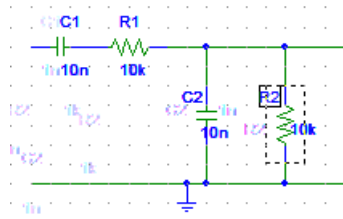


Figure 8-11 Wein-Bridge Filter

8.7 Studying a New Filter with PSpice.

Figure 8-11 shows the circuit of a new kind of bandpass filter called a Wein-Bridge filter. It consists of a high-pass filter (C1 and R2) wrapped around a low-pass filter (C2 and R1). Since both filters have the same cut-off frequency and the filters are not put one-after-another, we cannot use the rules from section 8.5. At this point, we can only guess what the overall transfer function will be. We shall investigate it with the AC Sweep analysis.

1) Frequency Response

Build the circuit and drive it from an AC voltage source. Then add voltage markers at the input and output as shown in Figure 8-12. Since the low-pass and high-pass filters each have the same cut-off frequency of 1591Hz we might as well use the same analysis parameters as last time. Make sure that the AC analysis is enabled and that the upper and lower frequencies are set to 10Hz and 1MHz respectively. Run the analysis and obtain the output shown in Figure 8-13..

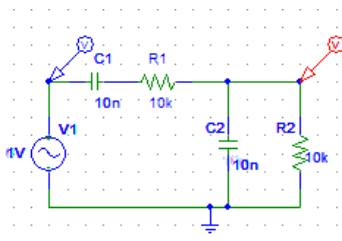


Figure 8-12

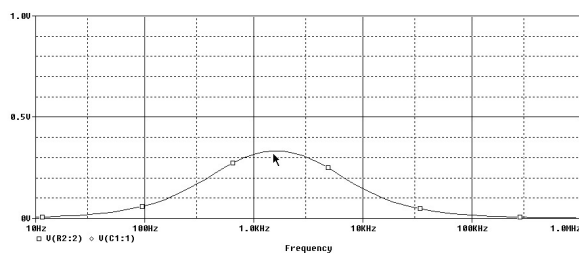


Figure 8-13

The output is clearly a kind of narrow band-pass output. The magnitude rises from almost zero at very low frequencies, peaks somewhat above 1kHz, and then falls back close to zero

2) Phase Response

We shall explore the region near the peak in more detail soon but first let us examine the phase. Add the phase trace as before to get a graph like this.

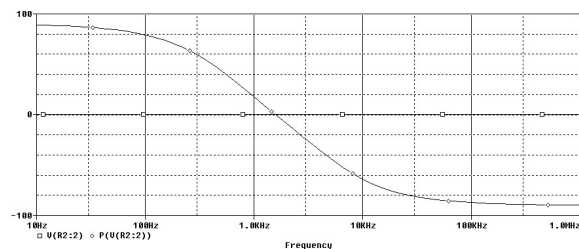


Figure 8-14

Here we see the phase go all the way from +90 degrees at very low frequency to -90 at very high. The transition is somewhat more abrupt than was the case for the simple low-pass filter and it passes through zero somewhat above 1kHz.

As before, we can't see both traces on the same scale and so it is hard to judge whether or not the phase passes through zero at the maximum of the amplitude or not. We can rectify that by adding a trace that scales up the output voltage. The maximum was a little below 0.5V so if we multiply the output voltage by 200 we should be able to see both traces on the same graph.

We can modify an existing trace by double-clicking on the legend entry. Double-click on the V(R2:2) label to bring up the Add Traces menu. Modify the Trace Expression to read $200 * V(R2:2)$ and press OK. This will rescale the amplitude trace to give the curved trace in the figure below.

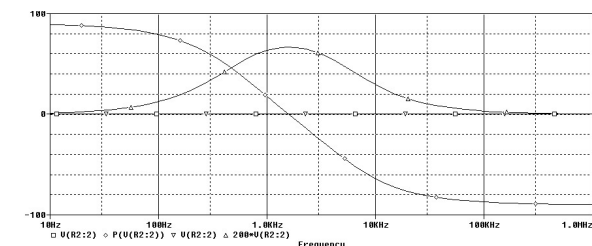


Figure 8-15

Now we can see that the voltage maximum and the phase zero do indeed occur at about the same place.

3) Zeroing in on the peak.

With only 11 points per decade we have too little information to study the region round the peak as closely as we would like. Accordingly, let us redo the analysis over a smaller range. Clearly, everything of interest occurs between 100Hz and 10kHz so return to the Analysis Setup dialog and adjust the AC Sweep parameters to put 101 points per decade and to set the region of interest between 100Hz and 10kHz. Rerun the sweep to get the result below.

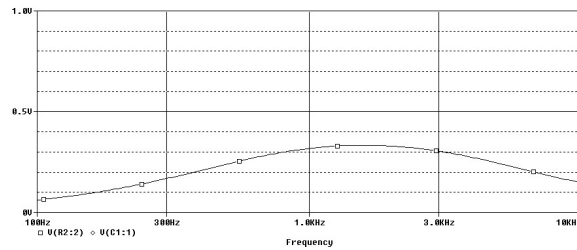


Figure 8-16

Now the peak is clearly visible but we can do better. The peak clearly lies between 1kHz and 3kHz and between about 0.25V and 0.4V. Double-click on one of the labels of the X axis to bring up the X axis dialog. Select the User Defined range and enter 1kHz and 3kHz as the limits. Then repeat the process for the Y axis, setting the limits to 0.25V and 0.4V. This will give the graph below.

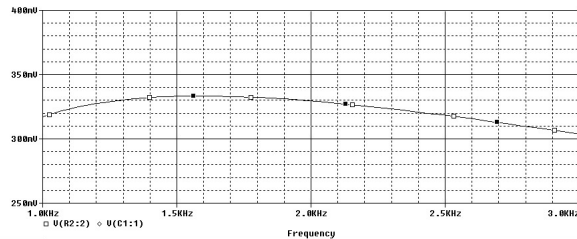


Figure 8-17

Now we can see that the peak is a little above 1500Hz, quite consistent with the expected 1591Hz. We can repeat our earlier trick of blowing up the V(R2:2) trace and adding the phase to get the last graph, Figure 8-18, which shows us that the peak output and the zero phase do indeed occur at the same frequency. The maximum output appears to be about 0.3V and theory confirms this result.

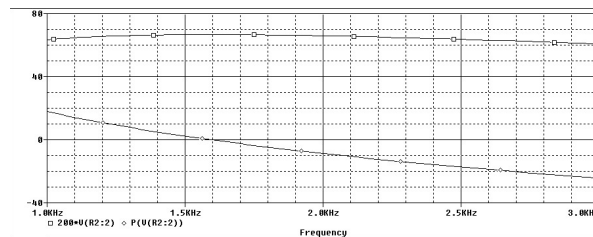


Figure 8-18

8.7.1 Transient Analysis

As discussed above, we use Transient Analysis to get a view of the actual voltages in the circuit. We shall return to the Low-Pass filter to explore this option since we have already studied the effect of passing square waves through a Low-Pass filter.

We need a different kind of voltage source to run this analysis; the VAC source is only useful for the AC Sweep analysis. This time we shall use a source called VPULSE which can generate square waves of any frequency and mark-space ratio. We shall use it to create a 50-50 square wave at 1kHz with an amplitude of 2V peak-peak.

1) Setup the Circuit.

First open the Low-Pass schematic, delete the AC voltage source, and add a VPULSE source. Then double-click the voltage source to bring up the Pulse dialog, Figure 8-19.

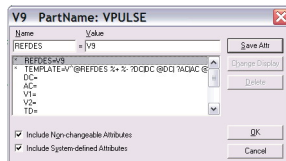


Figure 8-19 Pulse Dialog

There are a number of parameters that need to be set. Work through them and set the values as follows

DC=0, AC=1V, V1=-1V, V2=1V

TD=0, TR=10nS, TF=10nS, PW=0.5mS, PER=1mS.

These parameters have the following meanings:-

DC: the value for the power supply in the DC Bias analysis.

AC: the value for the source in the AC Sweep analysis. This allows us to use the one voltage source for both the transient and AC analyses.

V1: the lowest voltage of the pulse (or the highest, the order is irrelevant).

V2: the highest voltage of the pulse.

TD: the amount of time before the pulse train starts.

TR: the rise time of the pulse--the time it takes to go from Vlow to Vhigh.

TF: the fall time--the time to go from Vhigh to Vlow.

PW: the pulse width--the amount of time the pulse spends high.

PER: the period of the pulse train.

So we have specified a pulse train going between -1V and +1V with a period of 1mS which spends 50% of its time high and 50% low. The rise and fall times are both 10nS (which is quite fast but not really interesting for this application).

2) Setup the Analysis

Go to the Analysis menu and select Setup... to bring up the Analysis Setup dialog. Enable the Transient analysis and press the Transient button to bring up the Transient Setup dialog as shown in Figure 8-20 on the right. Once again there are lots of boxes to play with but we need only two of them.

The Print Step controls the interval at which results are stored from the simulation. This needs to be smaller than the smallest timescale in which we are interested. Since we are going to simulate a 1kHz wave we would actually be quite happy with any interval less than about 10uS. I went wild and used 10nS because that is the fastest timescale in the input wave (the rise/fall time).

The Final Time is just what it sounds like, the time at which the simulation stops. We are looking at a 1kHz wave so we want to see several cycles to get an idea of the typical behaviour. I picked 5mS to see 5 full cycles.

The other boxes are either needed only if we intend to perform a Fourier analysis on the result (a technique for looking at the frequency content of a wave rather than its shape) or for exerting unusually fine control over the simulation process. They can safely be left blank. Read the documentation if you want to know what they do.

5) Run the simulation in the usual way to obtain an output that should look like Figure 8-21 below. Note that if you asked for both an AC Sweep and a Transient analysis then you will have to select which result you wish to examine once the simulation is done.

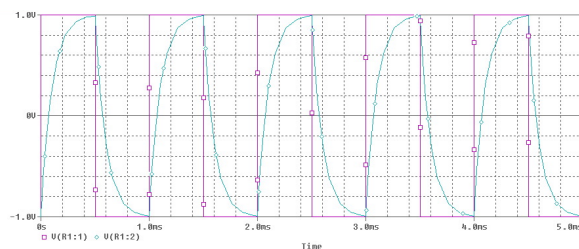


Figure 8-21 Square Wave through a Low-Pass Filter

Here we see the familiar low-pass filter behaviour. When the input signal changes sign the output starts off following it rapidly and then slows down as it approaches its final value and the current charging the capacitor gets smaller and smaller. Since this is the output of a low-pass filter we learn that the high frequencies are needed to form the sharp corners at the tops of the leading edges and the bottoms of the trailing edges. However, while it takes time for

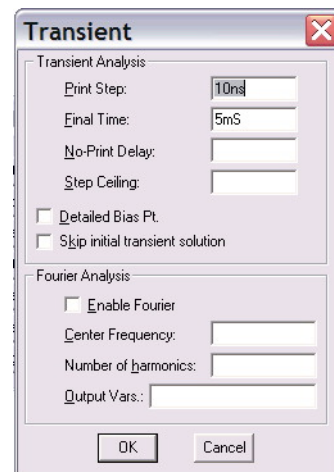


Figure 8-20 Transient Dialog

the output value to adjust to the sharp change of input, the slope of the output does suffer an instantaneous change when the input changes.

8.7.2 Square Waves and the Wein Bridge

Just for interest repeat the square wave transient analysis with the Wein Bridge. Since the bridge is a mixture of a low-pass and high-pass filter we may expect that the output will show characteristics of both low- and high-pass filters. The results bear this out. Figure 8-22 shows the output for the same 1kHz square wave that we used in the low-pass filter.

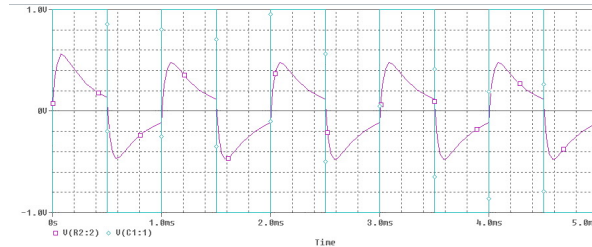


Figure 8-22 Square Wave through Wein-bridge Filter

The most notable difference is the reduced size of the output, barely half the size of the input. Otherwise the output is much as we expected. There is a sharp rise that begins to flatten off as it goes up just as we saw in the low-pass filter. However, the output does not stay at the maximum DC level. Instead, it immediately starts an exponential fall towards 0V, exactly what you would see in a high-pass filter.

Summary

A **Low-Pass Filter** allows sine waves to pass through without attenuation if their frequency is below a limit called the **Cut-Off Frequency**. Sine waves above that frequency are attenuated by at least 3 dB and the attenuation increases by 6 dB for every octave that the frequency increases. We can design either a low-pass filter or a high-pass filter for a given cut-off frequency with the following rules

- 1) Find the input impedance of the following circuit.
- 2) Choose $R < 1/10$ th of that impedance.
- 3) Select $C = \frac{1}{\omega R} = \frac{1}{2\pi Rf}$.

A **High-Pass Filter** attenuates all frequencies below the cut-off and lets through all those above. We can design a high-pass filter with the same rules as the low-pass filter.

A **Band-Pass** filter allows through a finite range of frequencies and so has two cut-off frequencies. So long as the two cut-off frequencies are not too close we can make a band-pass filter by following a low-pass filter with a high-pass filter. In that case we must be careful to make the resistor in the second filter at least 10x as large as that in the first filter!

Exercises

1. Design a low-pass filter with a cut-off frequency of 3000Hz that will deliver its output to a circuit with an input impedance of $1M\Omega$.
2. A 1kHz sinewave with an amplitude of 1V is applied to a low-pass filter made from a $10k\Omega$ resistor and a $0.2\mu F$ capacitor. What are the amplitude and phase of the output wave?
3. Design a bandpass filter to pass all audible frequencies and deliver them to an impedance of $100k\Omega$.
4. The input of an oscilloscope claims to be equivalent to a resistance of $1M\Omega$. in series with a $20pF$ capacitor forming a low-pass filter. What range of frequencies will the scope record without loss?

5. Redo the 2-section filter design of section 8.6 using 40kHz filter sections. Simulate the filter in PSpice and present a Bode plot to demonstrate that the resulting filter does indeed have a 20kHz cut-off frequency.

Chapter 9: The Diode

9.1 Introduction

The two components, the resistor and capacitor, that we have already met are linear components—the current flowing is a linear function of the driving voltage. Now we are going to meet our first non-linear component, the diode. This is basically the electrical equivalent of a one-way valve in plumbing. It allows current to flow fairly freely in one direction but prevents it from flowing in the other direction.

9.2 The ideal diode

An ideal diode would allow any amount of current to flow in one direction (called the **forward** direction) while allowing no current to flow in the other (**reverse**) direction (Figure 9-1).

Info What happened at 0V?

The device line is truly vertical at 0V. Any amount of current can flow in the forward direction without there being any voltage dropped across the component. This means that the slope resistance to current flowing in the forward direction is zero.

Similarly, the device line for negative voltages is horizontal and at 0A. Absolutely no current flows no matter how much voltage you put across the device. Since the slope of the line is zero, the resistance in the reverse direction is infinite.

The symbol for a diode (Figure 9-2) is designed to show the one way nature of its behavior. The arrow shows the direction in which current can flow. The diode conducts when the **anode**, the arrow end, is more positive than the **cathode**, the bar end. It acts as an open circuit when the cathode is more positive than the anode.

Bias: Forward and Reverse

Since the diode behaves differently depending on polarity of the voltage across it we commonly refer to the two different polarities as **Forward**, when the anode is more positive than the cathode and **Reverse** when the cathode is more positive. It is also common to speak of the DC voltage across a diode as the **Bias** across the diode. This leads to the common use of the terms **Forward Bias** for a voltage that is more positive on the anode end and **Reverse Bias** for a voltage that is more positive at the cathode end. A diode normally conducts under **Forward Bias** but not when **Reverse Biased**.

9.2.1 Rectification

If we connect an ideal diode in series with a resistor and apply a sinewave to the circuit (Figure 9-3) then we see a diode's basic behavior, called **rectification**. We will follow the behavior of the circuit as the voltage runs through a cycle Figure 9-4).

At the beginning of the cycle, the voltage at point a becomes positive with respect to ground. No current is yet flowing in the resistor so there is no voltage dropped across it and the voltage at point b is still zero. This means that point a is positive with respect to point b and current starts to flow through the diode in the direction of the arrow. When current is flowing in the forward direction, there is no voltage drop across the diode and points a & b stay at the same voltage. Thus, the output voltage at point b follows the input voltage.

Soon, the voltage at point b returns to zero and then goes negative. At the moment at which the voltage at b reaches 0 the current in the resistor reaches zero and the voltage at b is zero. Now, the voltage at point a goes negative and tries to make current flow through the diode in the reverse direction but the diode does not allow this. The voltage at point a continues to drop below zero but the voltage at point b stays at 0 since no reverse current can flow through the diode. So the output voltage follows the input during the positive half cycle but stays at zero through the negative half cycle. We call a circuit that turns an alternating—both positive and negative—voltage into a purely positive or purely negative voltage a **rectifier**.

Info A **linear function** of V is a function that depends only on V or any of its derivatives or integral. It may not depend on V^2 , V^3 , or any other power of V .

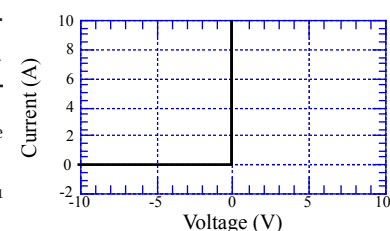


Figure 9-1 Ideal Diode I-V Curve

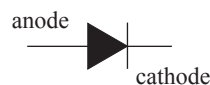


Figure 9-2 Diode Symbol

Info Occasionally you will also see or hear the term Back Biased which just means the same as Reverse Biased.

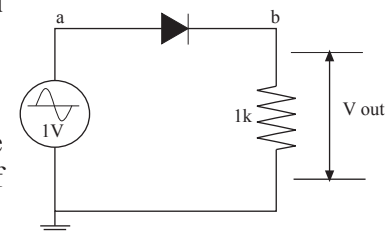


Figure 9-3 Ideal Rectifier

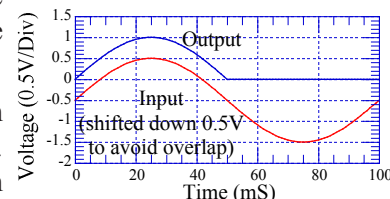


Figure 9-4 Ideal Rectifier Output

Note The input has been moved down for clarity. It really starts at the same level as input. This is the best way to show such signals on an oscilloscope. It is much easier to see the relationship between signals if they don't lie one on top of another.

9.3 The Real Diode

The ideal diode has no physical existence. However, there are real diodes that approximate the behavior of an ideal diode. As we shall see, there are a wide variety of real diodes available and their characteristics vary somewhat. Figure 9-5 shows the I-V diagram for a typical rectifier diode, a 1N4001, on the same graph as the ideal diode.

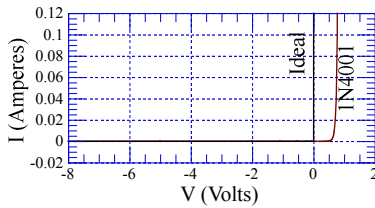


Figure 9-5 1N4001 I-V Curve

The major difference is easy to see. The real diode current does not rise at 0V but at some higher voltage, a little less than 1V. Moreover, the current does not turn on so sharply in the real diode. The ideal diode allows current to flow in the forward direction without any voltage drop across the diode. By contrast the real diode has a small voltage drop. The value depends strongly on the amount of current flowing through the diode. When the voltage across the diode is less than about 0.5V very little current flows (a few μA). Then current starts to flow until, by about 0.6V we say the diode has turned on and the current goes up *very* sharply, reaching hundreds of milliamperes by 0.8V.

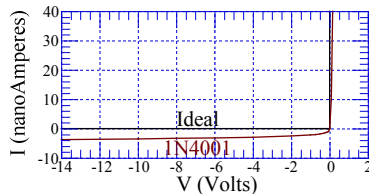


Figure 9-6 1N4001 leakage current

If we magnify the current scale enormously, by a factor of 1 billion, then we can see another difference (Figure 9-6).

The ideal diode does not conduct at all when the voltage is reversed. The real diode has a very small reverse current. This small reverse current is called **leakage current**, by analogy with a leaky valve in a plumbing system. Its value depends very strongly on temperature, which means that diodes are sometimes used to measure temperature.

In an expanded view of the forward bias situation (Figure 9-7) we see that very little current flows until the voltage across the diode gets to about 0.6V and then the current starts to rise. From this point on, the current increases extremely rapidly with increasing voltage—the slope of the I-V curve is very high so the slope resistance is very small.

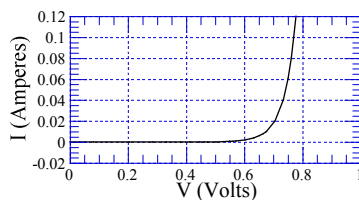


Figure 9-7 1N4001 turn-on voltage

Example

The data used to plot Figure 9-7 show that the current increases from 0.1A to 0.13A as the voltage increases from 0.770V to 0.782V so the forward slope resistance is

$$r = \frac{0.782 - 0.770}{0.13 - 0.10} = \frac{0.012}{0.03} = 0.4 \Omega$$

When the diode voltage is below the turn-on voltage (including reverse voltages), the I-V curve is essentially flat. Thus the resistance in the reverse direction is extremely high. This allows the diode to be used as a voltage controlled switch in some rather clever circuits.

Example

The data for Figure 9-6 show that the leakage current changes by 0.3nA when the voltage changes from 10V to 15V, so the reverse slope resistance is

$$R = \frac{5}{3 \times 10^{10}} = 17 \text{ G}\Omega$$

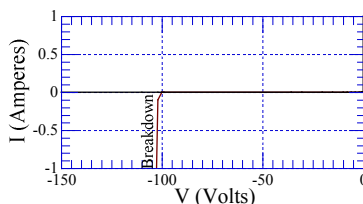


Figure 9-8 1N4001 reverse breakdown

Another difference between a real diode and an ideal diode can be seen if we extend the voltage scale in the reverse direction (Figure 9-8).

The ideal diode never allows any current to flow in the reverse direction, no matter how large a reverse voltage is applied. A real diode is not so obliging and eventually, if you apply enough voltage in the reverse direction, it will **break down** and start to conduct. The breakdown is very sharp; over a very small voltage range the current goes from near zero to several amps. This usually results in the total destruction of the diode as the power dissipated becomes enormous and the material from which the diode is made melts or even vaporizes! .

Note Breakdown itself does not harm a diode so long as the current is kept in check by the rest of the circuit. This is exploited in Zener diodes which are used to regulate the voltage across them

Example

For example, a 1N4001 breaks down at about 100V reverse bias. In a power supply circuit that can supply 1A of current, when breakdown occurs the power dissipated in the diode would rise to

$$P = 100V \cdot 1A = 100W.$$

That is enough to light a bright incandescent bulb and will destroy the diode in a few thousandths of a second!

Not only reverse breakdown destroys diodes. Too much forward current can be just as deadly. Each diode has a forward current rating that tells you how much current it can pass before it will suffer. For the 1N4001 this is 1A, but diodes are available for different uses with forward current ratings from a few mA for fast signal diodes to thousands of amperes for ultra-high power rectifiers.

9.3.1 The Real Half-wave Rectifier.

If we take our half-wave rectifier circuit and replace the ideal diode by a real diode then we get the circuit of Figure 9-9.

We can use the I-V curve for this diode (Figure 9-5 above) to deduce the output voltage for each value of the input voltage. There we see that, for low currents, the real turn-on voltage of the diode is close to 0.6V. This is the standard figure used for all silicon junction diodes, the most common kind of diode.

For voltages below 0.6V the diode does not conduct to any significant extent. Thus the voltage across the resistor stays at approximately zero volts and the voltage across the diode is just the input voltage.

For input voltages above 0.6V the diode curve goes up so sharply that we can treat as being vertical. That means that the output voltage $V_{out} = V_{in} - 0.6V$.

As we see in Figure 9-10, the real diode rectifier does not produce quite so much output voltage as the ideal diode rectifier. Instead of getting out all of the positive voltage, we only get output when the input exceeds 0.6V. If we want to have a half-wave rectifier that produces a 1V peak output voltage, then we have to use an input voltage of $\pm 1.6V$.

The half wave rectifier is a wasteful circuit since we get voltage out of it less than half of the time. We can do better than this by using more diodes, as we shall see in the next chapter, but we cannot eliminate the forward voltage drop in a rectifier that uses only diodes (but see Chapter 21).

9.4 Some Common Diode Circuits

This section includes a few of the most common circuit configurations for diodes. We shall meet a number of other uses in later chapters but these are some of the simplest.

9.4.1 The Diode Detector

One classic use of the small signal diode is as a detector for an amplitude modulated (AM) radio wave. This is a radio frequency (RF) signal, usually between 0.5MHz and 1.5MHz, whose amplitude or size is made to increase or decrease with the shape of a low frequency, audio, signal. Figure 9-11 shows a small portion of an audio frequency signal, in this case a sine wave at 2kHz, a high-pitched whistle.

Figure 9-12 shows the amplitude modulated radio wave that carries that signal. The radio frequency that I have used in this figure is much lower than would be used in practice, only about 40kHz. I have chosen this low frequency so that you can see the individual cycles of the RF frequency. In reality, the radio frequency would be so much higher than the audio frequency that the individual cycles would blur together into a solid color.

A typical AM radio signal propagates through the air quite well and can be collected and amplified by a radio receiver, basically a very narrow-band filter with a high-gain amplifier to select one particular signal out of the air and increase its size. However, in order to hear the audio information it is necessary to extract the audio signal from the radio signal with a **detector**. The most common detector is a half-wave diode rectifier followed by a low-pass filter.

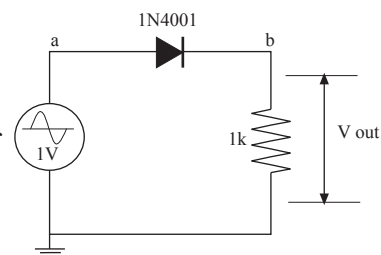


Figure 9-9 Real rectifier

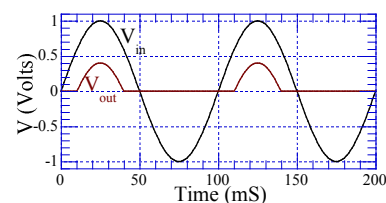


Figure 9-10 Real rectifier output

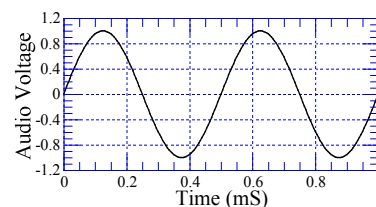


Figure 9-11 Audio Signal

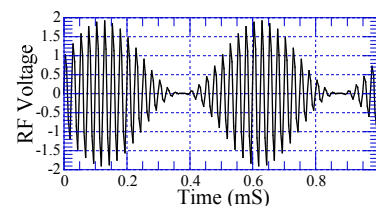


Figure 9-12 Modulated RF Signal

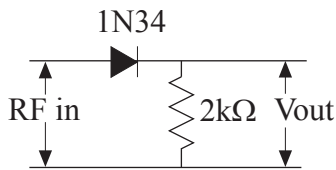


Figure 9-13 Basic Detector

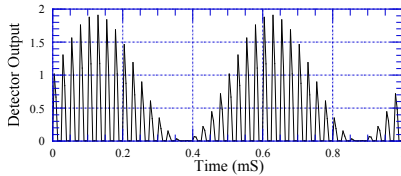


Figure 9-14 Rectified RF signal

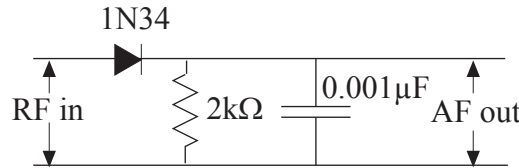


Figure 9-15 Filtered detector

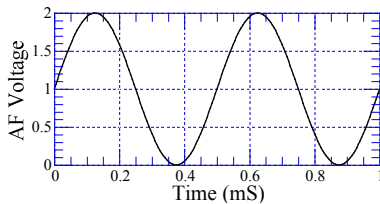


Figure 9-16 Detected Audio Signal

We start with the simple rectifier shown in Figure 9-13. The detector diode only conducts on the positive half-cycles, producing the output shown in Figure 9-14 below left.

This is still a radio-frequency signal at a much higher frequency than the audio signal. If we now add a capacitor in parallel with the 2k resistor, (Figure 9-15), the capacitor and resistor form a simple low pass filter and we have our complete detector. The filter removes out the radio-frequency part of the signal and only lets through the much slower audio-frequency (AF) voltage.

The result is the signal shown in Figure 9-16; a signal that looks very like the audio signal with which we started. The only difference is the DC level. Where the original signal went from -1V to +1V, the recovered signal goes from 0V to +2V; it has a +1V DC offset.

We can remove that DC offset by passing the signal through a **coupling** or **DC blocking** capacitor. This is a capacitor that is large enough to pass all the audio frequencies but which will not let through the DC offset. The result is a perfect copy of the original signal.

9.5 Diode Characteristics

We have already met most of the properties that characterize a real diode. Here is a more complete list, along with the values for some commonly encountered diodes (Table 9-1).

I_F , Forward current, continuous. The maximum current that the diode can pass in the forward direction for a long time. This is limited by the ruggedness of the diode's construction and by the amount of heat that the package can dissipate. Note that high power diodes may need to be cooled in some way to reach their rated current. We normally choose a diode with a forward current rating that is larger than the largest current that we expect to pass through the diode. For example, we might use a 1.5A diode in a circuit that will normally draw 1A.

I_F , Forward current, transient. The maximum current that the diode can pass in a very short burst. This is much higher than the continuous current since a transient does not have time to heat the diode significantly. It is limited only by the ruggedness of the diode.

Power dissipation. The amount of power that the diode can safely dissipate. This is determined both by the construction of diode chip and by the package around that chip. Small signal diodes can handle only a fraction of a Watt. Rectifier diodes are more robust and dissipate a Watt or two without external heatsinking. Large power diodes can dissipate enormous powers but require extreme cooling methods to reach their full power ratings

V_F , Forward voltage drop. The forward voltage drop across the diode when it is conducting. As we have seen, it is a somewhat imprecise quantity. The turn-on is not perfectly sharp but it is sharp enough to make it a very useful quantity. The value depends on the materials from which the diode is constructed.

At small signal levels, that is, currents of a few mA, all silicon p-n diodes have a forward voltage drop of about 0.6V while germanium diodes and Schottky diodes have a turn on voltage of about 0.3V. The forward voltage rises at higher currents so power rectifiers like the 1N4001 and 1N5817 have significantly higher operating forward voltage drops

V_R , Reverse breakdown voltage. The maximum reverse voltage that the diode can withstand without problems. You can usually put a somewhat larger voltage across the diode without problems but this is the maximum that the manufacturer guarantees the diode will survive. We choose this to be greater than the largest voltage that we expect the diode to see.

I_R , Reverse leakage current. The current that flows when the diode is reverse biased. This is quite constant regardless of the reverse bias. It should be very low compared to other currents in the circuit. In a high-power circuit this can be quite high without causing problems but in small signal circuit it must be very small.

C_o , Junction capacitance. Because of the way a diode is constructed, it acts as if there is a capacitor in parallel with it. This capacitor stores charge and affects how fast the voltage across the diode can change. High power diodes have large capacitances and so are somewhat slow to change voltage. Such a diode is useless at high frequencies. Small signal diodes can be made with very low capacitances so that they can be used up to hundreds or even thousands of MHz.

t_r , Switching time. The time it takes the diode to go from non-conducting to conducting as the voltage across it changes from reverse to forward-biased. Like C_o , this affects the maximum frequency at which the diode is useful as a rectifier. If we are trying to rectify a signal that is changing polarity every $0.1\mu\text{s}$, then we need to have a diode with a switching time that is much shorter than that. A 1N914 would be fine but a 1N4001 would not.

**Table 9-1: Some common diodes
1N914 fast, small signal, 1N4001 rectifier,
1N5817 ultra-fast (Schottky) rectifier**

	1N914	1N4001	1N5817
I_F (A)	0.2	1.0	1.0
I_p (A)	.45	0	25
I_R (μA)	3	10	1000
V_F (V)	0.62@5mA	1@1A	0.45@1A
V_R (V)	75	50	25
C_o (pF)	4	15	200
t_r (nS)	4	3500	0

9.6 Special Kinds of Diode

Depending on the details of their construction, diodes can be fabricated with a wide range of characteristics. We can group them into categories by their intended function.

9.6.1 Signal Diodes

The commonest kind of diode that we meet outside of power supplies is the small signal diode. As its name implies, it is designed to handle only very low currents, usually a few mA. Small signal diodes are chiefly used as switches and in diode detectors. They have small junction capacitances and very short switching times so that they can be used up to very high frequencies.

9.6.2 Power diodes

These are the workhorses of the diode race. They are used as rectifiers in all kinds of power circuits. The most familiar are small ones like the 1N400x family that can pass 1A of continuous current and which have reverse voltages of from 100V to 1000V. Higher currents and higher inverse voltages are available as are sets of diodes already connected up in various configurations such as the bridge rectifier (see chapter 11).

9.6.3 Varactor diodes

We have seen that one of the characteristics of a diode is its junction capacitance. When the diode is forward biased, the capacitance is hidden by the conduction but when the diode is reverse biased it behaves as a capacitor with a slight leak. The interesting thing is that the value of the capacitance depends on the reverse bias; the larger the bias the smaller the capacitance. This means that a diode can be used as a voltage controlled capacitance. Any diode will work in this way but there are a number of diodes that are constructed to enhance this effect and to have well-controlled capacitances. These are called **varactor** diodes and they are used in a

lot of tunable high frequency circuits. In particular, they are used for as the tuning element in electronically tuned radios.

9.6.4 Photodiodes

When a diode is reverse biased, only a tiny leakage current flows. The size of leakage current increases slightly as the temperature rises but it increases dramatically if light falls on it. Any diode is somewhat sensitive to light but there are some diodes that are constructed specially as light sensors. They are called **photodiodes** and they are discussed at some length in chapter 29.

9.6.5 Light-Emitting diodes

These are some of the most familiar diodes. An LED is diode that emits some of its waste energy as light rather than heat. LEDs are very efficient light sources, emitting far more of their total power dissipation as light than an incandescent bulb does. They are available in a variety of wavelengths from the infra-red up into the blue. LEDs are used in many situations where we need a small indicator light that is energy efficient and easy to turn on and off electronically. Unlike normal silicon diodes, with their 0.6V turn-on voltage, LEDs have high turn on voltages, typically around 2-3V for visible diodes. They usually operate on currents of 5-20mA.

9.6.6 Laser Diodes

The ultimate development of the light-emitting diode is the laser diode. Semiconductor manufacturers have managed to integrate light-emitting diodes with high efficiency mirrors to make a device that emits coherent visible light, laser light. Diode lasers are available in a much more limited range of wavelengths than regular LED. The most common types operate in the near infra-red and are commonly found in CD players. More recently red diode lasers have become widely available. These are found in DVD players and in the common laser pointers. Shorter wavelength diode lasers are under development with blue lasers built from Gallium Nitride starting to appear.

9.6.7 Zener Diodes

All diodes suffer from reverse breakdown if you put enough voltage across them. Zener diodes are specially processed to make the reverse breakdown voltage very constant and very repeatable from one unit to the next. They are designed as voltage regulators. The reverse breakdown voltage is extremely abrupt. Once the diode has reached breakdown it will conduct quite freely while holding a very constant voltage across its terminals. Zener diodes are available with breakdown voltages from about 2V to 75V and with current ratings from 5mA to more than 50mA.

Zener Diode Voltage Regulator

If we connect up a Zener diode in series with a resistor, to limit the current through the diode so that it doesn't burn up, we can build a simple voltage regulator (Figure 9-17). Such a voltage regulator can only deliver a small current because the diode must be able to carry the full load current.

The Zener regulator works because the diode passes as much current as is needed to keep its terminal voltage equal to its breakdown voltage. As the load draws more current, the diode draws less current and the voltage stays constant.

When we design a Zener diode voltage regulator we have to make several choices. First we must choose an output voltage, V_z , for which a Zener is available. For example, see ?????? for a table of common Zener voltages.

Second we must determine the power that the Zener must dissipate and make sure that a suitable diode is available. The power is determined by the largest current that we wish to draw from the regulator, I_{Max} . When the load current is zero, all of I_{Max} flows through the Zener so that the power dissipated is a maximum, P_{Max} .

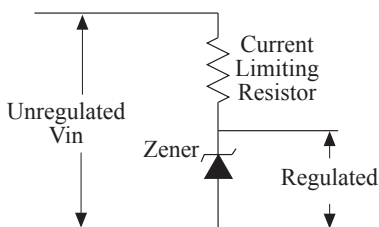


Figure 9-17 Zener diode voltage regulator

$$P_{Max} = V_Z \times I_{Max}$$

We must choose a Zener diode which is rated to dissipate at least that much current. If this is not possible (or too expensive) then we need to use a more elaborate regulator such as those described in Chapter 26.

Finally we must choose the series resistor. The current through this resistor is constant at I_{Max} so long as the regulator is working. Thus, if the input voltage is V_{in} and the output voltage is V_Z then

$$R = \frac{V_{In} - V_Z}{I_{Max}}$$

The Zener regulator is not a very good voltage regulator.

1) As the current drawn from the regulator varies, the current flowing in the diode varies. Because the reverse breakdown of a diode is not perfectly sharp this makes the output voltage vary. A Zener diode data sheet will normally specify this by giving the output resistance of the diode, R_{Out} . This allows us to calculate how much the output voltage will vary for a given change in the output current.

$$\text{Change in V out} = R_{Out} \times \text{Change in I out}$$

In this situation the regulator acts like a voltage source with an internal resistance (Thevenin resistance) approximately equal to the R_{out} for the diode.

2) If the load tries to draw too much current then the diode current drops to zero. At this point regulation stops, and the output voltage falls drastically. Now the system acts like a voltage source with output resistance equal to the current limiting resistor.

We shall learn how to build much better regulators that can avoid these problems in Chapter 26.

9.7 Using PSpice to Study a Diode

The Student Edition of PSpice comes with models for four different diodes. These models have names that start with D and then have the standard part number. There are models for the 1N914 and 1N4148 small-signal diodes (for rectifiers and logic elements), the 1N750 Zener diode, and the 1N4002 rectifier diode. This is a slightly higher reverse-breakdown voltage version of our favorite 1N4001 and it makes a good place to start.

9.7.1 The forward-biased diode characteristic

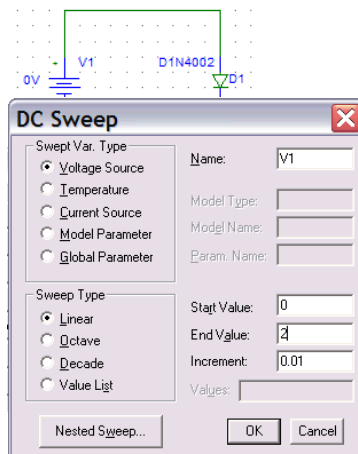
Spice is a great way to examine component characteristics. It simulates the electrical behavior of the device without any of the real-world side-effects such as overheating. It can simulate the diode up to currents that would destroy the real-world device if allowed to persist for more than a tiny fraction of second—too short for us to make practical measurements.

We measure the characteristic curve by applying a voltage to the diode and measuring the resulting current flow. In the lab we have to do this with voltmeters and ammeters. In PSpice we just tell the plotting routine what values to graph and do not have to model the instruments at all.

1) Get a D1N4002 diode model and a VDC voltage source from the Part Browser and construct the circuit on the right. Note that this is a circuit that we could never use in real life because there is no resistor to limit the current flowing through the diode and we should destroy it very quickly as we raised the forward bias.

It would be possible to perform this analysis using the Transient Analysis mode with a suitable linear ramp voltage but there is a better way. Spice provides a DC Sweep analysis which changes a single DC voltage source according to a pre-determined rule. We shall use that analysis.

2) Bring up the Analysis Setup dialog and click on the DC Sweep button to enable the analysis and to bring up the DC Sweep dialog. This is another complex dialog box but again we only need some of it.



We want to vary the voltage of the part named V1 so we check the Voltage Source and enter the Name V1. As you can see, the sweep can be linear, logarithmic (octave or decade), or can step through an arbitrary list of values. We want a linear sweep and I have decided to explore from 0V to 2V, hugely more than we have ever seen across a diode! For a sweep we must specify an increment to control the smoothness of the output plots. I have selected 100 points per volt.

3) Once you have dismissed the DC Sweep and Analysis Setup dialogs, run the simulation. You will, of course, get a blank graph since we have not told it to plot anything.

4) Choose Add Trace and select I(D1), the current in D1, as your Trace Expression. This will give you a graph like this.

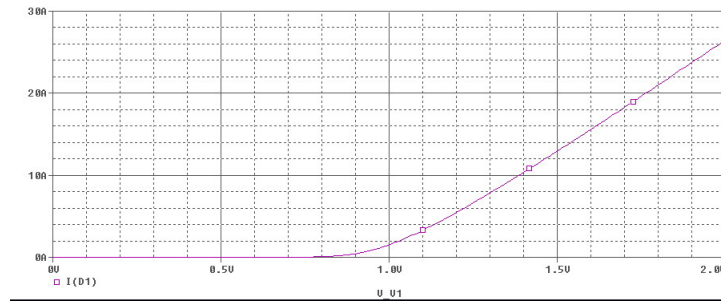


Figure 9-18

There are several remarkable things about this curve. First it does not seem to start until about 0.8V, way above the standard 0.6V diode voltage. Second, it reaches nearly 30A. This is FAR more current than you could push through this diode in real life. With 2V across the diode and 30A flowing through it, the diode would dissipate 60W, as much as a light bulb!

3) If we change the scale on the graph we can see the normal operating region in more detail and see that the diode does indeed behave as we would expect. Double-click on one of the x axis labels and reset the scale to run from 0V to 1V. You should get a picture like this.

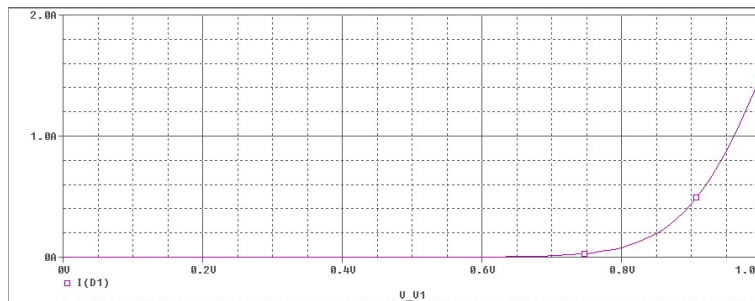


Figure 9-19 Diode Turn-on

This looks a lot more normal. Now we see the curve start upwards at 0.6V and reaching a little above 1A, the maximum normal operating current for the device.

9.7.2 The reverse-biased diode

Let us go back and re-do the sweep with negative voltages to see if we can see the reverse leakage current.

1) Go back to the DC Sweep dialog and set the sweep to run from -110V to 0V with an increment of 0.1V.

2) Rerun the analysis and play the same Add Trace trick to get a reverse-bias plot that should look like this.

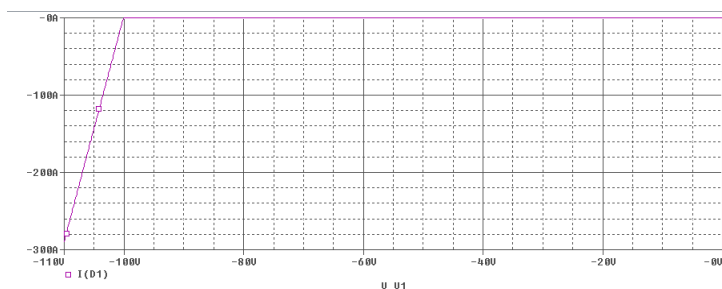


Figure 9-20 Reverse Biased Diode Breakdown

Clearly the diode breaks down catastrophically at -100V. Once it does so it acts like an extremely small resistor and huge currents flow. This is why exceeding the reverse breakdown voltage of a diode normally destroys the diode practically instantly.

3) We can see the reverse leakage current if we change the scale to -80V to 0V.

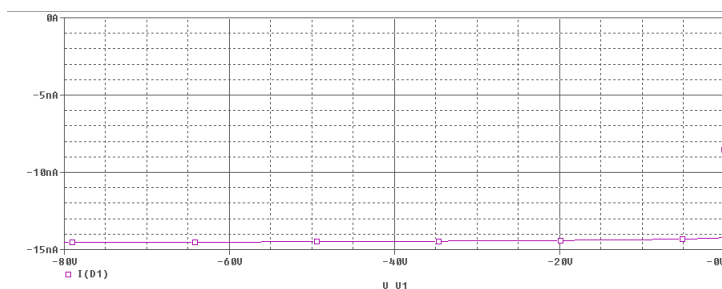


Figure 9-21 Diode Reverse Leakage

Here we see the reverse leakage current very quickly fall from 0 to its steady value of about 15nA and can see why we could not measure the current in the lab, where our instruments only work down to the μV level.

You could also play with the scale of the graph to look at the reverse-breakdown point in more detail or to look at the region near 0V where the leakage turns on.

9.7.3 Modeling the Half-Wave Rectifier

If we add a resistor to the circuit and change the source to an AC source then we can look at the behavior of the half-wave rectifier.

- 1) Either make a new schematic or modify your diode curve schematic. Add a small resistor and replace the DC voltage source with a VSIN sine wave source. Add voltage markers as shown.
- 2) Double-click on the VSIN source to bring up its parameter dialog. Set the amplitude, offset and frequency values. I chose to look at a 20V 60Hz sine wave with zero offset.
- 3) Bring up the Analysis Setup dialog and select a Transient Analysis. I chose to run the analysis for 0.02s (slightly over 1 period) and to collect data at 0.1 μs intervals.
- 4) Run the analysis to get an output. Mine looked like this.

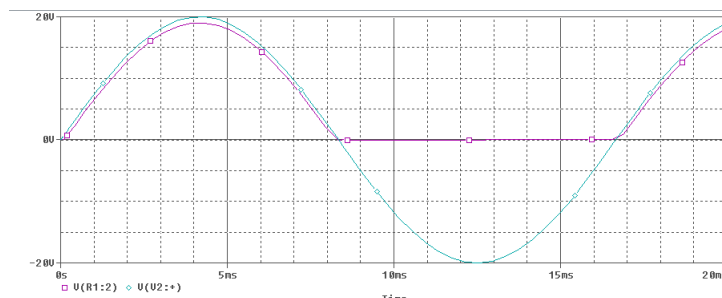
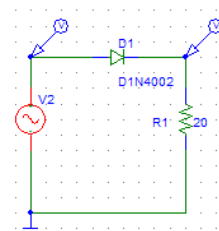


Figure 9-22 Half-wave Rectifier Output

The output can be seen following the input in during the positive half cycles and staying zero during the negative ones. The output does not turn on instantly when the input goes positive but trails by a small amount as the roughly 0.6V diode turn-on has to be crossed first. Once the output turns on it stays about 0.5-1V lower than the input as we expect.

9.8 The Physics of Diodes

Back in chapter 2 we looked at the mechanism of conduction and met those strange materials, the semiconductors. Now we shall examine how putting two pieces of semiconductor together can produce a diode.

9.8.1 The p-n junction

A piece of n-type semiconductor contains donor impurities that insert some extra electrons into the crystal, electrons that are then free to move through the crystal and carrying a current. A piece of p-type semiconductor contains acceptor impurities. It conducts using the free holes introduced by the acceptor atoms. If we put a piece of n-type material in contact with a piece of p-type material then something interesting happens (Figure 9-23). At the junction free electrons drift from the n-type material into the p-type material and there encounter the free holes. The new electrons fill up those holes and a region develops where there are no free holes and no free electrons. This is called the **depletion zone**.

Info At first glance, it seems as though this process should just continue until the depletion zone fills the whole semiconductor as all of the extra electrons migrate from the n-type material to the p-type and fill up all the holes. However, this cannot happen because of the charges left behind. Remember that each extra electron in the n-type material comes from an impurity that also has an extra positive charge. Similarly, each hole in the p-type material comes from an impurity that has one positive charge too few. When the electrons and holes get together in the depletion zone, they leave behind their charged impurity atoms. Thus the n-type material becomes positively charged, attracting its remaining electrons more strongly, while the p-type material becomes negatively charged, attracting its remaining holes. These charges prevent the remaining electrons and holes from migrating and so the depletion zone remains as a narrow band either side of the junction. We are left with a device that has conductive ends separated by an insulating depletion zone.

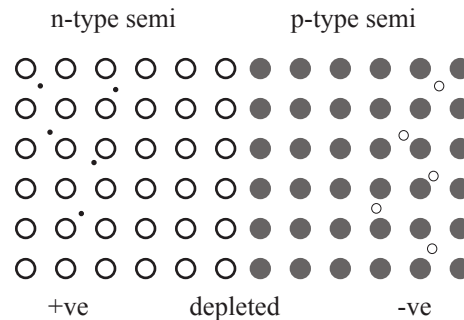


Figure 9-23 p-n junction

If we now make connections to the outside ends of this device, we can apply an external voltage to the device (Figure 9-24). What happens next depends on the direction in which we apply the voltage. If we make the n-type end positive and the p-type end negative, then the applied voltage exaggerates the natural effect, pulling electrons further into the n-type material and holes further into the p-type material. The depletion zone gets wider and no current flows.

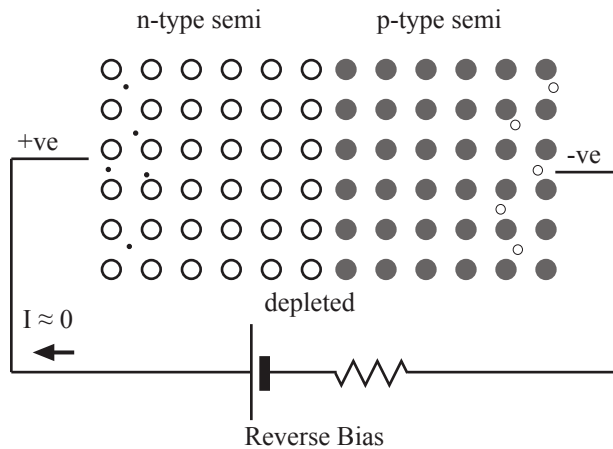


Figure 9-24 Reverse Biased p-n Junction

If we reverse the external potential, making the n-type end slightly negative and the p-type end slightly positive, then the situation is very different (Figure 9-25). Now the applied voltage acts to counter the natural charges. The electrons in the n-type material are pushed away from the

Note The reverse current is not exactly zero. The depletion zone behaves exactly like a piece of intrinsic semiconductor and so there are a few thermally excited electrons and holes around to carry a current. This tiny thermal current is the leakage current of the diode.

metal contact, towards the p-n junction. At the same time, the holes are pushed away from the metal contact on the p side towards the p-n junction so the depletion zone gets narrower. There is still no current flow. The depletion zone is narrower but it is still there.

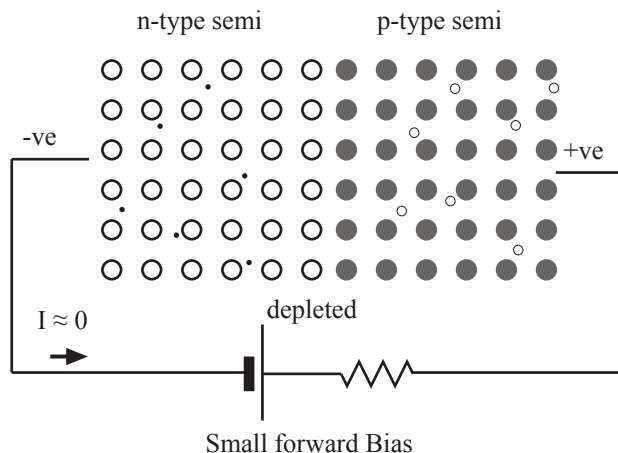


Figure 9-25 Forward Biased p-n Junction

When we increase the forward bias, the depletion zone gets narrower and narrower until finally it disappears. At this bias, about 0.5V in silicon, electrons from the n-type region can meet holes from the p-type material and consume each other. A small current flows; the diode has started to conduct. The depletion zone has now been replaced by a zone where electrons and holes are combining with each other; the **recombination zone** (Figure 9-26). In the n-type semiconductor the current is carried by electrons moving from left to right. In the p-type semiconductor the same current is carried by holes moving from right to left. The electrons and holes meet at the recombination zone and combine releasing energy, mostly as heat. As the electrons and holes consume each other, they make room for more electrons and holes to flow into the region and a steady current can flow.

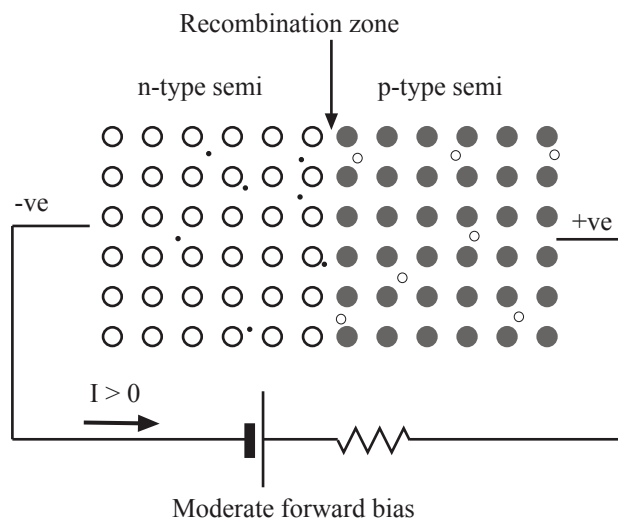


Figure 9-26 p-n Junction with more Forward Bias

As we increase the forward bias still further, the current rises extremely steeply. You cannot exceed the turn-on voltage by more than a few tenths of a volt without destroying the diode because of the amount of heat that is liberated. Almost all of the heat is given off as the electrons and holes recombine so that the heating is concentrated in the recombination zone. It is easy to get this region so hot that the diode destroys itself.

So we have seen how a p-n junction has a natural asymmetry. Left to itself it sets up a depletion zone across which no current can flow. If we apply a bias in the reverse direction then the depletion zone just gets wider; still no current flows. If we apply a forward bias, then the

Note In an LED a lot of the energy released by the recombination appears as light rather than heat. This light has a color which is characteristic of the energy given up when the electron and hole recombine. Thus each type of diode emits a characteristic color. We get the different colors by using different materials for the diodes.

depletion zone gets narrower until it disappears and current starts to flow. The I-V curve turns suddenly upwards and the diode turns on, allowing current to flow freely.

9.9 Manufacturing a diode

The very first diodes were made by trial, error, and luck. The p-n junction was formed by pressing a fine metal point, called a **cat's whisker**, against a lump of impure semiconductor and hoping to form a p-n junction. The material most often used was the naturally occurring ore of lead called Galena. These cat's whisker diodes were used in some of the earliest commercial radio receivers and people used to have sit by the radio, wearing heavy headphones, and playing with the setting of the cat's whisker to get a good signal. Because the detector was a crystal of Galena these radios were called **Crystal Sets**.

It was soon found that the p-n junction was the important part and manufacturers started making reproducible diodes and radio became a much more reliable means of communication. Today a diode is made by taking a piece of pure silicon and processing it carefully. The piece of pure silicon is put in an oven and exposed to low-pressure of n-type and p-type dopant atoms (Figure 9-27). In the high temperature of the oven, some of the dopant atoms stick to the silicon surface and diffuse into the metal forming an n-type or p-type layer on top of the existing silicon.

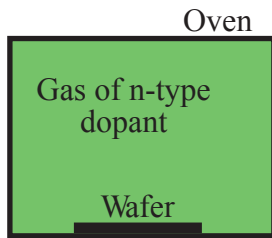


Figure 9-27

Diodes are not made one at a time. Instead, a large, thin disk of silicon, called a **wafer**, is processed all at once. First, it is doped with an n-type impurity and then a thin layer of light sensitive plastic, a **photoresist**, is poured onto the surface (Figure 9-28).

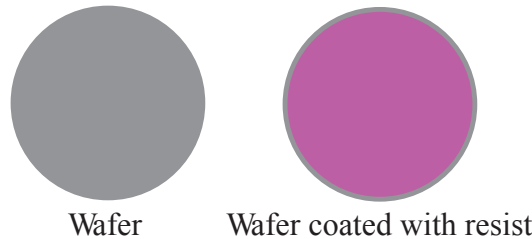


Figure 9-28 Wafer Processing 1

Then the resist coated wafer is exposed to bright UV light through a mask, which lets light fall on some parts of the surface and not on others. The mask for a set of diodes is a very simple one. It is mostly clear with an array of little black dots on it. The plastic resist hardens where the light falls and stays soft where the light was blocked. Next, the wafer is washed in a solvent and the soft parts of the resist wash away. The hardened parts remain, leaving a pattern of resist and holes covering the surface of the wafer (Figure 9-29).

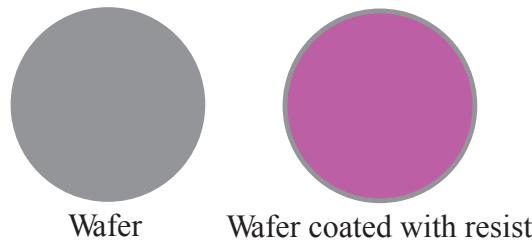


Figure 9-29 Wafer Processing 2

This resist masked wafer is put back into the oven and exposed to the gas of p-type dopant. The resist blocks the gas from the silicon surface underneath it but the holes let the gas through. In those places where the wafer is exposed, the p-type dopant diffuses into the surface and forms a p-type well in the n-type background with a nice p-n junction at the boundary. Once the p-type wells are formed, the remaining resist is removed and the wafer is cut up into individual chips, which are then mounted in packages for sale. Figure 9-30 shows the process as seen in a cross-section of the wafer at high magnification.

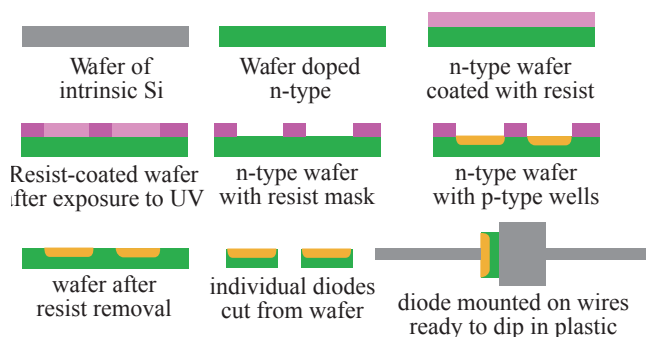


Figure 9-30 Stages in Making a Diode

The last stage is to mount the individual diode chip onto a metal header and attach leads to the two terminals and then to encapsulate the whole thing. Small signal diodes are often encapsulated in glass but higher power diodes usually use plastic packages and the highest power diodes use metal cases with glass seals. In these the diode chip is bonded to the metal case to facilitate heat removal. The highest power diodes can be quite large, about the size of a hockey puck, and require water cooling to handle their full rated currents of thousands of amperes.

The diode manufacturer can control the properties of the diode by adjusting the dopants and the way that the doping is done. This can lead to thinner or thicker wells, more or less heavy doping, and, of course, larger or smaller diodes, depending on the mask that was used.

Summary

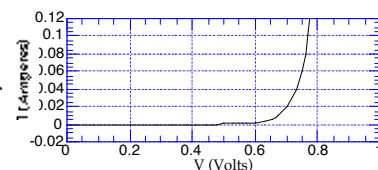
A diode is an electronic one-way valve. Its symbol is designed to show that current can only flow in one direction.



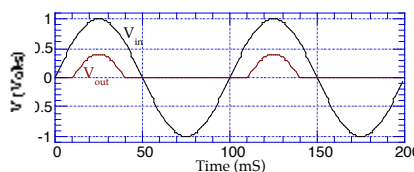
A diode is formed when there is a junction between a piece of p-type semiconductor and a piece of n-type semiconductor. The p-type material forms the anode of the diode and the n-type forms the cathode.

When a voltage is applied so that the cathode is more positive than the anode, no current flows. We say that the diode is **reverse biased**. If the reverse bias voltage is increased enough then current will start to flow and we say that **breakdown** has occurred. Unless precautions have been taken breakdown usually destroys the diode.

When a voltage is applied in the other direction, the anode is more positive than the cathode, we say that the diode is **forward biased**. In his case, a current flows that depends strongly on the bias voltage. For bias voltages less than about 0.6V very little current flows. For biases of 0.6V and above, a lot of current flows. Indeed, a diode conducts so strongly that it maintains a rather steady 0.6V across its terminals when it is conducting. Here is a typical diode forward I-V curve.



If you apply an alternating voltage to a diode, the diode conducts only one half of the time. The resulting waveform, consisting only of half sinewaves, is said to be **rectified** and the circuit is called a **half-wave rectifier**. Here is the output of a half-wave rectifier when the input is quite small, only just large enough to turn the diode on.



A Zener diode is constructed to have a very precise breakdown voltage and is used in reverse bias as a voltage regulator. The diode must be chosen to be able to dissipate a power of at least

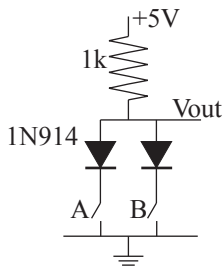
$$P_{Max} = V_Z \times I_{Max}$$

and must be put in series with a resistor of value

$$R = \frac{V_{In} - V_Z}{I_{Max}}$$

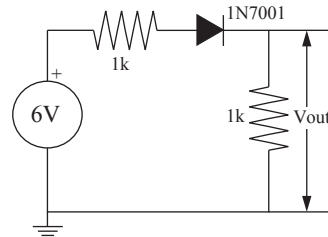
Exercises

1. Give a circuit for a half wave rectifier with a negative output voltage and explain how it works.

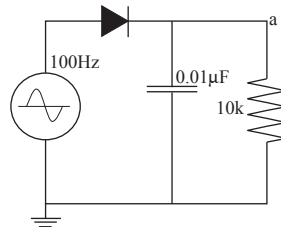


2. Using what you know about diodes, figure out the output voltage for all 4 different combinations of the two switches in the figure on the left.

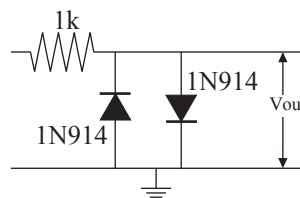
3. Find the output voltage from the circuit on the right



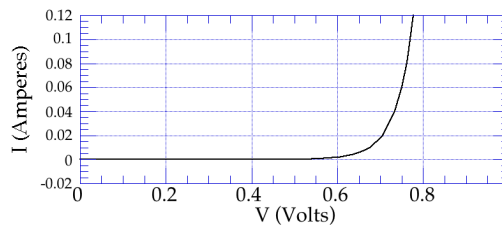
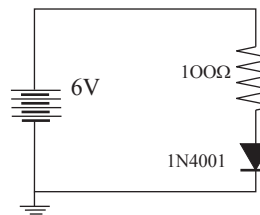
4. Use your knowledge of the behavior of capacitors and diodes to predict the voltage at point a (figure right) as a function of time. Assume that the signal source has a very low internal resistance. Repeat the calculation for Capacitors of $1\mu\text{F}$ and $100\mu\text{F}$.



5. Draw the output from the circuit on the right when it is driven by 1kHz sinewaves of amplitude 0.4V, 0.7V, and 1.0V. NOTE that the 1N914 is a fast, small-signal, silicon diode with a switching time of 4nS.



6. Use the graph below to find, as accurately as you can, the voltage at point a in the figure on the right.



Chapter 10: Power Supplies I

10.1 Introduction

All electronic circuits require a source of power to function. That source needs to provide one or more constant DC voltages each capable of supplying some predetermined amount of current. There are two ways to achieve this, batteries or mains power. If the equipment is to be portable then we normally use batteries, but they are expensive and have a limited life span. We normally operate fixed equipment from mains power through a power supply. The problem with mains current, which comes from a **generator** at a power plant, is that it provides an alternating voltage of $\pm 163\text{V}$ at 60Hz while we need a DC supply at a level such as $5\text{-}20\text{V}$. We face three hurdles; getting the voltage to the right level, converting it from alternating to direct current, and cleaning up the final voltage.

10.2 Transformers

The transformer is a device for converting one AC voltage into another. It consists of two or more coils of wire wound on a magnetic former, usually built up from thin steel strips cut into simple shapes. One coil, called the **primary**, is connected to the mains, which causes an alternating current to flow in the wires. That current sets up an alternating magnetic field, which passes through the turns of the other coil, called the **secondary**, causing a current to flow in them. The secondary current is an alternating current of the same frequency as the primary current but its voltage depends on the number of turns in the coils. The construction is reflected in the symbol for the transformer (Figure 10-1).

If the primary coil has N_1 turns and the secondary has N_2 turns, then the output voltage is

$$V_{\text{out}} = \frac{N_2}{N_1} \times V_{\text{in}}$$

This means that you can get almost any output voltage by choosing the number of turns appropriately. A transformer whose secondary has fewer turns than the primary, so that the secondary voltage is lower than the primary voltage, is called a **step-down** transformer. One whose secondary has more turns than the primary is called a **step-up** transformer. Moreover, a single transformer can have more than one secondary coil so that you can get several different output voltages from a single transformer. For example, a moderate power radio transmitter that uses vacuum tubes in its final power stages needs several power sources. It needs a high voltage for the vacuum tubes, a low voltage for the vacuum tube heaters, and a separate low voltage for the transistor circuitry in the early stages of the transmitter. You might use a transformer like the one in Figure 10-2.

Because of the law of conservation of energy, there is no way to get more power out of the transformer than goes into it. Thus, if the input current is I_{in} then the maximum output current is

$$I_{\text{out}} = \frac{N_1}{N_2} \times I_{\text{in}}$$

In fact, there are always some small losses in a real transformer so that the actual output current is always a little less than the maximum. Thus we always have

$$V_{\text{out}} \times I_{\text{out}} < V_{\text{in}} \times I_{\text{in}}$$

Small instrument transformers, those used in normal electronic devices, have efficiencies of between 80% and 90% while the large transformers used in the electrical power distribution grid can have efficiencies of up to 98%.

The limited efficiency of real transformers shows up most clearly in the behavior of the voltage as we try to draw large currents from the transformer. As the current draw increases, the peak voltage decreases. We account for this effect by ascribing to a transformer an **output impedance**. Physically this impedance has its origin in a large number of physical effects; the

Info A **generator** converts mechanical energy into electrical energy. It consists of a large coil that is rotated between the poles of a powerful magnet. The changing magnetic field pushes current through the wires and converts the mechanical energy needed to make the coil turn into electrical energy

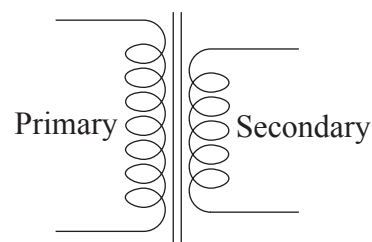


Figure 10-1 Transformer

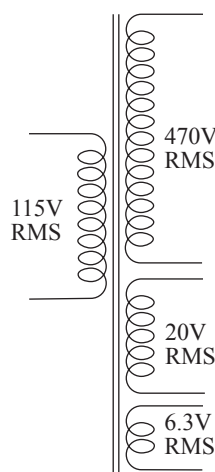


Figure 10-2 Multi-Secondary Transformer

Info Transformers are usually rated in terms of the RMS output voltage and the amount of current that can be drawn from the secondary. For example a transformer may be rated at 9V and 1A. It will provide an output of 9V RMS, so that the output varies from $-9\sqrt{2} = -12.7V$ to $+9\sqrt{2} = +12.7V$, while supplying up to 1A on a continuous basis or several times that for a short period.

actual resistance of the wire in the coils, magnetic losses in the iron core of the transformer, capacitance between the turns of the coils, etc. Practically, we can treat a transformer in a circuit by replacing it by its Thévenin equivalent, a sinusoidal voltage source of the rated output voltage in series with the Thévenin resistance that represents all the loss effects (Figure 10-3).

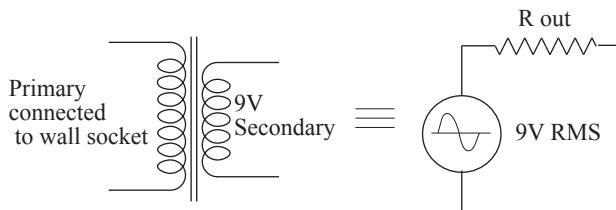


Figure 10-3 Thévenin Equivalent of a Transformer

For a typical small instrument transformer the internal resistance is usually a few Ohms. For large power transformers it can be milliohms or smaller.

10.2.1 Tapped transformers

It is possible to get several voltages from a single secondary by **tapping** the secondary. That means making a connection to one of the intermediate turns of the secondary coil, as you can see in the symbol in the figure below. Then the voltage between the bottom of the coil and the tap depends on the number of turns between the bottom and the tap while the output voltage between the first and last turns remains unaltered.

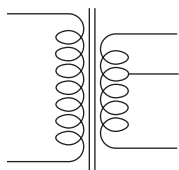


Figure 10-4 Tapped Transformer

One of the most common kinds of tapped transformer is the center-tapped transformer, which has the tap at half the number of turns as the name implies. We shall see one use for such transformers when we look at the full-wave rectifier. Another common use to generate equal positive and negative voltages from a single transformer. Such a transformer is usually described by the voltage on one side so, for example, one might have a 10-0-10V transformer that would produce $10V_{RMS}$ between the center-tap and each end tap or produce $20V_{RMS}$ between the two end taps.

Example

We can make a transformer that has output voltages of 10V RMS and 15V RMS for an input of 115V RMS if we use a transformer whose primary coil has 200 turns and a secondary with

$$15 \times \frac{200}{115} = 26 \text{ turns}$$

which we then tap at

$$10 \times \frac{200}{115} = 17 \text{ turns.}$$

That gives us a coil that has 17 turns on one side of the tap and 9 turns on the other side for a total of 26 turns. This is cheaper than using two separate coils, one of 17 turns and one of 26 turns, but the wire has to carry the current from both circuits so that it may need somewhat heavier wire.

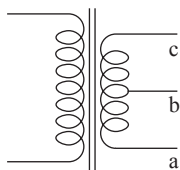


Figure 10-5 Center-Tapped Transformer

Probably the most common configuration for a tapped coil is the **centre-tapped** coil, which has the same number of turns on each side of the tap (Figure 10-5). This produces a secondary that is ideal for either dual-polarity power supplies or power supplies that use a full-wave rectifier as described below. Such a transformer secondary is usually called, for example, a 9-0-9V secondary. That would mean a single secondary that produced a total of 18V with a tap at 9V. If we look at the voltages on various pairs of terminals then we can see how the same coil can be used as either 9-0-9V or as 0-9-18V. If we label the terminals as shown in Figure 10-5, then we get the output voltages of Figure 10-6. I have plotted the outputs choosing the center terminal, b, as the zero of voltage. That way we can see that the voltage at point a is always exactly the negative of the voltage at point c. When the voltage at point c is positive, that at point a is negative and vice versa. The end result is that the total voltage from a to c is always twice the voltage from a to b or from c to b.

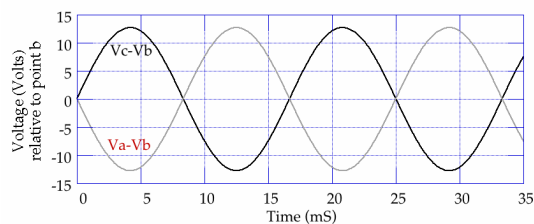


Figure 10-6 Center-tapped Output

We shall see why this is important in the next section.

10.3 Rectifiers

Once we have the correct voltage, we need to convert it to direct current. We have already seen one circuit to do this, the half-wave rectifier of Figure 10-7.

As we saw before, the half-wave rectifier produces the intermittent output voltage of Figure 10-8 so it is rather hard to make this into a smooth DC voltage.

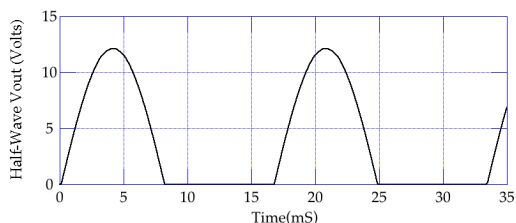


Figure 10-8 Half-Wave Rectifier Output

10.3.1 The Full-Wave Rectifier

A better solution is to use a **full-wave** rectifier, which produces an output during both halves of the cycle. There are two ways to achieve this. The older one uses a center-tapped transformer and two diodes connected so that one half of the transformer produces output on one half cycle and the other half of the transformer produces output on the other half cycle (Figure 10-9).

During the half cycle when the voltage at a is positive with respect to that at b, the lower diode conducts and the voltage across the load resistance is just $V_a - V_b - 0.6V$ (the 0.6V is the diode forward voltage drop). While this is happening, the voltage at point c is negative with respect to point b so that the upper diode is reverse biased and no current flows in it.

On the next half cycle, the voltage at point a is negative with respect to point b so that the lower diode is reverse biased and does not conduct. However, the voltage at point c is now positive with respect to point b, so the upper diode does conduct and the top of the resistor is again raised to a positive voltage, $V_c - V_b - 0.6V$.

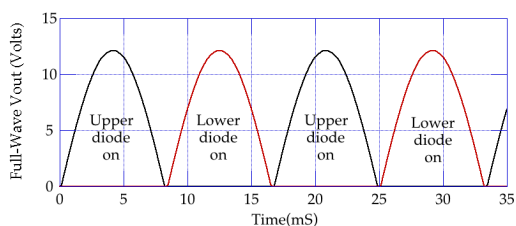


Figure 10-10 Full-Wave Rectifier Output

Thus, we get the output of Full-Wave Rectifier Output, where the output is shown in black when the upper diode is conducting and gray when the lower diode is conducting. This output is clearly a lot better than that from the half-wave rectifier since the voltage is much more continuous. It is still not a smooth DC voltage though; we will need the circuitry of the third section to reach that.

Remember If a transformer is labelled something like 9-0-9V then it is center-tapped and the total voltage between the outside ends is $9+9=18V$.

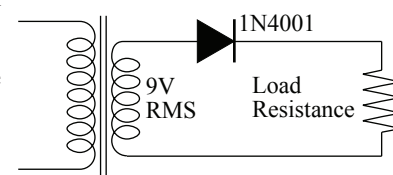


Figure 10-7 Half-Wave Rectifier

Remember The peak output voltage from the half-wave rectifier is 0.6 volts less than the peak output voltage from the transformer because of the diode forward voltage drop.

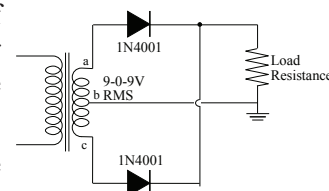


Figure 10-9 Full-Wave Rectifier Circuit

Two important facts about this circuit are not apparent at first glance. First, the fact that only one side of the transformer secondary is supplying current at any instant means that the power being drawn from the secondary is only I_{out} times the voltage across one side of the secondary. For example, if we draw 1A from our 9-0-9V secondary in the above circuit then the power drawn is only 9W, not 18W. This means that the transformer only has to handle half of its rated power.

Second, the maximum reverse voltage across each diode is the full 25.4V (= $E_2 - 18V$) rather than 12.7V because each diode is connected essentially across the whole secondary, from a to c not from a to b or c to b.

Note To be precise, the maximum reverse voltage is 25.4V - 0.6V because the two diodes are across the whole secondary and one conducts, dropping 0.6V, leaving only 24.8V across the other.

10.3.2 The Bridge Rectifier

The full-wave rectifier using a centre-tapped transformer was extremely popular when diodes were expensive. Today, the transformer is by far the most expensive component in a power supply so the **full-wave bridge** rectifier circuit shown below is much more common. This can produce a full wave rectified voltage from an untapped transformer but needs four diodes to do it. The output voltage from this is almost the same as that from the center-tapped full-wave rectifier but the mechanism is quite different.

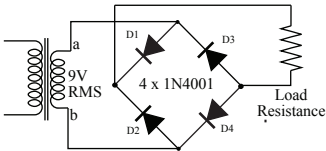


Figure 10-11 Bridge Rectifier Circuit

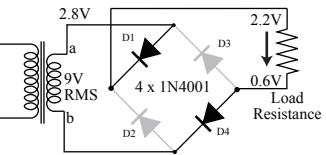


Figure 10-12

Let us follow in detail what happens as the input voltage goes through one cycle. Early in the cycle point a is positive with respect to point b. This means that diodes D1 and D4 are forward biased while D2 and D3 are reverse biased. At first nothing happens because the voltage between a and b has to overcome TWO diode forward drops before any significant current starts to flow. Once the input voltage exceeds 1.2V, current starts to flow, and voltage appears across the load. Figure 10-12 shows the situation when point a is 2.8V positive with respect to point b and makes clear the path along which the current flows.

Time passes, the voltage across the load rises to 11.5V and falls again, reaching 0V when the voltage across the secondary reaches 1.2V again. Now no diodes conduct and the output voltage stays zero. The output voltage remains zero while the secondary voltage continues to fall and reverses sign. Now that point a is negative with respect to point b, diodes D1 and D4 are reverse biased and diodes D2 and D3 are forward biased. No current flows, however, until the secondary voltage passes -1.2V. Once the voltage is large enough, current starts to flow along the path shown in Figure 10-13.

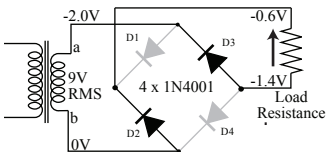


Figure 10-13

Again, the voltage rises and falls, and the current continues to flow along this path until the voltage across the transformer again falls below 1.2V. Then the diodes turn off and there is another short gap of zero volts before the whole process repeats. Current flows in the same direction through the resistor no matter which path the current is following. Thus, the voltage at the top of the resistor is always positive with respect to the bottom of the resistor. The voltages shown look a little peculiar because we have chosen to call the lower end of the transformer 0V. In practice, we usually connect the junction of diodes D3 and D4 to ground, setting it equal to 0V (Figure 10-14).

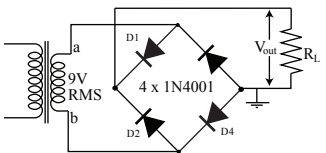


Figure 10-14

The output voltage is measured across the terminals to which the resistor is connected. As shown in Figure 10-15 it looks very similar to the output from the center-tapped full-wave rectifier.

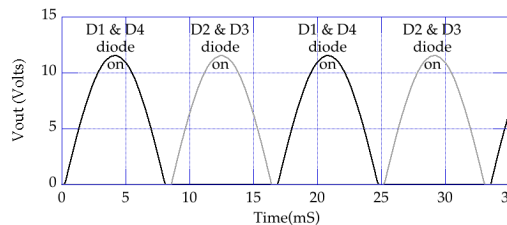


Figure 10-15 Bridge Rectifier Output

Note Because the current must flow through two diodes, the voltage across the load is 1.2V less than the transformer voltage. Thus the maximum voltage that this circuit can deliver is $E_2 - 9 - 1.2 = 11.5V$.

If we look at the first period in detail, we can see the short intervals of 0V output. These occur when the secondary voltage is smaller than 1.2V and so neither side of the bridge is conducting and there is no current flowing in the load.

The full-wave bridge rectifier has two advantages over the center-tapped full-wave rectifier. First, we can use a simpler and lighter transformer since we don't need a center tap and can use a transformer rated to carry the full load power rather than twice the load power. Second, we can use lower voltage diodes since each diode sees only the output voltage across it, one half of the voltage seen by each diode in the center-tapped circuit. The transformer does have to use slightly thicker wire in the secondary since the output current flows through the whole secondary all the time. In the simpler rectifier, only one half of the secondary carries current at any time. The output voltage of the bridge circuit is also 0.6V lower than that from the center-tapped circuit. That is easily remedied by adding a few extra turns to the secondary to make up the lost voltage.

10.4 Smoothing the output

All of the rectifier circuits that we have seen produce pulsed DC outputs not the steady outputs that we want. We have to add further components to smooth out the bumps. What is needed is a component that can store up charge during the pulses and then supply charge to keep up the output voltage during the gaps. We need a capacitor, called a **filter capacitor**. This capacitor is placed directly across the output of the power supply as shown in Figure 10-16. For a typical small power supply supplying about 10V at an amp or less we might use a capacitor of about 10,000 μ F, far larger than we shall meet in any other context.

Let's see how the capacitor works and how to choose a capacitor for a particular use. We will replace the transformer and rectifier circuit by its Thévenin equivalent, a rectified sine wave in series with a resistor of about 1 Ω , and the load by its Thévenin equivalent, a resistor. Let us work with our 9V transformer and bridge rectifier and have a load that will sink 1A at 9V so it has a resistance of 9/1 = 9 Ω . Here is the Thévenin equivalent of our filtered bridge rectifier (Figure 10-17)

Note The ideal diode is there to prevent current from flowing backward through the 1 Ω resistor and voltage source. It represents the bridge rectifier, which, in the real circuit, prevents any current from flowing backwards into the transformer.

Now let us follow the circuit through several cycles of the voltage source. We will start at the moment the circuit is turned on and current starts to flow. At this instant, the capacitor is uncharged so the output voltage is zero and current flows through the 1 Ω resistor into the capacitor and 9 Ω resistor. In order to calculate the way that the capacitor charges we must further simplify the circuit. First, we rearrange the circuit to see that the source of current charging the capacitor is a voltage divider (Figure 10-18).

It is important to realize that this circuit is identical to the circuit of Figure 10-17. Current does not go first to the capacitor and second to the resistor. They are connected together in parallel so *the order does not matter*.

When the voltage source is supplying current, as is this case at this early stage in the process, the diode is conducting and so acts as if were a wire (remember, this is an *ideal* diode, not a real one). Thus, the circuit is equivalent to Figure 10-19, where I have enclosed the circuit that supplies current to the capacitor in a dotted box.

We see that the circuit that supplies current to the capacitor is a voltage divider, which we can replace by its Thévenin equivalent (Figure 10-20). The Thévenin voltage of the equivalent has the same shape as the real voltage but the peak voltage is only $11.5 \cdot 9/10 = 10.35$ V.

The Thévenin resistance is the parallel combination of 1 Ω and 9 Ω ,

$$R_{th} = \frac{1}{\frac{1}{1} + \frac{1}{9}} = \frac{1 \times 9}{1 + 9} = 0.9\Omega.$$

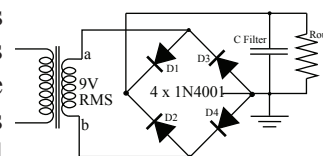


Figure 10-16 Bridge Rectifier with Filter Capacitor

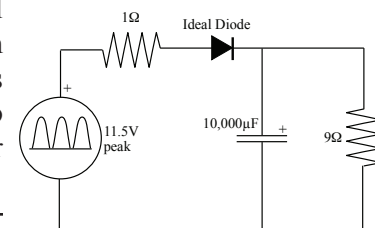


Figure 10-17 Thévenin Equivalent of Bridge

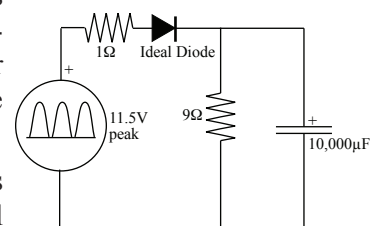


Figure 10-18

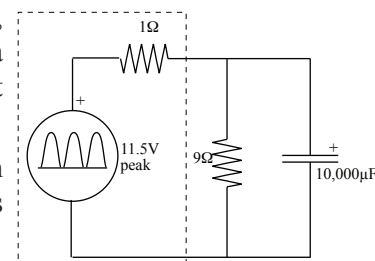


Figure 10-19

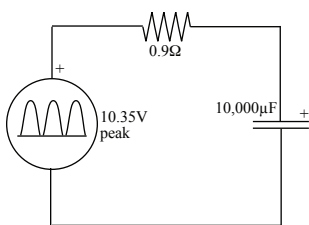


Figure 10-20

So we get the circuit of Figure 10-20. When a $10,000\mu\text{F}$ capacitor is charged from 0.9Ω , the time constant is $0.9\Omega \cdot 0.01\text{F} = 9\text{mS}$. Since the input is a 60Hz sine wave, it takes 4.2mS for the voltage to reach its peak. Thus, the capacitor only has time to charge to some fraction of 10.35V but will not get nearly all the way. The precise details of the shape of the voltage vs. time curve for the capacitor and its final voltage are tricky to compute but the result is shown in Figure 10-21.

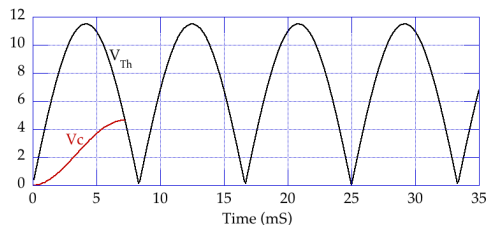


Figure 10-21

Once the sine wave passes its peak, the source voltage starts to fall. For a while it is still greater than the voltage on the capacitor so the capacitor continues to charge. Eventually, the source voltage falls below the capacitor voltage and the capacitor starts to discharge. At this point, we have to go back to the circuit to see what happens. Once the capacitor starts to discharge, current flows the opposite way and the diode turns off, preventing current from flowing back into the source. The discharge circuit is quite different from the charging circuit since the capacitor is discharging only through the 9Ω resistor (Figure 10-22).

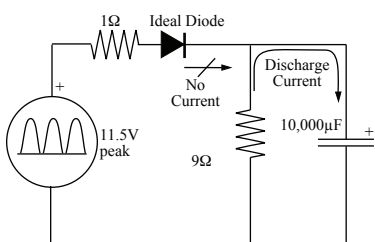


Figure 10-22

The time constant for this process is $9\Omega \cdot 0.01\text{F} = 90\text{mS}$. This time is very long compared with the 8.3mS left until the voltage peaks again so the voltage across the capacitor only has time to sink by a small amount. The capacitor was filled up with charge from a low resistance source during the charging phase and is now supplying current at a much lower rate through the higher load resistance so that the voltage lasts very well (Figure 10-23).

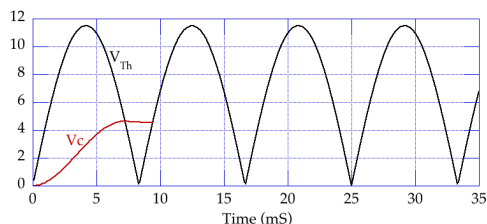


Figure 10-23

Next, the source voltage turns round and rises again. It soon passes the output voltage and starts to charge the capacitor again. Because the voltage has only fallen a small way, the next peak is higher than the first one, and the next higher still. So we get the output of Figure 10-24.

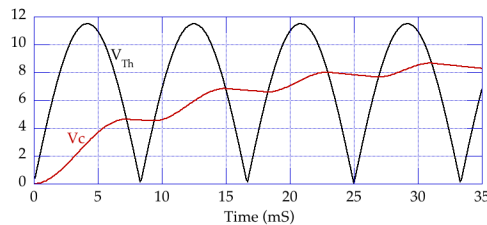


Figure 10-24

Over a period of several cycles the pattern settles down to the output voltage shown in Figure 10-25. The output is still not perfectly smooth. It rises and falls by a small amount called the **ripple voltage**.

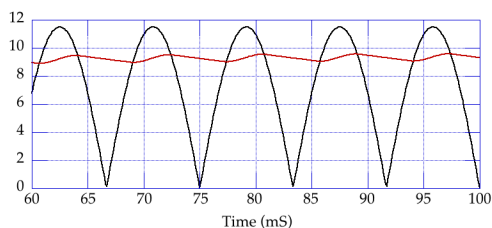


Figure 10-25

By increasing the size of the capacitor we decrease the ripple voltage but we cannot make it go away completely. If we need a really smooth voltage then we need to use a voltage regulator as in chapter 25. We can calculate the ripple voltage fairly accurately by assuming that the discharge current is approximately constant (valid because of the long discharge time). In our example

$$I_{out} = \frac{10.35V}{9\Omega}$$

We know that the current flowing out of a capacitor is equal to the rate at which charge is decreasing on the capacitor plates, so we can calculate the rate at which the voltage is decreasing

$$\frac{dV_c}{dt} = \frac{1}{C} \times \frac{dQ}{dt} = \frac{1}{C} \times I_{out} = \frac{1.15A}{0.01\mu} = 115V/\mu S$$

Now, the time during which the capacitor is discharging is at most 8.3mS so the maximum voltage drop is

$$V_{ripple} = 0.0083s \times 115V/\mu S = 0.95V$$

In general the ripple voltage from a capacitor smoothed power supply can be estimated with the formula

$$V_{ripple} = \frac{I_{out}}{f \times C}$$

where I_{out} is the maximum current drawn from the supply, C is the value of the smoothing capacitor and f the frequency of the unsmoothed supply. For a half-wave rectifier f is the same as the input power-supply frequency but for a full-wave rectifier of either kind f is twice the frequency of the input power supply.

10.5 Simulating the Full-Wave Rectifier

We can use PSpice to examine the voltages and currents inside a model of a working full-wave rectifier.

10.5.1 Unsmoothed

We will start with the unsmoothed full-wave rectifier circuit.

1) PSpice does not provide a model of a transformer so we shall have to make our own using the Thevenin circuit from Figure 10-3. In this case we shall use a 15-0-15V transformer with an internal resistance of 2 Ohms so the model uses two 15V VSIN sources at 60Hz in series with two 1 Ohm resistors. Then we shall use two 1N4002 diodes for the rectifier and simulate the load with a 20 Ohm resistor, R_L , as shown.

2) We are interested in seeing the relationships between the input and output voltages so we put markers on the transformer leads (represented by the ends of R_1 and R_2) and upon the load (R_L).

3) We need to set the voltage sources up to have an amplitude of 15V, an offset of 0V, and a frequency of 60Hz. We need not bother with the DC or AC properties since we are doing either a DC or AC Sweep.

4) We want to look at the output waveforms so we select a Transient analysis and ask for 25mS of simulation. The output is shown in Figure 10-27 overleaf.

Info The commonest power supply for all small electronic instruments is the wall brick or “wall wart”, a transformer and rectifier circuit built into a large plug. These are available in a wide variety of voltages and currents and there are even adjustable versions with a switch to select one of a set of common voltages such as 3V, 4.5V, 6V, 9V, 12V (all voltages that can be achieved with regular batteries!). One of these makes a very good basic power supply for a small instrument, although the ripple is usually appalling. I highly recommend these for small projects with common power needs, especially if you add an external filter capacitor to clean up the ripple.

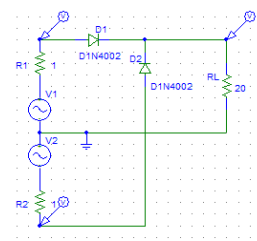


Figure 10-26

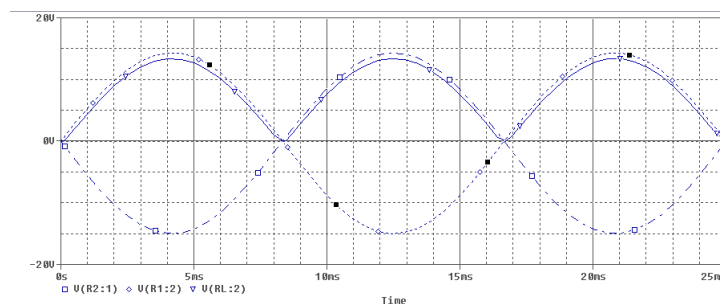


Figure 10-27 Simulated Rectifier Output

The dotted and dot-dashed lines are the two inputs and the solid line is the output. As we expect, the output is always positive and follows 1 diode voltage drop below the more positive input. The separation between input and output curves increases slightly towards the top of the wave as the current flowing in the diode increases.

It is instructive to look at the current flow in this situation. I deleted all the lines from the graph and then replotted it showing only the currents flowing in the two power supplies.

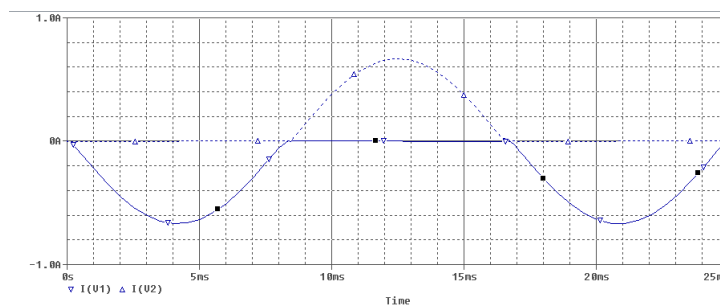


Figure 10-28

We see clearly that current flows in only one side of the transformer at once. Indeed there is a short period, when the voltage is too low to turn on either diode, when neither side conducts. Note that the current is shown as flow INTO to the positive terminal of each VSIN. Thus, during the first half-cycle when V1 conducts and current flows OUT of it, into the rectifier, the solid curve lies below the axis. In the second half-cycle, V2 conducts and current flows out of its negative terminal. That means that current flows INTO the positive terminal and so the dashed line lies above axis. This reflects the fact that, from the point of view of the transformer, current flows round the circuit in opposite directions on each half cycle.

10.5.2 The Effect of a Smoothing Capacitor

Now let us add a capacitor to smooth out the ripple voltage. How large should the capacitor be?

First we must decide on the maximum ripple that we can tolerate. Let us choose the rather large value of 0.5V so that we shall be able to see the ripple when we simulate the circuit.

Taking into account the diode voltage drop, the peak output voltage is about 14.4V so that the peak output current is 720mA.

Given those facts we compute the capacitance needed as

$$C = \frac{I_{out}}{f \times V_{ripple}} = \frac{0.72}{60 \times 0.5} = 24,000 \mu\text{F}$$

That is not a standard value so we shall use the next larger standard value, 25,000 μF .

Here is the simulation output with the capacitor added.

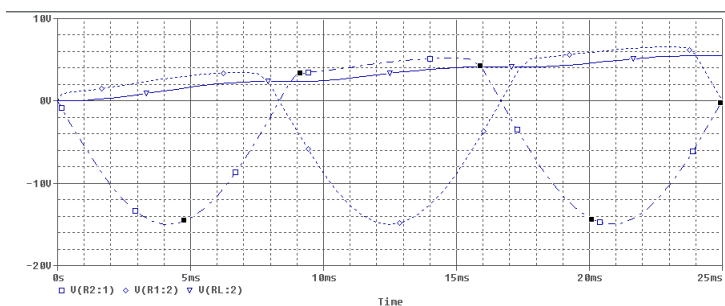


Figure 10-29 Smoothed Full-wave Output

That is somewhat surprising. The output (the solid line) is barely making it to 5V. If we look back at figures 10-23 and 10-24 we can see what is happening. The smoothing capacitor has not had time to charge to its full value yet. If we re-simulate the circuit over a longer timescale then we should see the output settle down to a voltage near 14V. Here is the result of simulating the first 1/3second after turn on.

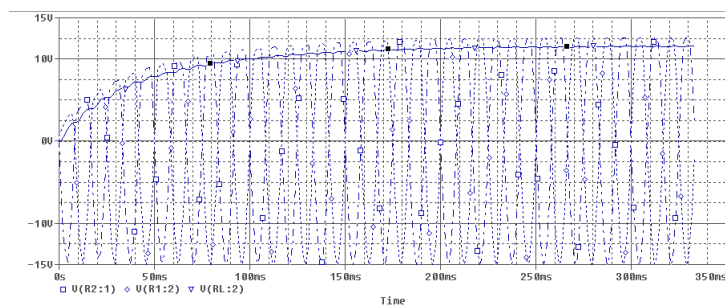


Figure 10-30

This looks better. After about 200mS the output has settled down to a level just below (about 1 diode drop) the peaks of the input. It shows a small ripple but the size is hard to estimate. By altering the graph axes we can get a better look at the ripple.

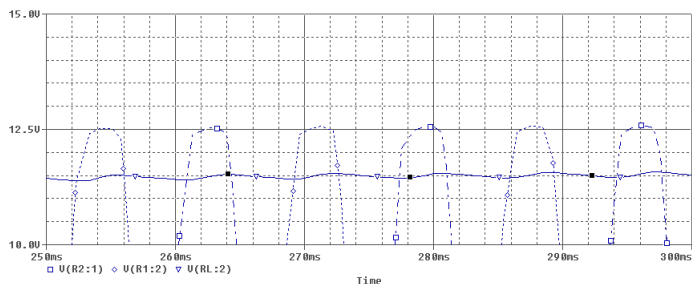


Figure 10-31

Here we see that the ripple is actually quite a bit less than 0.5V (1 small division on this scale) and the actual output voltage is closer to 12V than 14V. This is because of the internal resistance of the transformer. During the times that the diodes conduct very large currents flow to charge the capacitor and these give large voltage drops in the internal resistances, R1 and R2, which bring the peak transformer voltage down to about 12.6V leaving only 12V for the final output. We can see the truth of this by plotting the transformer currents again.

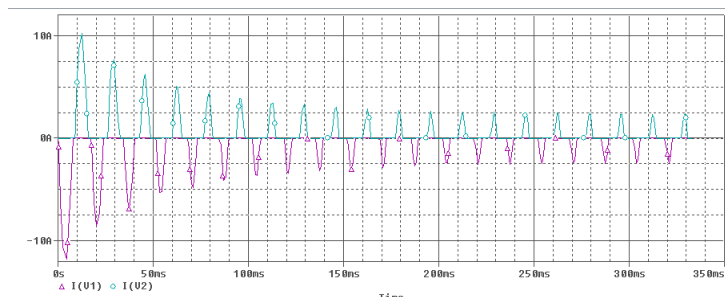


Figure 10-32

As expected, only one side of the transformer conducts at a time. The numbers are quite surprising though. During the first few cycles after turn-on the transformer is delivering about 10A pulses to charge the filter capacitor. Even when things settle down the pulses are still about 2.5A tall, comfortably accounting for the 2.4V drop seen in the peaks of the output.

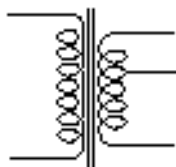
Note: The very high initial current seen in this system is actually very common. Because the excessive current draw only lasts for a few cycles it does not cause overheating problems. However, it does blow fuses. Remember that more current flow in the secondary of the transformer means more current flow in the primary, thus the device pulls an extra large current from the mains supply during this period. If we put in a fuse that is small enough to protect the device during normal operation then it will probably blow every time you turn the machine on. Instruments with large smoothed power supplies thus have to use “slo-blow” or time-delay fuses. These are fuses that will allow excessive current to flow for a fraction of a second and will only blow if the condition continues for more than about 200mS.

Summary

A mains operated power supply consists of three pieces, a transformer to change the incoming high AC voltage to the appropriate voltage for the task at hand, a rectifier to make the current all flow in one direction, and a filter capacitor to smooth the output voltage so that it is nearly constant with only a small ripple voltage.

A transformer uses Faraday’s law of electromagnetic induction to change one alternating voltage into another. When a voltage V_1 is applied to the primary coil, with N_1 turns, then a voltage V_2 appears across the N_2 turns of the secondary coil where

$$V_{out} = \frac{N_2}{N_1} \times V_{in}$$



The symbol for a transformer looks like the coils from which it is made. Transformers are rated in terms of the RMS voltage output from each set of windings and the amount of current that you can safely draw from each winding.

A **half-wave rectifier** supplies current only on every other half-cycle of the supply voltage. A transformer with an RMS rating V_{RMS} used with a single diode and a filter capacitor C will produce a peak output voltage

$$V_{Pk} = (V_{RMS} \times \sqrt{2}) - 0.6V$$

with a ripple voltage

$$V_{ripple} = \frac{I_{out}}{f \times C}$$

The diodes must be rated to handle the full output current and a peak inverse voltage greater than V_{Pk} .

A **full-wave rectifier** supplies current on both half cycles of the supply voltage. A center-tapped transformer rated at V_{RMS} -0- V_{RMS} used with two diodes and a filter capacitor C will produce a peak output voltage

$$V_{Pk} = V_{RMS} \times \sqrt{2} - 0.6V$$

with a ripple voltage

$$V_{ripple} = \frac{I_{out}}{2 \times f \times C}$$

The diodes must be rated to handle the full output current and a peak inverse voltage greater than twice V_{Pk} .

A **bridge-rectifier** is a special form of full-wave rectifier. It uses an untapped transformer and a 4-diode bridge. A transformer with an RMS rating of V_{RMS} used with a 4-diode bridge and a filter capacitor C will produce a peak output voltage of

$$V_{Pk} = V_{RMS} \times \sqrt{2} - 1.2V$$

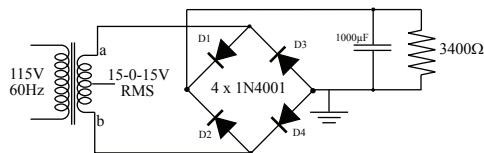
with a ripple voltage

$$V_{ripple} = \frac{I_{out}}{2 \times f \times C}$$

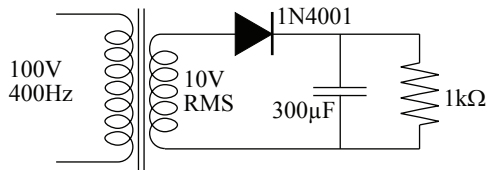
The diodes must be rated to handle the full output current and a peak inverse voltage greater than V_{pk} .

Exercises

1. What is the peak voltage output from a 15V RMS transformer?
2. On average, do you need larger or smaller filter capacitors for a bridge rectifier compared with a half-wave rectifier? Explain your answer.
3. On average, do you need larger or smaller filter capacitors for a bridge rectifier compared with a full-wave rectifier? Explain your answer.
4. Estimate, as accurately as possible, the output voltage and ripple voltage for the circuit below.



5. Calculate the peak voltage and the ripple voltage output from the circuit below.



- 6) Design a simple power supply to deliver 15V DC at 100mA to its load. Choose an appropriate transformer, diode(s), and filter capacitor and draw the complete circuit.

Chapter 11: The Field-Effect Transistor

11.1 Introduction

All of the components that we have met so far are passive ones—they can affect the shape, phase, amplitude, etc. of a signal but they can only reduce the amount of energy in the signal. In order to increase the energy in a signal, to **amplify** it, we need **active** components. The most important active component that we shall meet is the **transistor**, which has dominated all of electronics since the 1960's when the vacuum tube was phased out as the principle active component.

There are two fundamentally different kinds of transistor, which differ in the way they work and in the ways that they interact with the circuitry around them, but which have certain common features. Transistors are 3-terminal devices in which a small signal applied between one pair of terminals controls a much larger current which flows between the other pair. All transistors require a steady DC source of power to operate at all and then need smaller, time-varying signals to do their work. We shall deal here with the class of transistors called **field-effect** transistors.

Info The name transistor originated with the first experiments to produce three-terminal semiconductor devices conducted at Bell Laboratories in the 1950's. These experiments proved that a voltage applied to one terminal could control the resistance between the other two terminals. This property was called **transfer resistance** and the name **transfer resistor** soon became shortened to transistor. Our FETs are distant descendants of these early devices.

11.2 The FET

A field effect transistor has three terminals, called the gate, source, and drain. A voltage applied between the gate and source terminals alters the resistance of the device to current flowing between the drain and source terminals. Almost all the current in the device flows in the channel between the source and drain terminals, little or no current flowing into or out of the gate terminal. The device gets its name because it is the electric field set up by the gate voltage that controls the channel resistance. Figure 11-1 shows the basic symbol for an FET, although there are minor variations to distinguish different types of FET (see below).

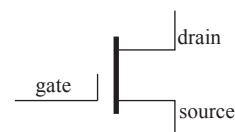


Figure 11-1 Basic FET Symbol

FETs come in several varieties with slightly different characteristics. First, there are **n-channel** and **p-channel** FETs, which differ in the doping of the channel that carries the current between the source and drain. In an n-channel electrons flowing in an n-type semiconductor channel carry the FET current. In a p-channel FET, the current is carried by holes flowing in a p-type channel.

Second, the gate terminal can be either completely isolated from the drain and source terminals or can be made to form a diode with them. An FET made with the first kind of gate is usually called a MOSFET, standing for **Metal Oxide Silicon Field Effect Transistor**, because it is made from layers of metal, silicon dioxide (quartz or glass), and silicon. In this kind of FET the gate draws zero static current since there is an insulating glass layer to prevent current flow. In fact, the gate acts like a capacitor. This separation between gate and channel is shown by the gap in the MOSFET symbol (Figure 11-2).

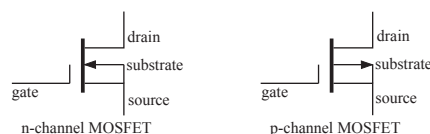


Figure 11-2 MOSFET Symbols

An FET built with the second kind of gate is called a junction FET or JFET because a diode junction is formed at the gate. This kind of FET must be operated with the gate-source junction reverse biased or current will flow into the gate and destroy the device. Even reverse biased,

the junction FET does draw a very small current through the gate because of diode leakage. However, that current is extremely small compared to the drain-source current.

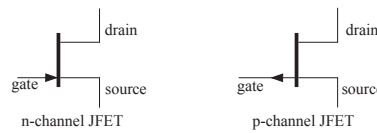


Figure 11-3 JFET Symbols

Finally, a MOSFET can be either an **enhancement-mode** or a **depletion-mode** device. Depletion-mode MOSFETs are quite rare. In them a large current (10's of mA) flows when the source and gate are at the same voltage and you have to apply a voltage to the gate to cause the current to decrease. Most MOSFETs are enhancement mode devices. These do not conduct at all when the gate is at the same voltage as the source and then start to conduct as the gate-source voltage is increased. By contrast, junction FETs are always depletion mode devices, since they would need to be forward biased to run in enhancement mode.

We shall deal almost exclusively with enhancement mode MOSFETs of both polarities.

11.3 The water model MOSFET

Even the MOSFET has a water model analogue. As you might expect, it is quite a bit more complex than anything we have met so far. Figure 11-4 is a picture of what a water model of a depletion-mode MOSFET would look like.

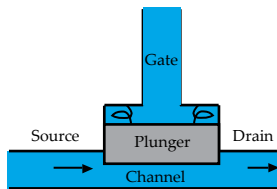


Figure 11-4 Water Model of a MOSFET

Note It is just as easy to construct a water model of an enhancement-mode FET but it is *much* harder to draw a 2-D picture.

When there is no water pressure in the gate chamber, the springs hold the plunger open so that water can flow freely through the device. When water is let into the gate chamber, the pressure pushes the plunger back against the spring and starts to constrict the channel. The greater the water pressure, the larger the area of channel blocked and the higher the resistance of the device to the water flowing through it. There is no connection between the gate chamber and the channel so water cannot flow between the gate and either the source or the drain. However, it does take up some water to fill the gate chamber and provide the pressure to move the plunger so the gate chamber acts like a capacitor—the more pressure you apply the more water you need to put in the chamber. A very tiny, transient, flow of water in the gate chamber can control the much larger steady flow of water through the channel in the plunger.

Remember Except during rapid changes in gate voltage, **no current flows into or out of the gate of a MOSFET.**

11.4 3-terminal device characteristics

Back in Chapter 4 I mentioned that a 3-terminal device is really a special case of a 4-terminal device because the input has to be applied between two terminals and the output taken from two terminals (Figure 11-5). In the case of the FET, the input is applied between the gate and source and the output is taken from the drain and the source. The simple I-V curve of a 2-terminal device is no longer sufficient to describe the behavior. We use three types of information to describe a 3-terminal device

1. the input characteristic
2. the output characteristic
3. the transfer function

The **input characteristic** tells us how the device affects the circuit that it is connected to. This treats the device as a 2-terminal device and shows how the input current is related to the input voltage. In theory this needs to be done for many different sets of operating conditions but in practice it is fairly constant for many kinds of device.

The **output characteristic** tells us how the two output terminals behave and how they affect the circuit to which the device is connected. Again, this treats the device as a 2-terminal device and shows how the output current and voltage are related. For most interesting devices the

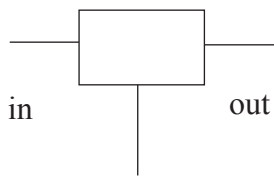


Figure 11-5 3-Terminal Device viewed as a 4-Terminal device.

output characteristic depends on the conditions on the input, often quite strongly. This means that we don't have an I-V plot with only one line on it but a plot with many lines, each giving the behavior for a different set of conditions at the input terminals.

A **transfer function** is more complex. It tells us how some property of the output terminals depends on some property of the input terminals. For example, in an FET it is common to plot the output current (drain-source current) as a function of the input voltage (gate-source voltage). Such a plot often shows several different lines taken under different conditions of, for example, the output voltage. There are four possible transfer functions that we can plot, only one of which is usually specified for a given kind of device.

1) **Voltage gain**—plots the output voltage as a function of the input voltage. The ratio of output voltage to input voltage (the slope of this line) is called **voltage gain** and it is a dimensionless quantity.

2) **Current gain**—plots the output current as a function of input current. The ratio of output current to input current is called **current gain**. This is also a dimensionless quantity.

3) **Transresistance**—plots the output voltage as a function of the input current. The ratio of output voltage to input current is called **transresistance** or **transimpedance** since it has the same units as resistance or impedance but the voltage and current are measured in different parts of the device. The units of transimpedance are Ohms.

4) **Transconductance**—plots the output current as a function of the input voltage. A current divided by a voltage has the units of 1/resistance, which is called **conductance**. A conductance computed from an output current divided by an input voltage is called a **transconductance**. The units of transconductance are Mhos.

$$1 \text{ Mho} = \frac{1}{1 \text{ Ohm}}$$

Just as we usually use the symbols R or r for resistance, we usually use G or g for conductance.

Info Trans means across.

Note Mho is Ohm spelled backwards. The official unit of conductance is the Siemen—1 Siemen = 1 Mho—but everyone still uses the Mho!

11.5 Characteristics of an FET

Let us look at the characteristic curves of a typical FET. We will look at the 2N7000, a typical small-signal enhancement mode n-channel MOSFET.

There is no input characteristic curve for a MOSFET since the input draws no DC current. Instead, the input behaves like a capacitor. The data sheet for the 2N7000 tells us that the capacitance is 60pF max. Thus we know that a 2N7000 will appear like a capacitor of 60pF so far as the circuit driving it is concerned. We shall see later how this will affect the behavior of FET circuits.

11.5.1 FET Output Characteristic.

The output characteristic is shown in Figure 11-6 as a plot of drain current (I_{DS}) vs. drain-source voltage (V_{DS}) for various values of the gate-source voltage (V_{GS}).

Look at a single line on this graph, for example, the $V_{GS} = 8\text{V}$ line. The curve can be divided into two regions with a small transition region between them. For drain-source voltages (V_{DS}) between 0V and 4V the line is moderately straight. Its slope varies from about $1\text{A}/2\text{V} = 0.5 \text{ Mho}$ at 0V to about $10.5\text{A}/4\text{V} = 0.12 \text{ Mho}$ at 4V. These slopes correspond to resistances of 2Ω to 8Ω . So, for moderate values of the drain-source voltage the channel of the FET behaves like a nearly constant resistance.

Remember The resistance of a component is inversely proportional to the slope of its characteristic curve in an I-V plot

Once the drain voltage rises above 4V, the resistance rises sharply to a huge value—the current reaches a maximum value and then stays almost the same regardless of further increases in drain-source voltage. We call this phenomenon **saturation**. It occurs because there is a fixed limit to how much current can flow through the channel which is set by the structure of the channel (see later) and is not affected by the drain-source voltage. In this region the I-V line is flat and so the resistance is extremely high.

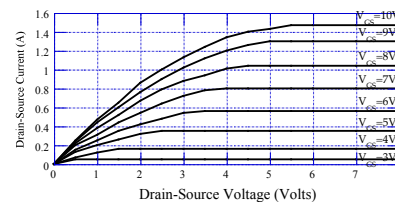


Figure 11-6 Output Characteristic for 2N7000

Note The data in Figure 11-6 cover currents that you cannot reach with a real device. These data come from the Motorola data sheet and may have been measured on bare devices with special cooling arrangements. They were certainly measured with very short pulses of current. Once the chips are encapsulated in plastic they can only tolerate continuous drain currents of 0.2A

Info Most amplifier circuit configurations work with the FETs in the saturation region so that they operate like constant current sources rather than like low resistances. By contrast, switching circuits operate in the linear region so that the FET acts like a low valued resistor.

Now look at the other lines, which correspond to successively lower gate-source voltages. As the gate source voltage gets smaller, the channel gets thinner and the saturation current falls. At the same time the linear region gets shorter and shorter, saturation happens at lower and lower voltages. In fact, saturation generally occurs at approximately one half of V_{GS} . However, the general shape of the curves is the same; a nearly linear low resistance region then a very high resistance saturation region.

Thus, the channel of an FET behaves like a small, fairly constant, resistance for values of $V_{DS} < 0.5 V_{GS}$ but at high V_{DS} the drain current saturates; I_{DS} ceases to be affected by the V_{DS} and the resistance gets very large.

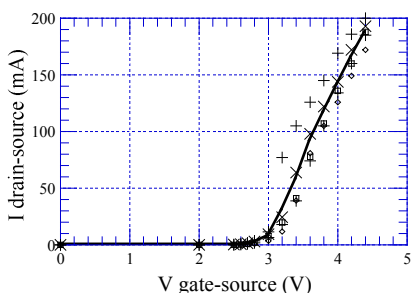


Figure 11-7 2N7000 Transfer Function

11.5.2 FET Transfer Function

The other characteristic plot is the transfer function. For an FET this is a plot of drain current against gate-source voltage. It tells you how the input voltage affects the output current. There are many possible curves depending on the drain-source voltage at which they are measured but the most common practice is to plot the curve under saturation conditions. I measured the transfer curve of some 2N7000s at $V_{DS} = 10V$ (Figure 11-7). The figure shows data from 5 randomly chosen 2N7000s with an average line added.

As you can see, no drain current flows until the gate-source voltage reaches the **turn-on voltage** or **threshold voltage** of about 2.5V. After that the saturation current rises quite steadily as the gate-source voltage rises. The rate of rise is not quite constant—the curve is not quite a straight line—but the non-linearity is fairly small. The slope of the curve is called the **forward transconductance**, symbol g_m .

For the 2N7000 under the conditions in Figure 11-7, g_m varies from about 10 mMHos at turn-on to about 100 mMHos in the linear region from about $V_{GS} = 3V$ to $VGS = 4V$. This tells us that, if the drain voltage is high enough to hold the drain current in saturation and the gate-source voltage is in the linear region, then the drain current will quite faithfully follow changes in the gate-source voltage. This is the basis of amplification as we shall see below and in Chapter 18.

Info Why g_m ?

Well g is for conductance and μ is the Greek form of the letter m. Another term for forward transconductance is the mutual conductance so the g is for conductance and the μ for mutual.

Note The transconductance curve of Figure 11-7 was measured for 5 randomly selected 2N7000s. It is more detailed than, but agrees in general form with, the curve in the data sheet for the 2N7000 FET published by Motorola. As the figure shows, there are significant variations between individual devices of the same type. There are also significant variations in the behavior of a single device at different temperatures. The data sheet gives data for devices under ideal conditions such as constant chip temperature. These data are not such a good reflection of the real packaged devices that I used. For instance, the published characteristic curves include data taken at much higher currents than would be safe to put through a real, packaged device. In fact, the data sheet lists the maximum continuous drain current as only 200mA but shows currents up to 1.5A (in very short pulses) on the data sheet!

11.5.3 P-Type FETs

So far we have looked only at the more common n-type FET. Although there are somewhat more common, many circuits use both polarities of FET so we should take a quick look at the p-type.

A p-type FET conducts when its gate is made **negative** with respect to its source. P-type FETs operate completely upside down. The source is made more positive than the drain and the FET is turned on by making the gate more negative than the source. Thus the transfer function is inverted relative to the n-type.

The most obvious result of this is that circuits using p-type FETs are upside-down compared to their n-type counterparts. For example, Figure 11-8 below shows simple circuits to measure the transfer functions of both n-type and p-type FETs.

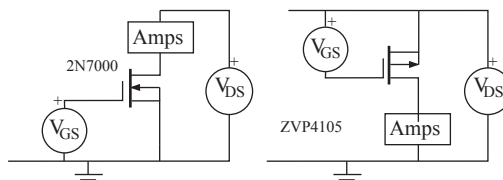


Figure 11-8 Equivalent NFET and PFET circuits

The p-type FET has its source connected to the positive power supply line and current flows into its source and out of its drain. The gate-source voltage is applied between the positive supply and the gate so that the gate is negative with respect to the source.

In general, p-type FETs have somewhat poorer characteristics than their n-type counterparts. At the same magnitude of gate-source voltage, but of opposite signs, a p-type FET will tend to have rather higher resistance than a similar n-type FET.

11.5.4 Power, Heat, and Temperature

When current flows through an FET power is dissipated as always. The power, $I_{DS} \cdot V_{DS}$, appears as heat in the silicon making up the FET. As the FET heats up its characteristics change. In general, as the FET gets warmer its channel resistance gets higher and its threshold voltage gets lower. This means that the two effects somewhat cancel out since a lower threshold voltage means that the FET will be turned on more for the same V_{GS} but a higher channel resistance means that the current flow will be lower for the same V_{DS} . At low currents (less than about 200mA for the 2N7000) the threshold effect dominates and the drain current tends to rise somewhat with temperature. At higher currents the drain resistance dominates and the current falls as the temperature rises. This is a very desirable characteristic. When an FET is overloaded—when it is forced to dissipate too much power—it responds by *decreasing* the amount of current that can flow and so decreasing the power dissipation. It makes FETs fairly hard to destroy by overloading.

Because temperature changes the operating characteristics of the FET it is a little tricky to make good measurements of the characteristics. If you simply run a steady current through the device and measure the voltage (or vice versa) then you get the characteristic at that operating power level. The more current and voltage you apply, the higher power level. So operating curves recorded in this way (like the one in Figure 11-7 above) are not perfectly accurate measures of what the device will do under conditions when the voltage and current are changing rapidly. Then only the average power level is important; short periods of high or low power will not affect the temperature much. You only worry when designing fairly precise circuits.

11.6 Simple FET circuits

Here are some of the simplest common FET circuits. We will look at an FET used as a switch, then as a constant current generator, and finally as a simple amplifier.

11.6.1 Simple FET Switch

Figure 11-9 shows a simple FET switch, which allows us to turn a light bulb on and off with an electronic signal. The light bulb is designed to light up when it draws 100mA from a 15V supply. Thus, it cannot be lit by the 0-5V coming from V_{in} . The FET acts as a voltage controlled switch, allowing the low-voltage, low-current source to control a high-voltage, high-current load.

When V_{in} is 0V the gate-source voltage of the FET is 0V, which is below the threshold so no drain current flows. That means that the drain voltage rises up to 15V and the light bulb is OFF.

When the input voltage goes to 5V the gate source voltage follows it and the FET starts to conduct, since the gate-source voltage is well above threshold. Looking at the output characteristic for $V_{GS}=5V$ shown in Figure 11-10, we see that with a gate-source voltage $V_{GS}=5V$ the saturation current is about 150mA. That is greater than the current that the light bulb draws (0.1A max) so we look along the output characteristic and find that when the drain current is 0.1A then the drain-source voltage (at a gate-source voltage of 5V) is about 0.3V. We call this the **operating point**.

So when the FET is turned on there is about 0.3V across it leaving 14.7V for the light bulb which turns on and draws about 100mA of current through the FET.

If the input voltage drops back to zero then the FET ceases to conduct and the light bulb turns off again. Thus we can control a 15V, 0.1A light bulb, with a 5V signal that supplies NO current to the switch! We shall look at the turn-on and turn-off processes in more detail in the next chapter.

11.6.2 Constant Current Source

Back in chapter 3 we met the idea of a constant current source. This is a two terminal device that keeps a constant current flowing between its two terminals regardless of the voltages on

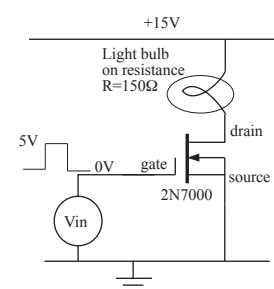


Figure 11-9 Simple FET Switch

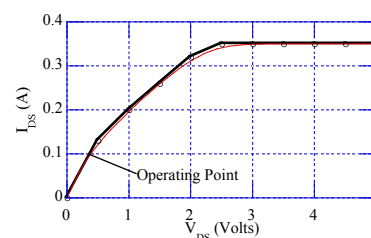


Figure 11-10 Output Characteristic and Operating Point

the terminals. It has an I-V curve that is a straight, horizontal line and thus exhibits an infinite slope resistance. As we have seen in section FET Transfer Function above, an FET in saturation exhibits this behavior. So we can make a constant current source by applying a fixed V_{GS} to an FET and making sure that the V_{DS} never gets small enough for the FET to leave the saturation regions. When we combine the current limiting behavior of a saturated FET with a normal power supply then we get a current source as in Figure 11-11 below.

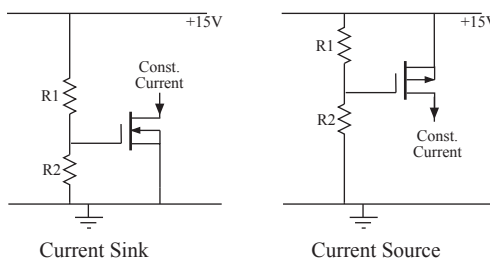


Figure 11-11 FET Current Source/Sink

In each case we keep the source voltage fixed and control the gate voltage with a voltage divider in order to set the operating current. The resulting current flows into or out of the drain. The voltage on the drain can vary over a very wide range without altering the drain current and so we have a constant current source. In Chapter 20 we shall use such circuit to improve an amplifier.

11.6.3 Simple Amplifier

Figure 11-12 shows the circuit of a simple 1-FET amplifier.

Here a 2N7000 is put into a network of resistors that allows it to provide gain in a linear way. That is, it increases the size of a signal without altering its shape. The output voltage is a magnified copy of the input voltage.

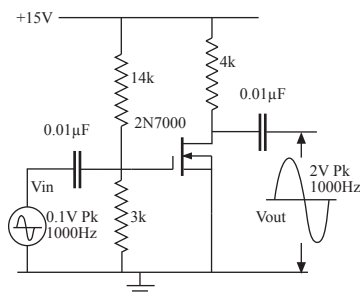


Figure 11-12 FET Amplifier

Note The circuit of Figure 11-12 is an overly simple amplifier circuit. You can build it and get good performance from it if you are lucky but the resistor values may need to be tweaked slightly to get the performance described here. I built the circuit with several different 2N7000s from the same batch and got different gains and different voltages at point b with each one. In Chapter 18 we shall see ways to make amplifiers that are less sensitive to the choice of FET. However they all work according to the principles described here.

To understand the amplifier circuit we first look at the simpler circuit of Figure 11-13.

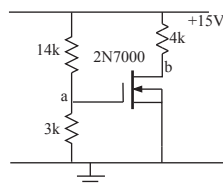


Figure 11-13

The 3k and 14k resistors form a **bias network**. That is, they provide a voltage that holds the FET at a useful point in its operating characteristics. Since the FET gate draws no current, the voltage at point a is found from the voltage divider equation

$$V_a = \frac{3}{3 + 14} \times 15 = 2.65V$$

This is slightly above the threshold so the FET turns on a little and current flows in the 4k drain resistor. The exact amount of current depends on the transfer function of the specific FET. For the FET that I used 2.75mA flowed so that the voltage at point b became

$$V_b = 15V - 4000 \times 0.00275 = 15V - 11V = 4V$$

It is the purpose of the bias network to hold point b at this voltage. When a signal is applied that makes the input voltage go up and down by a small amount, the output voltage can go both up and down. Without the biasing this would not be possible. If we did not have the bias network then the FET would be turned off for all input voltages below about 3V and so the voltage at point b would be 15V and would not vary. For higher input voltages, there would be a little current flow in the FET and the output voltage would fall, creating an output signal.

Now let us return to the full circuit of Figure 11-12. Here a sinusoidal input voltage is added to the fixed bias voltage through a DC blocking capacitor. The capacitor allows the alternating current caused by the signal to flow through it, altering the voltage at the gate of the FET. So

Note Because of the large variations between nominally identical transistors, most of the numbers in this example are specific to the transistor that I used to test the circuit

long as the signal varies rapidly enough, the effect is to add a fraction of the input voltage to the bias voltage so the voltage at the gate oscillates from 2.61V to 2.69V, an amplitude of 0.04V. Now the drain-source current follows this variation in the gate voltage because the transconductance is essentially constant for these small variations of input voltage. At this very low drain current the transconductance is quite low, only about $12.5 \text{ mMho} = 0.0125 \text{ Mho}$ so the drain current varies by $\pm 0.04 \times 0.0125 = \pm 0.0005 \text{ A} = \pm 0.5 \text{ mA}$. That current has to flow through the $4 \text{ k}\Omega$ drain resistor and so the voltage across the drain resistor varies by $4000 \times \pm 0.0005 = \pm 2 \text{ V}$. Thus, a 0.1V sinewave going into the circuit produces a 2V sinewave at the output. The gain of the whole circuit is $2/0.1 = 20$. Note that the $0.01 \mu\text{F}$ capacitor at the output is there to remove the DC component from the output. As the drain voltage varies from 2V to 6V, the capacitor removes the average value so that the voltage at the output of the capacitor varies from -2V to +2V. This is the **AC coupling/DC blocking** capacitor described in chapter 3.

11.7 Thévenin Models of an FET

When we are analysing the workings of an FET circuit it is often useful to have a simple Thévenin model for the device. We can easily construct such a model using the characteristic curves.

11.7.1 DC Thevenin Model

On the input side we know that the FET draws no current from the source and so has an infinite input resistance. On the output side, the FET acts as either a constant current source or a resistance. In either case the value of the current source or resistance depends on the gate source voltage so we have the alternate models of Figure 11-14 below.

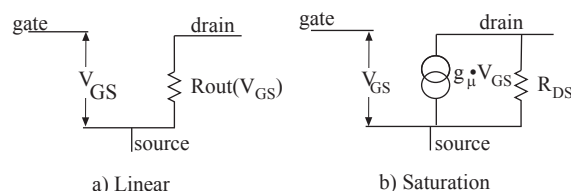


Figure 11-14 DC Thévenin Models of an FET

In each case the input is simply an unconnected wire. The output resistance or current depends on the input voltage, V_{GS} . In the case of the saturated FET we show the variation in its simplest form as a transconductance (g_m) times the the input voltage. This ignores the variation of g_m with V_{GS} and is an approximation that is most useful when the FET is fairly well turned on. The drain resistance, R_{DS} , is included to account for the slight variation in I_{DS} when the drain voltage varies. It usually has a very high value and can be estimated from the slope of the output curves of the FET.

11.7.2 AC Model of an FET

So far we have dealt only with steady currents and voltages. Most interesting circuits operate with time varying currents and voltages. We can improve our model of the FET to take account of these by adding some capacitors to the model. As we shall see in section 11.9 below, the construction of the FET leads to the formation of capacitances between the various terminals. The most obvious ones are the capacitances from the gate to the drain and to the source, but there is also a capacitance from the drain to the source. In addition to the capacitances that are an intrinsic part of the semiconductor device, there are capacitances associated with the leads and the packaging. There is no way to distinguish between these in a real device and so we have a so called **lumped** model where each capacitance shown is really made up from several different physical capacitances.

We sometimes show the capacitances on the normal FET symbol, as in Figure 11-15, but more often show them on the Thévenin model, like this.

Info These capacitances are often called parasitic capacitances. They detract from the performance of the FET but they come along as an integral part of the FET much as the fleas come along with an ill-kept dog.

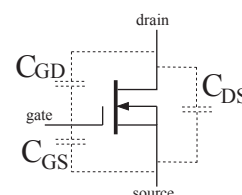


Figure 11-15 FET Symbol showing Parasitic Capacitances

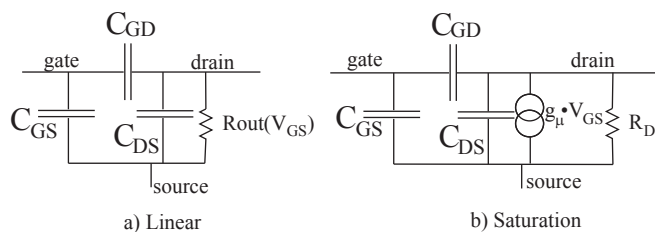


Figure 11-16 AC Thévenin Models of an FET

Note The Motorola data sheet for the 2N7000 uses different names for these capacitances. They call C_{GS} the input capacitance, C_{DS} the output capacitance, and C_{DC} the reverse transfer capacitance.

For a typical small signal FET such as the 2N7000 the data sheet gives values of $C_{GS} = 60\text{pF}$, $C_{DS} = 25\text{pF}$, and $C_{DG} = 5\text{pF}$.

These capacitances are quite small but they are the key factors in determining the high speed behavior of the FET. For instance, if you wish to change the voltage on the gate of the FET then you must charge or discharge the input capacitance, C_{GS} . In order to charge the capacitance, current must flow into or out of the gate lead. This violates our usual rule that no current flows into or out of the gate lead. For example, if we try to apply a 100kHz, 1V sine wave to the gate of the FET then a current of $38\mu\text{A}$ must flow into and out of the gate.

In Chapter 12 we shall see how the input and output capacitances have a profound effect on the speed at which the FET can switch and in Chapter 18 we shall see how these capacitances determine the frequency response of amplifiers.

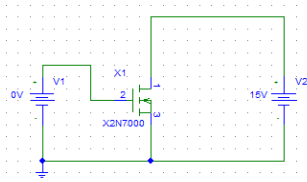
11.8 Using PSpice to Study a MOSFET

Spice comes with a set of very elaborate models for MOS FETs. These have been highly refined to meet the needs of integrated circuit designers. It is enormously expensive to construct a new IC, \$10k-\$100k in addition to the engineering costs to design the thing. Researchers have thus expended considerable resources to refine the models of the components used in IC design, especially the MOSFETs. The resulting models are extraordinarily complex, needing dozens of parameters to describe a single device. This means that you can't just take a generic MOSFET model and tweak it slightly to get a particular device. Fortunately, many manufacturers have done the work to build good models of their own devices and then made those models available (usually on the web). We have such a model of the Zetex 2N7000 n-channel MOSFET. I got it from the ZETEX website and used it to build a PSpice component so you can select a 2N7000 from the parts list without having to worry about how it got there.

11.8.1 The FET transfer function

The most useful FET characteristic is the forward-transfer function; a plot of the drain-source current as a function of the gate-source voltage. We expect that no current will flow until a threshold, in the 2-3V range, is reached and then the current will rise rapidly. We can use the DC Sweep analysis to generate such a function.

1) Build the circuit shown on the right. You will need to set the DC voltage of the drain-source power supply to a suitable voltage; I chose 15V. You can leave the gate-source voltage at zero for now.



2) Go to the Analysis Setup dialog and select a DC Sweep. Choose V1 as the source and set the voltage to go from 0V to 5V in 0.02V. Dismiss the dialogs.

4) Put a current marker on the -ve terminal of the battery. We actually want to see the current that flows out of the FET but Spice won't let us put a marker on the FET model. However, all the current that flows into the drain also flows into the battery so we can put the marker there. (If you put the marker on the +ve terminal the graph will be upside down!)

4) Run the simulation. This should give you a graph like this.

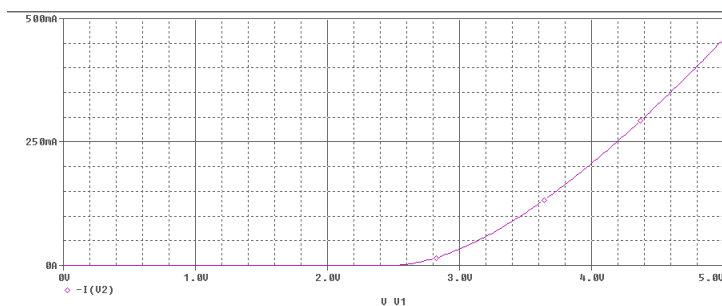


Figure 11-17 Simulated FET Transfer Function

This is exactly what we expected—the current starts to rise at about 2.5V and the rise gets steeper as the gate-source voltage gets higher. The advantage of simulating this curve rather than measuring it is that the simulated transistor does not heat up. When we tried to measure this curve in lab the transistor got hotter as the drain current rose and so different parts of the curve were measured at different temperatures. This is VERY different from the normal operation of the FET as an amplifier, when the variations of current are so fast that the FET temperature cannot respond and is simply set by the average power dissipation. So the simulation gives a BETTER view of the normal behavior of the FET than our simple lab measurement could.

11.8.2 The Output Characteristic

The output characteristic is even more sensitive to the temperature of the FET so simulation is an even better idea for this. We can make this measurement with the same circuit. We simply fix the gate-source voltage and set the drain-source as the swept voltage. Set your V1 source to a value near threshold and set up a DC Sweep of V2 from 0 to 15V. You should get a result something like Figure 11-19.

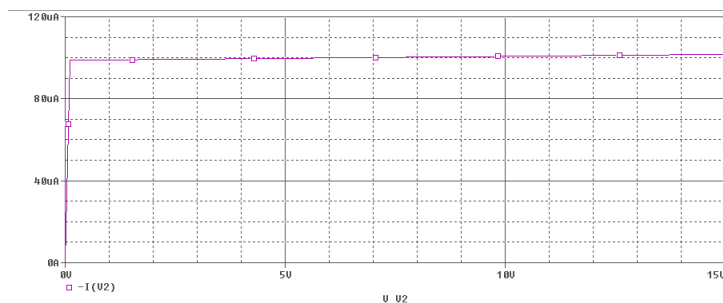


Figure 11-18 Simulated FET Output Curve

You can clearly see the small linear region followed by the nearly flat saturation region.

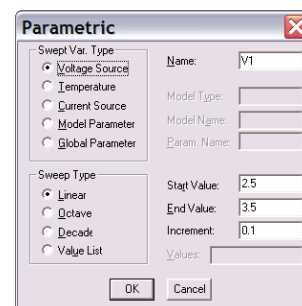
11.8.3 Plotting Multiple Output Curves at Once

The output characteristic of MOSFET depends strongly on the gate-source voltage. What would be really nice would be a way to plot the output curve for several different values of the V_{GS} . Fortunately, an analysis called a Parametric Sweep will do exactly that; I make several runs of another analysis, changing a single value each time.

- 1) Start with the same circuit.
- 2) From the Analysis Setup dialog first make sure that you have set up a DC Sweep of V2 from 0 to about 15V. I used a 0.1V increment. Dismiss the DC Sweep dialog but keep the Analysis Setup.
- 3) Double-click the Parametric... button to bring up the Parametric dialog. We are going to use this to vary the gate-source voltage, V1, so set the name field to V1 and the Swept Var. Type to Voltage Source.

The threshold voltage for this FET is around 2.5 volts so I chose to generate 10 curves at 0.1V increments from 2.5V to 3.5V.

- 4) Dismiss the Parametric and Analysis Setup dialogs.



5) Run the analysis and bring Probe to the front if necessary. It will not be showing a graph but will show a list of datasets or *Sections*.

We want to plot all of these on the same graph so make sure that they are all highlighted (press the All button) and press OK to dismiss the dialog. Probe will put all of the traces on a single graph like Figure 11-20.

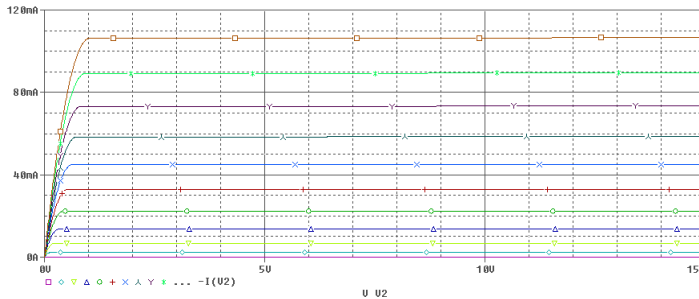
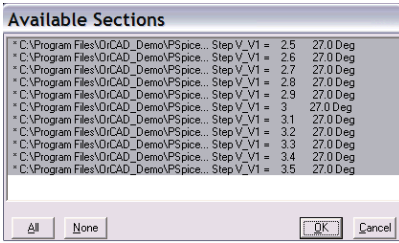


Figure 11-19 Simulated FET Output Characteristics

11.9 The Physics of FETs

Like diodes, transistors are semiconductor devices. Indeed, the first transistors were made by opening up the cases of commercial diodes and adding a third wire. Field-Effect Transistors are rather simple devices to understand, though each different type is constructed a little differently. We shall start with our favorite, the n-channel Metal Oxide Silicon FET.

11.9.1 The N-Channel Enhancement MOSFET

Figure 11-21 shows a side view through an n-channel enhancement MOSFET. The n-type wells under the source and drain electrodes are heavily doped so that they make good low-resistance connections with the metal electrodes. The metal gate is separated from the semiconductor by a very thin layer of silicon dioxide. Silicon dioxide is quartz, an excellent insulator, like a very pure window glass. It is the metal gate and the oxide layer that give the Metal Oxide Semiconductor manufacturing process its name.

Info Substrate

The substrate is like a baseplate—it provides mechanical support for the active components but does not play an electrical role in the device. Most MOSFETs connect the substrate to the source internally but we shall see that some integrated circuit MOSFETs have their substrates connected to other points in the circuit. From our point of view, the substrate is just a way to tell the source from the drain.

In a MOSFET symbol, the gate is the line that does not connect to the rest of the symbol. The thick vertical bar represents the channel and the arrow represents the substrate. The source is the electrode connected to the substrate and the drain is the one that is left!

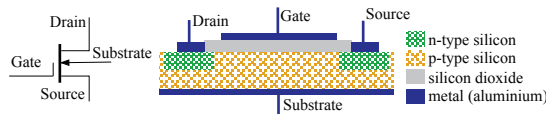


Figure 11-20 Construction of an n-channel MOSFET

In operation, we know that the drain is held positive with respect to the source, which is (usually) connected internally to the substrate. This is important because the n-type wells at the drain and source form p-n junction diodes with the p-type substrate on which the FET is constructed. These diodes must be turned off for the FET to work. The drain-substrate junction is reverse-biased by the drain voltage and the source-substrate junction is zero-biased. Thus, neither diode conducts.

Because both of the p-n junctions are reverse biased, no current can flow from the drain to the gate, no matter how much voltage is applied to the drain, so long as there is no voltage on the gate. However, if we put a positive voltage on the gate then we can alter the characteristics of the p-type material just under the gate. When we make the gate positive, a positive charge builds up on the metal of the gate and it attracts electrons in the p-type substrate towards the gate oxide and repels the holes down toward the bottom of the substrate. We can actually pull enough electrons into a thin layer near the top of the substrate to make the material weakly n-type in this region (called an n- region), leaving a layer of the p-type substrate with a larger density of holes (called a p+ region).

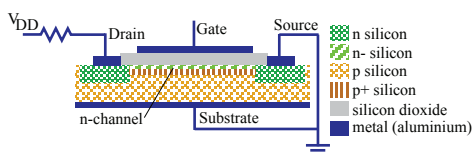


Figure 11-21 t

The channel can carry current from the n-type drain region to the n-type source region. It acts like a high value resistor because of the low density of conduction electrons and the narrowness of the channel. As the gate voltage is increased, more electrons are drawn up into the channel region, which gets thicker and contains more conduction electrons, so the resistance falls.

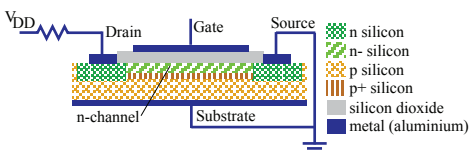


Figure 11-22

If we hold the gate voltage fixed and vary the drain-source voltage then at first the current rises. However, a given gate voltage only creates so many mobile electrons. As the channel current rises, we eventually run out of conduction electrons and the channel current stays constant even though the drain source voltage continues to rise. The FET has entered saturation.

We can see the origin of the gate capacitance in the structure of the MOSFET. The gate electrode is separated from the conductive channel by the insulating gate oxide forming a simple parallel plate capacitor. In order to generate the electric field that pulls the electrons into the n-type channel, you must put a charge onto the gate electrode and charge up the capacitor. The gate oxide makes sure that no DC current flows in the gate but there has to be a transient current flow both to the charge the gate and to discharge it again. This charging current often limits the speed of FETs especially in high power switches.

11.9.2 The N-Channel Depletion MOSFET

The key difference between a depletion mode MOSFET and its enhancement mode cousin is that the depletion mode FET has an n-type channel diffused into the top of the substrate, just under the gate oxide. This means that there is a channel even when there is no gate voltage and so the FET conducts at zero gate voltage.

If the gate is made positive then the channel will grow and the FET on-resistance will fall and its saturation current rise. If the gate is made negative then holes will be attracted to the gate by the negative charge. The holes will consume some of the channel electrons, narrowing the channel and increasing the on-resistance.

11.9.3 The P-Channel MOSFET

A p-channel MOSFET is built and operates in exactly the same way as an n-channel FET except that the polarities are reversed. This time the substrate is n-type material and the drain and source wells are p-type material as is the channel that forms.

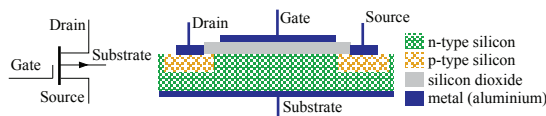


Figure 11-23

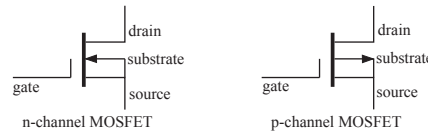
In a p-channel MOSFET, the substrate must be connected to the most positive point in the circuit and is often connected to the source. The drain and gate operate at negative potentials relative to the source and substrate and so the p-n junction diodes are held reverse biased. The negative charge on the gate now forms a strong electric field at the surface of the substrate that

Info This raises the question of why the voltage on the gate is strong enough to create conduction electrons in the p-type semiconductor while the similar voltage between the drain and source is not. The answer lies in the different distances involved. The voltage drop between the drain and source is spread over the whole distance between them, typically $0.25\mu\text{m}-1\mu\text{m}$. The voltage from the gate only acts across the thickness of the gate oxide, about 1nm , so that the electric field which it creates is about 1000 times as great, though it only penetrates a very short distance into the substrate. Thus the gate field is strong enough to pull nearby electrons up into the conducting state and to make the top few layers of silicon atoms into n-type silicon while the drain-source field is only strong enough to move the conduction electrons around once they are formed. The other interesting question is why the conduction electrons end up moving sideways under the influence of the weak drain-source field rather than dashing vertically across the channel in response to the strong gate-substrate field. This time the answer lies in the insulating layers above and below the channel. Above the channel is the gate oxide through which no electrons can pass and below the channel is the p-type substrate. The n-channel and the p-substrate form a reverse biased diode and so no electrons can pass that way.

pushes electrons away so hard that extra holes are formed in the top layers of semiconductor and a p-type channel forms.

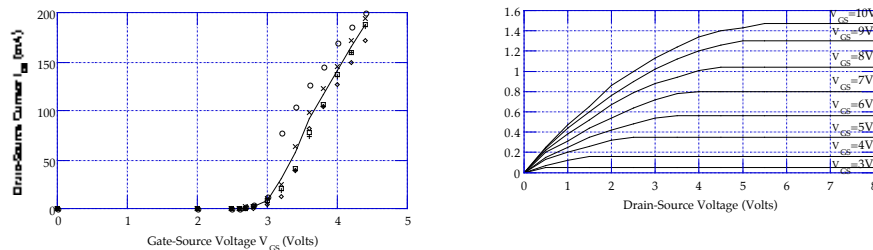
Summary

An FET is 3-terminal device that allows a voltage across the **gate-source** input to control the current flowing between the **drain** and **source** terminals. The symbol shows the channel with the gate isolated from it and also shows, with an arrow, the **substrate** on which the device is built. This is not a lead that emerges from the device but is internally connected to the source.



Beware: Do not confuse the substrate with the gate. The gate is a real input but the substrate is an internal connection and is mostly included on the symbol to make the source different from the drain.

The drain-source channel acts like a resistor whose value depends on the gate-source voltage. The gate-source input looks like a capacitor and draws no DC current. For gate-source voltages below the **threshold** the FET does not conduct. For voltages above threshold the output current is proportional to the input voltage and the proportionality constant is called the **transconductance**. Here are plots of the output current vs. gate-source voltage (transconductance curve, left) and the output current vs. drain-source voltage (output curve, right)



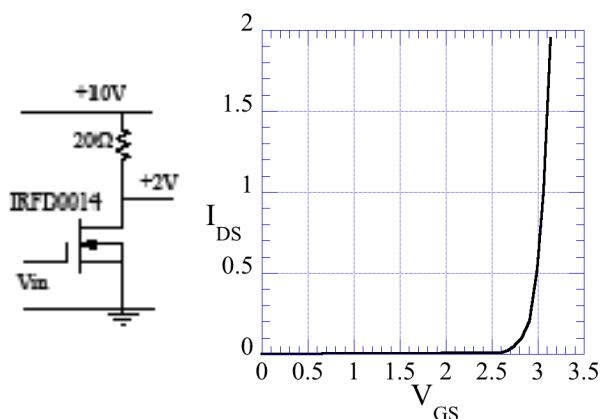
An FET obeys a few simple rules:-

1. No DC current flows into or out of the gate lead.
2. When the gate-source voltage is below the **threshold**, the FET is turned OFF and no current flows between drain and source.
3. When the gate-source voltage is above **threshold**, the drain-source behaves like a resistor whose value is controlled by the gate-source voltage. The higher the gate-source voltage the lower the drain-source resistance.
4. When the FET is turned on, the drain-source behaves like a resistance until the drain source voltage is so high that the current **saturates**. Further increases in drain-source voltage cause no further increase in drain-source current.

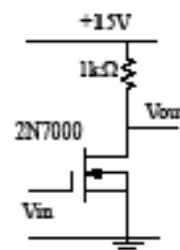
An FET can be used as a simple switch to allow a low voltage, low current signal to control a high voltage, high current load. An FET can also be used with a bias network to build a simple linear amplifier.

Exercises.

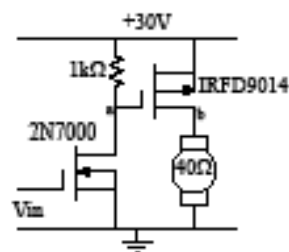
1. Use the I/V curve below right to find the value of V_{in} in the circuit below left.



2. Use your understanding of the FET, and the trans-conductance curve from Q1, to plot V_{out} as a function of V_{in} as V_{in} is varied very slowly from 0V to 5V in the circuit on the right.



3. In the circuit on the left, the funny rounded object labeled 40Ω is an electric motor with a resistance of 40Ω . The PFET, IRFD9014, is a moderate power device that can handle currents of more than 1A and which has an on resistance of 0.5Ω and a threshold voltage of about -3V.



- a) Initially, V_{in} is 0V. What are the voltages at point a and at point b?
- b) Will the motor turn?
- c) What is the current flowing through the S2N7000 and in the IRFD9014?
- d) What is the power dissipated in the S2N7000 and in the IRFD9014?

Later the input voltage is raised to +5V.

- e) What then are the voltages at a and b?
- f) Will the motor turn?
- g) What is the current flowing in each FET?
- h) What is the power dissipated in each FET?

Chapter 12:FET switches.

12.1 Introduction

A switch is a device that can take on one of two states, turned on and turned off. An FET makes a very good switch. When the gate-source voltage is below threshold, no current flows between drain and source. When the gate source voltage is well above threshold, the device enters saturation and allows current to flow from drain to source with a minimal voltage drop across the device. The gate circuit draws no current (except during switching, see below) and so uses practically no power and yet it controls a higher power current flow through the FET.

There are two basic kinds of FET switches, power switches and logic switches, which differ in their output and in their use. Power switches are current switches. They control the flow of current in some external device, turning it on and off. Logic switches are voltage switches. They take one or more on/off inputs and produce an on/off voltage output. The one thing they have in common is that the transistors in them are only ever used in two states, off and on. The FETs in them are always either turned off or are turned on so hard that they enter the saturation region. For this reason all these circuits are called **saturated switches**.

Power switches are often found connecting computers to the outside world. They are devices that use a low voltage, low current input to control a higher power device. For example, in chapter 11 we saw how a 0-5V signal supplying no current could turn on and off a light bulb that took 15V at 0.1A. The commonest place to find a power switch is between a computer and a real world device such as a light, a furnace, or a motor. FET switches are available over a wide range of powers from simple low power ones, like the 2N7000 circuit that we have already seen, to power FETs that can switch hundreds of volts at currents of several amps. In this chapter, we shall look only briefly at power switches but we shall return to them in Table 12-6.

Logic switches lie at the heart of all computing devices. They are circuits that use two different voltages to represent the two different states of a logical value—false and true. Logic switches are circuits whose inputs and outputs are both logical values and in which there is some special relationship between the input and output. The simplest example is a switch called a NOT gate, whose output is true if its input is false and vice versa. These have to operate rapidly and to use very little power so that they don't heat up too much. For example, a modern microprocessor may have more than a million such switches so that each one must use only a few microwatts of power or the whole microprocessor will get very hot. This means that each switch must operate at very low currents.

Note Until recently, almost all logic devices operated from 5V power supplies and so used 0V and 5V as the two logic levels. Recently there has been a move to lower and lower operating voltages. First the power supply dropped to 3.3V and, more recently, 2.5V, 1.8V and even 1.3V devices have appeared. These lower voltage chips use less power than their high voltage relatives and so run cooler and less wastefully

12.2 Power Switches

Despite their greater size and power handling, power switches are typically easier to understand than logic switches, so we shall look at them first. At its simplest, an FET power switch consists of a single FET connected in series with a load and driven directly from the input signal (Figure 12-1). We looked at the operation of such a switch in Chapter 11 and now return to it briefly to examine some of the limits to its performance.

The first limiting factor is the FET that we have chosen. According to the data sheet, the 2N7000 can withstand a drain-source voltage of at most 60V, can carry a drain current up to 200mA continuously, and can dissipate up to 350mW (and that would make the FET run extremely hot!). This clearly limits the size of load that it can drive. If we want to switch larger loads then we must choose a more powerful FET.

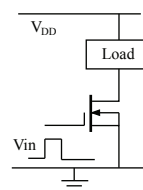


Figure 12-1 Simple FET Power Switch

Table 12-1 compares the 2N7000 to a few examples from the vast range of power FETs that is available.

Note The more power the FET has to handle, the hotter it will get. Higher power FETs come in large, rugged, metal cases and must be cooled to reach their full power rating. The case is typically bolted to a large piece of metal, called a **heat sink**, which carries the heat away. Moderate power devices use large pieces of aluminum with fins on them to help remove the heat. Higher power devices add a fan to blow over the heat sink and some even use water cooling.

Table 12-1: Some Power FETs

Device	$V_{MAX}(V)$	$I_D(A)$	$P_D(W)$	$R_{ON}(\Omega)$
2N7000	60	0.2	0.5	4
IRFZ14A	60	10	30	0.14
IRFZ44A	60	50	126	0.024
IRF510A	100	5.6	33	0.4
IRF550A	100	40	167	0.04
IRF610A	200	3.3	38	1.5
IRF650A	200	28	156	0.085

The second limitation is the drive—the voltage available to turn the FET on. In the most common case we have a logic level (0-5V) signal and want to control a power hungry load. If the load is small enough to use a 2N7000 then all is well. 5V of gate drive are sufficient to push the FET into saturation and so the only limit to the current is the power rating of the FET. However higher power FETs require a lot more drive. For example, the threshold voltage of the common IRF510 is 4V and it requires a V_{GS} of 10V to saturate the device and switch the full rated current. That is too much for our logic level source. In that case we must either use a special low threshold power FET or use a more complicated circuit.

12.3 Logic Switches

A logic circuit is one in which each signal can only be one two distinct states as opposed to an analogue circuit in which a signal takes on a continuous range of values. The two states are usually called **true & false** or **high & low**. They are usually represented by two different voltages. The most common logic circuits use 0V for low and 5V for high. A system in which all signals have only two levels is called **binary** system. We are going to look at binary logic systems for the next few chapters. A single signal in a binary system can represent one fact such as

one bit of a binary number
 whether a lamp is on or off
 whether a switch is on or off
 etc.

A logic circuit is made up of electronic switches. Each switch has one or more **input** signals and (usually) a single output whose state depends on the values of the inputs. Each input and output can be in only one of the two states at any time. We call such a signal-controlled switch a **gate** and we give different kinds of gates different names depending on how the inputs affect the output.

12.3.1 Positive and negative logic

Although the terms high and low map uniquely to the voltages 0V and 5V, there is no unique map from high/low to true/false. It is probably more common to use the map

high = true
 low = false

which is called **positive logic**, but it is equally possible to use the map

high = false
 low = true

which is called **negative logic**.

In order to make it easier to tell which way a given signal works we use a notation called **assertion logic**. In this scheme a signal with a plain name is true when it is high and false when low, a signal with a bar over it is true when it is low and high when it is false.

Example

Consider a signal called Closed that tells us whether a particular switch is open or closed.

If the switch is connected as shown in Figure 12-2a, then the signal Closed will be high when the switch is closed and low when the switch is open. We call a signal like this a **positive logic** (or +ve logic) signal since the signal is true when it is high.

By contrast, if the switch and resistor are reversed we get the circuit of Figure 12-2b. Now, the output is low when the switch is closed and high when it is open. This means that Closed is **true** (that is, it is true that the switch is closed) when the signal is **low** and so the signal is a **negative logic signal**. We show this by putting a bar over the name of the signal.

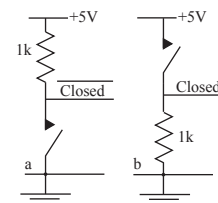


Figure 12-2 +ve (a) and -ve (b) True Switches

12.3.2 A simple logic switch

The simplest logic switch has a single input and a single output. It can work in one of two ways.

The first way is to have the output equal to the input; if the input is true then the output is true, if the input is false then the output is false. The second way is to have a true input make a false output and vice versa.

There are two ways to build the first type of switch. The simplest way is just a piece of wire. The more complex way is called a **buffer** and has the symbol shown in Figure 12-4. A buffer differs from a wire in two ways. First, it is a one way device. The output follows the input but nothing that happens at the output can affect the input. By contrast, a piece of wire is a two-way device. A voltage at either end affects the voltage at the other. Second, a buffer can increase the power of a signal. The input of a buffer draws very little current from the input signal but the output can supply much more current to the circuit following to it. A wire can only carry current from one end to the other.

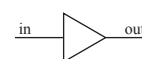


Figure 12-3 Buffer Symbol

The second type of 1-1 switch is called a **NOT gate**, or an **inverter**. When the input is true, the output is **not** true etc. The symbol for an inverter is a small circle drawn in the signal line. However, the usual symbol for a complete not gate is shown in Figure 12-5, where the inverter is added to a buffer symbol. This shows that the gate not only inverts the signal passing through it but also buffers its output from its input, protecting the input from changes at the output and allowing a low current input to drive a high current output.



Figure 12-4 Logic Inverter

12.4 FET Logic Switches

The simplest logic gate that we can build is a NOT gate, which can be made with only two components. We shall look first at a couple of simple inverter circuits and then combine them to get the CMOS inverter that lies at the heart of all modern digital circuitry.

12.4.1 The NMOS inverter

Our first attempt at an inverter uses an n-channel MOSFET and a resistor (Figure 12-6). It is called an NMOS inverter because it uses only an n-channel MOSFET. The first thing to notice about it, in contrast to the simple inverter symbol above, is that it has not only an input wire and an output wire but also has wires for a 5V power supply and ground line. These are *essential*. No real circuit can operate without power and so wires must be present to supply that power. In addition, a real signal voltage only has meaning with reference to some other fixed point in the circuit, a ground point. All signal voltages are defined between the signal wire and the ground. The power and ground wires are often not shown in logic circuit diagrams because they clutter the picture without supplying any real information. However, they must always be present and connected whether they are shown or not!

The NMOS inverter is obviously a variant of the simple FET switch that we have already discussed.

- When the input is 0V, the FET is turned off. Since no current can flow in the FET no current flows in the resistor and the drain voltage rises to 5V.

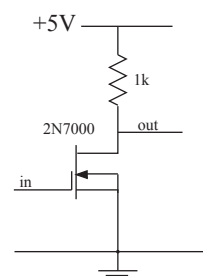


Figure 12-5 NMOS Inverter

Warning The most common place to forget the extra wires is when connecting signals from one board to another. In this case it is vital to connect at least a ground wire as well as the signal wires, otherwise the two boards don't know what each considers to be ground and the signals have no meaning. Thus, it takes at least n+1 wires to carry n signals from place to another.

Note The actual voltage can be found from the voltage divider equation for a 1k resistor in series with resistor representing the turned on FET. In this case the on resistance is of order 5Ω and so we have

$$V_{out} = \frac{5\Omega}{1000\Omega + 5\Omega} \times 5V = 0.025V$$

- When the input goes to the high state, 5V, the FET turns on since its threshold voltage is only about 2.5V. That means that the FET allows as much current as is available to flow. Thus, the drain voltage falls to nearly 0V (see box, right).

This circuit suffers from two kinds of problem. First, it dissipates a fair amount of power when it is turned on. A current of at least $5V/1000\Omega = 5mA$ is drawn from the power supply regardless of how much current is drawn by the output. Second, it is somewhat slow to turn on and off as we can see in Figure 12-7, which shows the output and input when the circuit is driven by a 1 MHz square wave.

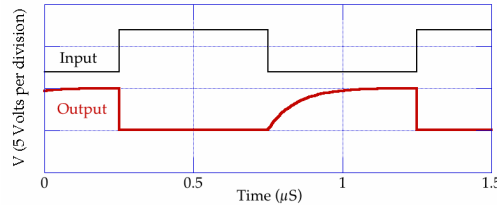


Figure 12-7 NMOS Inverter Timing

Figure 12-6 NMOS Inverter Timing

The first thing to notice is that this is shown as you would see it on an oscilloscope screen. The input trace is not really at 7-12V but was simply moved up there with the offset knob, to get it out of the way of the output. The input really goes from 0V to 5V. Next, we see that the circuit does the right thing. When the input is high, the output is low; when the input is low, the output is high. The trouble is the time it takes to get there. The 1-0 transition is very sharp but the 0-1 transition takes a very long time. The curve is exponential with a time constant of 90nS.

We could improve the turn-off time of this circuit by reducing the value of the drain resistor but that would increase the power drain of the circuit and we do not want that. What we really need is a drain resistor that is very low when the input is low, so the load capacitor can charge quickly, but goes very high when the input is high and the FET is turned on. That way the power drain would be kept low. In other words, we need another switch that works the opposite way.

12.4.2 The PMOS Inverter

We can make such a switch using a p-channel MOSFET, which operates with exactly opposite voltages from an n-channel MOSFET. Figure 12-8 shows a typical circuit. As you can see, the drain is now **negative** with respect to the source. The FET turns on when we make the gate sufficiently negative with respect to the source.

If we drive this circuit with our 0-5V signal we find that when the input is 0V, the gate-source voltage is -5V (since the source is at +5V) and the FET is turned **on**. When the FET is turned on, it pulls the output voltage up to nearly 5V (Figure 12-8).

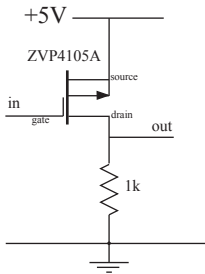


Figure 12-7 PMOS Inverter

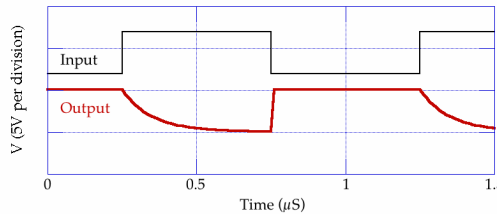


Figure 12-10 PMOS Inverter Timing

Figure 12-8 PMOS Inverter Timing

When the input goes to +5V, the gate-source voltage falls to 0V, the FET turns **off** and the output voltage is pulled to ground by the resistor. Again, the circuit acts as an inverter and again, if we look at the time dependent behavior, it is quite slow (Figure 12-10).

The behavior is quite comparable to that of the NMOS circuit. This time it is the high-low transition that is slow. When the FET turns on the output capacitance can charge through the low resistance of the turned on FET. When the PFET turns off, the capacitance has to discharge through the 1k resistor. The PMOS inverter is a little slower than its NMOS cousin. The fall time is about 100nS, 10% slower than the NMOS inverter. This is presumably due to the slightly larger output capacitance of the FET itself.

12.4.3 The CMOS Inverter

On its own the PMOS inverter is no better than the NMOS inverter, indeed, it may be a little worse. However, something magic happens when we combine an NMOS inverter with a PMOS inverter to form what is called a CMOS inverter (Figure 12-11).

When the input is at 0V, the NFET is turned off but the PFET is turned on. That means that the output is connected to +5V through a low resistance, $V_{out} = 5V$. No current flows through the FET totem pole because the NFET is turned off.

When the input goes to +5V, the NFET turns on and the PFET turns off. Then the output is connected to ground through a low resistance and $V_{out} = 0V$. Still, no current flows through the totem pole because now the PFET is turned off. The output load capacitance is always charged or discharged through the low resistance of a turned-on FET so the switch changes states rapidly. This gives us a nearly ideal inverter.

1. It changes state very rapidly.
2. The output voltage swings very close to both ground and the supply voltage.
3. The output can source or sink moderate currents.
4. The input draws no current when it is in either a high or a low state.
5. The circuit draws no operating current when it is in either the high or low state.

Because the switch is built from two complementary FETs, one p-type and one n-type, it is called a **Complementary Metal-Oxide Silicon** switch or a **CMOS** switch. Almost all logic circuits built today use CMOS switches.

12.4.4 Switching the CMOS Inverter

The CMOS switch is not perfect. It draws no static current but it does draw current when it changes state. This happens in two ways. First, there is the current needed to charge and discharge the input capacitance of the circuit that the gate drives. That current is drawn from the +5V or ground lines through the FET that is turned on at the time. Second, there is a switching current in the totem-pole itself. The input voltage does not pass instantaneously from 0V to 5V; it passes through all voltages between. There is a small amount of time when one FET has not finished turning off before the other starts to turn on. During that time a small current pulse passes down the totem-pole. If we plot the current drawn from the power supply while the switch changes state as a function of time, we get something like Figure 12-12.

This transient current draw means that CMOS circuits draw very little power so long as the operating frequency is low. If the operating frequency gets high, however, the average current gets much larger.

This transient current draw of CMOS logic circuits leads to another important effect. Every time the gate switches and pulls current from the power supply, the power supply tends to dip slightly as the supply readjusts to the changing load. Thus the power lines of a CMOS circuit can get very noisy. Indeed, left unchecked the noise can get large enough to cause gates to switch when they should not because a large noise spike arrives at the power supply leads. We combat this effect by putting capacitors on the power supply lines of logic chips as close

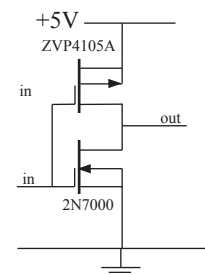


Figure 12-9 CMOS Inverter

Info This output arrangement, with one FET stacked on-top-of another, is called a **totem-pole output**. The idea is that the two FETs connected one on top of another look like the heads in the totem poles carved by the Indians of the American Northwest.

Warning This last statement is only true if the output is connected to another logic circuit and so carrying no current. If the output is connected to a load resistance then the output current will flow in or out through one of the two FETs. In this case the circuit will use some current. So long as the output is only connected to other logic circuits, though, it really takes 0 current to keep it in either the on or the off state.

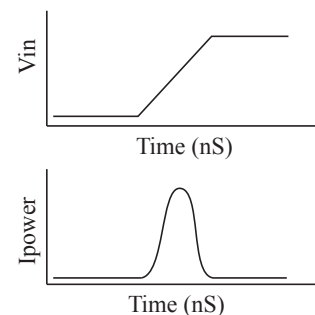


Figure 12-10 CMOS Switching Transient

as possible to the power and ground leads. That way, the capacitor can supply the little current pulses and prevent the noise from getting out into the system. We usually use a high quality 0.01μF ceramic capacitor for every one or two CMOS packages. Large CMOS chips such as microprocessors may need more capacitance and desktop CPU chips may need their own power supply regulators separate from the rest of the circuit.

Example

If the current pulse is 10mA for 10nS and the operating frequency is 10kHz then there are two current pulses in each 0.1mS cycle. That means that a total charge of $2 \times 10\text{mA} \times 10\text{nS} = 200\text{pC}$ flows in each period giving an average current of $200\text{pC}/0.1\text{mS} = 2\mu\text{A}$. That is a very low current and could be drawn from batteries for months or years without problems.

When the operating frequency is 10Mhz, the time per period is 0.1μS instead of 0.1mS and the average current is now 2mA which is getting serious.

This kind of speed/power relationship can be seen in the power supply characteristics for the PIC16C56, a small single chip microcomputer. It draws only 80μA at 32kHz but the current drain rises to 9mA at an operating frequency of 20Mhz.

12.5 More complicated gates

As we shall see in the next few chapters, there are thousands of more elaborate logic switches that we can build in the same CMOS technology as the inverter. We shall examine only two more in detail, looking at the transistors from which they are formed, as these are the building blocks from which all the rest are made.

12.5.1 NAND

Figure 12-11 (left) shows the circuit for a CMOS NAND gate.

Since this circuit has two input wires and each input can be in one of two states, high or low, there are four possible input combinations or **states**. We will work out what happens for each of the four possible input states.

Input A = Low, Input B = Low Output HIGH

The gates of both the NFETs are at ground potential so that the two NFETs are both turned off. The gates of both PFETs are also at 0V so that there is a -5V gate-source potential to turn them on. That means that the output is connected to the 5V supply through two parallel low resistances and is disconnected from ground.

Input A = Low, Input B = high Output HIGH

The gate of the upper NFET, Q3, is at 0V so Q3 is turned off. Because input B is high, the lower NFET, Q4, is turned on. Since Q3 and Q4 are in series, the output is disconnected from ground. In the upper half, Q1 is turned off because its gate is high, at the same voltage as its source. Q2 is turned on as before. Q1 and Q2 are in parallel and a very low resistance in parallel with an infinite resistance is a very low resistance so the output is connected to the 5V supply and isolated from ground.

Input A = High, Input B = Low Output HIGH

This is just the same as the previous case. One of the series NFETs is turned off so the output is isolated from ground and one of the parallel PFETs is turned on so the output is connected to 5V through one low resistance.

Input A = High, Input B = High Output LOW

This time both NFETs are turned on. Q3 is turned on by the 5V on its gate while its source is at 0V. That causes Q3 to conduct heavily so the voltage at its drain falls to nearly 0V and brings the source of Q4 to 0V. The 5V on input A turns on Q3 and the output is connected to ground through two turned-on FETs. Both upper PFETs are turned off because their gates are at 5V, the same potential as their sources. That means that the output is connected to ground and isolated from 5V.

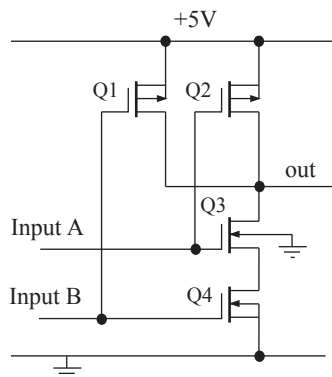


Figure 12-11 CMOS NAND Gate

Note Where two or more wires connect there is a blob. Where two lines merely cross there is no electrical connection; the wires merely pass each other. The substrate of the upper NFET is shown tied to ground rather than being connected to the source. This is a direct result of the way the FETs are made on a single piece of silicon. It has the desirable effect of making sure that the substrate is always the most negative point in the device and so holding the internal p-n junctions in reverse bias.

A	B	Out
L	L	H
L	H	H
H	L	H
H	H	L

Table 12-2: NAND Logic

We can summarize this behavior in a **truth table**; a table that shows the output state for all possible input states. We can write the truth table in several different ways. Table 12-2 presents the table with the two states labeled H and L. This is an un-ambiguous method of displaying the behavior since it does not depend on the choice of logic. A high level is a high level whether it represents a 1 or a 0. Since we usually use positive logic it is common to show the truth table with 1's and 0's as in Table 12-3.

Figure 12-14 shows the symbol for a NAND. We recognize the little circle at the output as our symbol for an inverter and so deduce correctly that this symbol is made up of some other symbol followed by an inverter. We shall return to this in the next chapter.

A	B	Out
0	0	1
0	1	1
1	0	1
1	1	0

Figure 12-12 NAND Symbol

Table 12-3: NAND Truth Table

12.5.2 NOR

Figure 12-13 shows the other type of basic gate. As you can see it has a structure like the NAND gate except that now the NFETs are in parallel and PFETs in series. This is a **nor** gate. It has the property that the output is low unless both inputs are low.

Input A = Low, Input B = Low Output HIGH

Both NFETs are turned off by the low voltage on their gates. Both PFETs are turned on by the 0V on their gates making their gate-source voltages -5V. Thus the output is connected to 5V through a low resistance path and isolated from ground.

Input A = Low, Input B = High Output LOW

Q3 is turned on and Q4 is turned off so there is a low resistance path to ground from the output. Q2 is turned off while Q1 is turned on. Because the PFETs are in series the output is isolated from 5V.

Input A = High, Input B = Low Output LOW

This time Q4 is turned on and Q3 turned off so there is again a low resistance path to ground from the output. Q1 is turned off and Q2 is turned on. Because the PFETs are in series the output is isolated from 5V.

Input A = High, Input B = High Output LOW

Now, both NFETs are turned on so there is a low resistance path to ground from the output. This time both PFETs are turned off so again the output is isolated from 5V.

That gives us the truth tables of Table 12-4 and Table 12-5.

Figure 12-14 shows the standard symbol for a NOR gate. Again, it has the form of another symbol followed by an inverter.



Figure 12-14 NOR Gate Symbol

12.6 Connecting Switches Together

Individually, logic switches have very limited utility. Their real power emerges when we can connect several switches together to form more complicated logic circuits. We shall study the mathematics of logic circuits in the next chapter but first we need to look at how to connect two switches together.

12.6.1 Naive Interconnection

So long as two different switches operate from the same power supplies it would seem to be trivial to interconnect them. For example, we could combine a NAND gate with a NOT gate to make an AND gate like that in Figure 14-17.

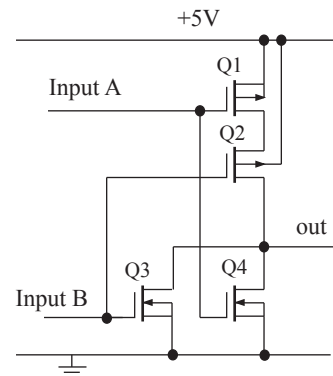


Figure 12-13 CMOS NOR Gate

Note This time it is the substrate of the lower PFET that is not connected to its source. Instead, it is connected to the positive power supply. In part this is a result of the way in which the transistors are grown on a single piece of silicon but it is also a necessary connection to keep the substrate more positive than any other terminal of the FET.

A	B	Out
L	L	H
L	H	H
H	L	H
H	H	L

Table 12-4: NOR Logic

A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0

Table 12-5: NOR Truth Table

When you connect gates like this, it is vital that the ground lines of the two gates be connected together so that both switches agree on the meaning of 0V. However, it would be possible to have the NAND and NOT gates supplied by two different 5V power supplies. All that matters is that the voltages be the same and that they agree on the meaning of 0V.

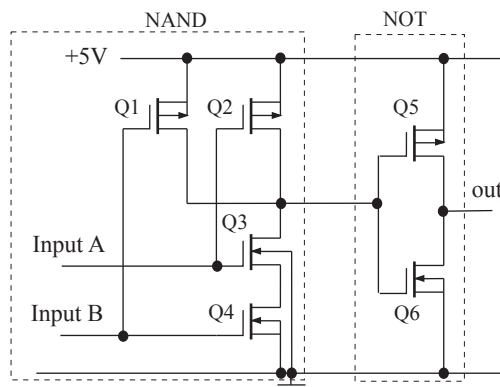


Figure 12-15 NAND+NOT=AND

This circuit looks as though it will work well, since the output levels of the NAND gate are perfectly matched to the input levels of the NOT gate. In practice, however, the unavoidable stray capacitances of the FET make this connection impractically simplistic. In order for the FETs to turn on and off very rapidly we have to charge and discharge their input capacitances very quickly. This means that we want a large pulse of current to come out of the NAND gate and we want to have the input capacitances of the NOT gate be as small as possible.

Unfortunately, to get large pulses of current we have to use physically large FETs and those large FETs necessarily have large input capacitances. Thus we need to make our switches out of FETs that are both large and small!

12.6.2 Buffered Logic

Info A **buffer** is a circuit that connects two subsystems together in such a way that the receiving circuit does not alter the output of the sending circuit. It usually turns out to be a kind of amplifier that does not alter the voltage of a signal but increases the current. That way the sender does not have to supply any current but the receiver can draw significant current from the signal.

Logic switch manufacturers solved this problem by adding extra stages of transistors called **buffers**. Figure 12-16 shows the complete circuit for a buffered NAND gate. Transistors Q1-Q4 form the basic NAND circuit. These are very small FETs with small input capacitances, fast switching times, and limited current carrying ability. Transistors Q7 and Q8 form the output buffer. These are larger FETs, capable of supplying much more current to the outside world at the cost of larger capacitances and slower switching times. Since Q7 and Q8 form an inverter we have to add Q5 and Q6 to recover the NAND function. These are made of the same sort of small FETs as Q1-Q4.

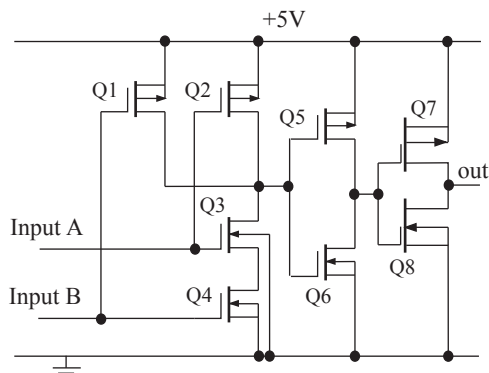


Figure 12-16 Buffered NAND gate

Buffered gates have several advantages over their unbuffered equivalents. They can provide more current to an external circuit than an equivalent speed unbuffered device. They turn out to be less sensitive to noise on the signal and power lines, and they allow a single output to drive several inputs without slowing the switching time. There is a slight loss of speed caused by the extra stages and large output FETs, but this is usually negligible because the input stages

can be made so much faster. In a device of any complexity, the delays in the input and intermediate stages usually dominate the switching time.

With the exception of the specialized 74HCU04 unbuffered hex inverter, all the 74HC family devices (see below) use standardized output buffers and so all have exactly the same output characteristics. They are guaranteed to deliver up to 4mA to an external circuit at either $V_{out} = 0V$ or $V_{out} = 5V$.

Note The unbuffered inverter is not intended for use as a logic gate. Instead, it operates as a high gain inverting amplifier (see Chapter 19) and is usually used to build oscillator circuits.

12.6.3 Connecting Two Outputs

One crucial restriction on the logic switch circuits that we have seen so far is that it is fatal to connect two outputs together. There are times when it seems as though it would be very convenient to connect several different outputs to a single input and only use one of them at a time. The trouble is that there is no way for one of our switches to produce **no** output. The output is always either a 1 or a 0. That is the whole point of a binary output. So long as the two outputs always agree on the value of the output there is no problem. Indeed, we occasionally parallel the outputs of two identical gates, with identical inputs, in order to deliver more current to a circuit. However, if the outputs disagree then disaster strikes.

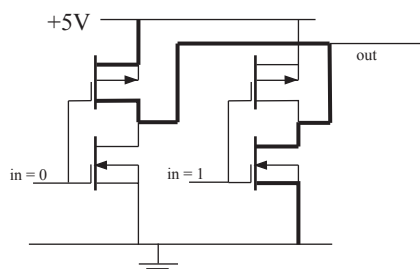


Figure 12-17 How NOT to connect two outputs

As shown in Figure 14-19 above, a path is created directly from +5V to ground through two turned-on FETs (bold lines). A very large current flows and both devices are destroyed.

Remember

Never connect two standard logic outputs together.

The only correct way to connect two normal CMOS outputs to a single input is through a **multiplexer**, as saw in section 14.4. However, the idea of connecting outputs is sufficiently useful that two modified output circuits have been developed to allow multiple outputs to be connected together in certain circumstances.

12.6.4 Open-Drain Outputs

The simpler, and older, of the two alternative output circuits is called the **open-drain output** or, for historical reasons, the **open-collector output**. An open-drain output is exactly what it sounds like. Instead of the output being taken from a pair of complementary FETs, it is taken from a single un-connected FET drain. Figure 14-21 shows the internal circuitry of one gate from a 74HC03 quad 2-input NAND gate with open-drain outputs.

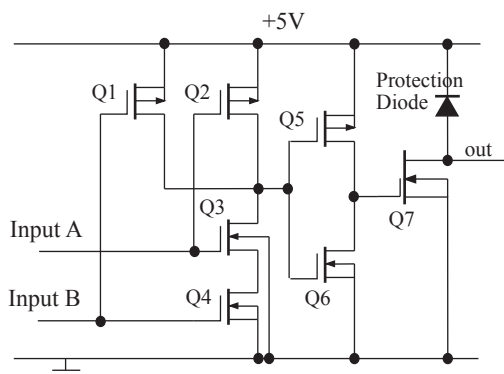


Figure 12-18 Open drain 2-input NAND gate

The output totem-pole has been replaced by a single n-channel FET. The diode connected to the positive rail plays no part in normal operation of the device. It is just there to prevent damage to the output FET. So long as the output voltage is less than +5V the diode is turned off and so acts as an open circuit. If some external circuit tries to pull the output to a voltage above 5V then the diode will turn on and prevent V_{out} from rising more than about 0.6V above +5V, thus preventing overvoltage damage to the output FET.

By itself this gate has two output states, $V_{out} = 0$ and $V_{out} = \text{unconnected}$. We can recover the logic 1 state by connecting a resistor (usually about 10k) to +5V and then the voltage at the output will switch between 0V and 5V according to the standard NAND truth table.

The magic comes when we connect two of these open-drain outputs to a single input. (Figure 14-22, left). Let us look at the truth table in terms of the intermediate points a and b. When $a = 0$ and $b = 0$ the outputs of the buffer inverters are both 1 and so the two n-FETs are turned ON. Each wants the output to be 0V and they are both happy with $out = 0$. Current flows through the output drain resistor to ground.

When $a = 0$ and $b = 1$ the upper inverter output is 1 and so the upper n-FET is turned ON. The lower inverter output is 0 and the lower FET is turned OFF. Since the lower circuit is now totally disconnected the output is set entirely by the upper circuit. It wants the output to be 0 and so we find $out = 0$. Again, there is current flow in the output drain resistor.

When $a = 1$ and $b = 0$ the situation is exactly the same except that now the lower FET is the only one turned on. Again $out = 0$.

When $a = 1$ and $b = 1$ both inverter outputs are 0 and both FETs are turned off. In this case no current flows in the output resistor and so there is no voltage lost across it. The output is pulled up to +5V. For this reason we call the output resistor a **pull-up resistor**.

The resulting truth table is shown in Figure 12-11. We recognize it as an AND gate and refer to this connection format as **wired-AND**.

Open drain gates have the big disadvantage that they draw static current when their outputs are in the 0 state and they are rarely used. The exception is found in a few circuits that connect directly to non-logic devices. Because the output n-FET is made rather larger than is usual, the output current in the low state can be up to 25mA instead of the 4mA limit of a standard CMOS output. Thus these devices are sometimes used to drive small lamps or large LEDs.

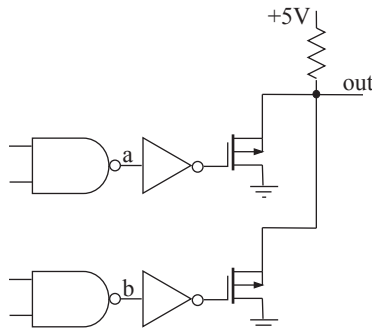


Figure 12-19 Using open-drain outputs

Table 12-6:
Wired AND

a	b	out
0	0	0
0	1	0
1	0	0
1	1	1

12.6.5 Tri-State Outputs

The second solution to the problem of connecting outputs was developed by National Semiconductor Corporation who gave it the trademarked name TRI-STATE. A tri-state output is a standard CMOS totem-pole output with extra logic circuitry to provide a third output state in which both output FETs are turned off (Figure 14-23).

This is a little more complicated than the open-drain output but we can work out its truth table with a little care.

When OE is high, the upper input to the NAND gate is 0. Examination of the NAND truth table shows us that in this case NAND output $b = 1$ and the p-FET is turned OFF. At the same time, the lower input of the NOR gate is a 1. Examining the NOR truth table we find that in this case NOR output $a = 0$ and the n-FET is turned OFF. Since BOTH of the output FETs are turned off, the output acts as though it were not connected. It is said to be in a **high-impedance state**.

When OE is low, the upper input to the NAND gate is 1. In this case we find that $b = in$. Similarly, the lower input to the NOR gate is a 0 and we find that $a = in$. A little thought shows that in this case $out = in$. The device is a simple straight-through buffer. Table 12-7 shows the resulting truth table. Note that the z marks the high-impedance states.

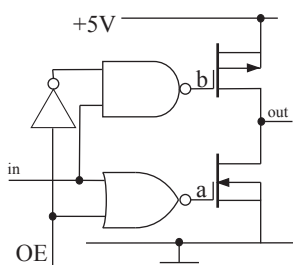


Figure 12-20 TRI-State CMOS Output

Info The final output totem-pole of a tri-state gate uses the larger, higher current FETs that we would expect to find in the output of a buffered CMOS gate while the extra gates involved in the tri-state switching use the smaller internal FETs. Thus real tri-state outputs are buffered in just the same way as regular totem-pole CMOS outputs.

The symbol for a tri-state buffer (Figure 14-24) is just the usual buffer triangle but with an extra wire coming into the triangle at an odd angle. These devices are available in various combinations of inverting/non-inverting and enable = high or enable = low.

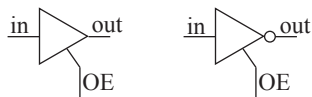


Figure 12-21 Non-inverting and Inverting Tri-state Buffer Symbols

Table 12-7: Tri-State Buffer

OE	In	Out
0	0	z
0	1	z
1	0	0
1	1	1

Many kinds of CMOS device, from simple gates to complex microprocessors, are available with tri-state outputs and there are special tri-state buffer chips such as the 74HC240, octal tri-state inverter, and 74HC241, octal tri-state buffer, that can make standard signals into tri-state signals. As we shall see later, these devices are found throughout modern digital computers and make possible such tricks as a single wire that sometimes carries signals from left to right and sometimes carries signals from right to left.

Example

Let us revisit our computer memory. We can convert it to a working system by adding tri-state buffers to the memory circuits and some extra circuitry to select which memory is active.

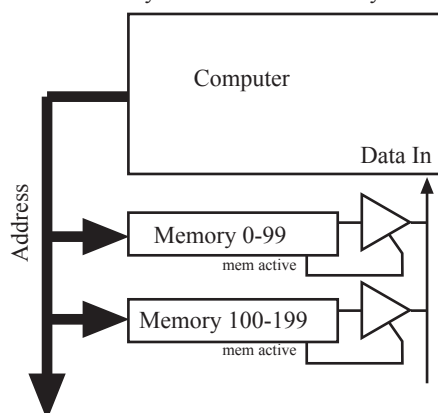


Figure 12-22 Improved Computer Memory

Each memory circuit now generates an extra signal, `mem_active`, that is true if the current address lies inside that circuit. Since only 1 memory can hold any given address, at most 1 such signal will be true at any instant. Which ever memory is active gets to set the state of the Data In line without any competition from the inactive memory circuits. They all have their output buffers in the high-impedance state and act as though they were disconnected from the circuit altogether.

In Chapter 15 we shall see one way to generate the `mem_active` signals needed for such a system. Apart from the unrealistically small sizes of the memories, this is now an accurate picture of how one bit of a computer's memory works when the computer is reading from memory.

12.7 Integrated Circuit Switches

Although the first transistorized computers, e.g. the UNIVAC model 80 and the CDC 1604, were built with individual transistors in the logic gates, a way was soon found to put several components on a single piece of silicon. That meant that you could put several gates in the same space, and with the same power consumption, as a single gate built with discrete components. A whole circuit built on a single chip of silicon was called an **integrated circuit**.

The individual components in an integrated circuit can be closely tailored to serve only one purpose since they are built into the circuit in which they will be used. This means that integrated circuits are not only smaller and less power hungry than discrete circuits but they can also be made to perform better since they are not made from components that have to be able to serve in many different circuits.

Info Integrated circuits first made their appearance in commercial systems in the mid 1960's. They started in specialized applications and slowly took over the whole system. Integrated circuit memory devices first appeared in the IBM 360 Model 91 where they were used as a small part of a larger magnetic memory. The first commercial all semiconductor memory seems to have arrived with the IBM 370 Model 145 in September 1970, which was probably the first all IC computer.

Example

Our discrete CMOS inverter takes about 20nS-155nS to turn on and off. A typical integrated circuit CMOS inverter, the 74HC04, claims a typical switching time of only 8nS, a factor of >2-20 better. The latest generations of CMOS integrated circuits feature extraordinarily small FETs that can switch still faster. The internal circuitry of a high-end desktop CPU in 2008 runs its switches at rates of more than 3GHz (3000MHz) and so requires switching times much shorter than 0.3nS. These FETs have channels that are only about 0.045 μ m long. That is only about 300-600 atoms in length!

At first, the number of components that could be put in a single integrated circuit was quite small; about a dozen transistors and resistors. Very rapidly ways were found to increase the number of components so that the number of transistors on a single chip has doubled roughly every 18 months since the early 1960's. 50 years later we have logic chips with more than 2 billion transistors on them and memories with over 64 billion transistors. Obviously the individual transistors have had to get smaller and smaller and that has had two benefits. First the smaller transistors have smaller inter-electrode capacitances so that they can turn on and off much faster. Second, the smaller the transistor, the less power it uses. That means that a modern notebook computer can run on batteries for up to 8 hours while providing thousands of times more computing power than a 1960s IBM mainframe computer that took up a whole room and required massive air-conditioning to get rid of the kilowatts of heat that it emitted.

12.7.1 Integrated Circuit Logic Families

Info A logic family is a set of integrated circuits that offer a wide range of functions and that are designed to be easy to connect together. All the members of a family operate from the same power supply (most often 5V), use the same definitions of logic 0 and 1, and share common speed and power dissipation characteristics.

The first integrated circuits were constructed the older style bipolar transistors mentioned at the start of Chapter 11. The very first logic family, RTL (**resistor-transistor logic**) used circuits very like our first NMOS switch (Figure 12-6) and suffered from similar problems. They were soon replaced by the TTL family (**transistor-transistor logic**) that uses only bipolar transistors in much the same way that CMOS uses only FETs. This family rapidly established itself as a standard and was in universal use at the beginning of the 1970s (except for some very esoteric high speed applications).

The most popular family of TTL logic devices came from Texas Instruments. Starting with the 7400 quad two-input NAND gate, they gave us a device naming scheme that endures to this day. Over the next decade new families of TTL devices became available that operated at higher speeds and used less power than the original 74xx series. The IC manufacturers cleverly adopted a naming scheme that was based on the original 74xx devices. Each device had a number that began 74 followed by some letters that identified the family, and the 2 or 3 digit code that told you what the device did. So '00 was always a quad 2-input NAND, '04 always a hex inverter (6 NOT gates), and so on. Moreover, the new devices were pin compatible with the old so that you could take a board designed for 74xx devices and plug in a set of 74LS devices and suddenly have a board that ran a little faster but used only 1/10th the power.

Meanwhile, RCA developed the first CMOS logic family, the 4000-series (because the numbers began at 4000 and went up slowly as more complex devices were released). These were extremely difficult to work with since tiny static charges could destroy the devices. They were also very expensive and very slow (they took about 1 μ S to switch!) but they drew almost no current from the supply. Unfortunately, there was no relationship between the part numbers of the 4000 series and those of the TTL 74xx series.

In the 1980s, dramatic improvements in FET design lead to the development of the modern high-speed CMOS devices that we shall be using. These devices were made to be directly compatible with 74 series TTL and so use the same family naming structure. Thus a 74HC04 is a CMOS hex inverter that is pin compatible with the original 7404.

Warning You should not mix 74HC devices directly with older TTL devices because the older devices have rather poor logic 1's (they are only guaranteed to be >2.4V) and often cannot switch the newer 74HC devices. However, modern high-speed CMOS devices are in almost all respects superior to the older TTL versions and are clearly the devices of choice for all new designs.

As we have seen, the CMOS devices have the happy characteristic that they draw current only when they change state and so are ideal for low-power devices operating at low speeds. However, the average current draw increases as the operating frequency increases. Once the clock speed gets too high, the power draw of CMOS can get significant. The latest trend is to lower the operating voltage of very high speed logic devices in order to reduce total power needs.

Example

An Intel Pentium4 operates internal circuits from a 1.5V power supply and it interfaces to the outside world with logic that operates at 2.5V. It still manages to consume about 30W of power running at a clock speed of 1GHz!

The operating characteristics of some common families are summarized in Table 12-7. If you are interested in the history of logic families then I would recommend the excellent discussion found in chapter 9 of Horowitz and Hill.

Table 12-8: Some common logic families

Family	Power (mW/tgate)	Speed (MHz)	Popular	Comment
74xx	high (10)	~20	1970-1980	The original: power hungry and fairly slow.
74LSxx	fair (2)	20	1976-1986	A lot less power hungry but no faster.
4000B	low (0.3)	0.3	1970-1984	Low power but very slow.
74HCxx	low (0.5)	30	1984-present	Low power but usable speed
74ACxx	low (0.5)	125	1990-present	Faster but more expensive version

All new circuits designs should use the 74HC or 74AC families.

12.7.2 Logic Complexity

Totally separate from the classification of logic IC's by family, there is a classification by complexity. This is measured in terms of the number of individual gates packed into a single piece of silicon.

SSI: Small-scale integration logic puts up to about ten gates in each package. It provides elementary functions so that each package usually contains several copies of a single simple type of gate. Examples include the 7400 quad 2-input NAND, the 7402 quad 2-input NOR, the 7404 hex inverter, and so on.

MSI: Medium-scale integration logic is a little more complex. Each package contains up to a few dozen gates. A single package holds a complete sub-circuit such as a decoder or a multiplexer. Examples include the 7447 that displays a number on an LED and the 7490 that can count up to 10.

LSI: Large-scale integration provides much more complex functions. A single LSI chip may have hundreds or even a few thousand gates. Small single-chip computers, disk-drive controllers, keyboard controllers, and older memory chips are examples of LSI circuits, as are programmable logic devices or PLDs.

VLSI: Very large-scale integration sits at the top of the complexity ladder. A VLSI chip contains anywhere from tens of thousands of gates up into millions of gates. A single VLSI chip may replace several complete LSI chips and their associated glue logic. Modern high powered microprocessors such as Intel's Pentium descendents and the various kinds of ARM are VLSI chips as are today's massive memory chips which are already reaching past 1 billion bits of memory on a single silicon chip.

SOC: System-on-a-chip is the building of complete systems on a single piece of silicon. These devices may have hundreds of millions of transistors on them. They may include not only a complete CPU and some memory but also a set of peripherals that would otherwise require their own VLSI chips. SOCs power iPods, iPads, many TV set-top boxes, many ultrabooks, and the ubiquitous smartphones. The most popular processor for these is the ARM, which is a design produced by a company that makes no hardware but simply licences the design to many other companies to manufacture into their own systems. Most of the major semiconductor houses have at least one family of ARM chips in their catalogue.

Over the past twenty years the commercial world has almost abandoned MSI and SSI devices because it takes many packages to make even a simple circuit. Each packages has to be installed on a board and connected to all the other packages. All this pushes costs up. The

modern trend is to put all of the logic into as few chips as possible. For example, a computer motherboard that had 150 separate IC packages on in in 1985 now has only a handful, despite the fact that all the peripherals that filled the card slots in 1985 are now on the mother board!

Part of the trend to more complex chips is the emergence of programmable logic chips. These put hundreds or thousands of gates on one chip along with a method of interconnecting them. One of these chips can be programmed to perform a vast range of different functions that would formerly have been implemented with a large number of SSI and MSI devices (c.f. Chapter 15).

12.7.3 *Small-scale logic*

The smallest SSI circuits provide individual logic gates. These were the first chips built and also the first to be replaced. The original 7400 family contained a bewildering array of these including such oddities as the 7451 expandable 2-wide 2-input and-or-invert gates and their sister 7460 dual 4-input expander gates. The appearance of MSI logic and the rise of programmable logic have reduced the need for SSI functions. Only a few simple devices remain in today's logic families. This loss of so many devices accounts for the large gaps in the 74xx numbering scheme.

There is still an occasional need for tiny amounts of logic and there are now logic families that take this to the logical extreme and package a single NAND or NOR gate into a very tiny surface mount package. Many of these newer families also operate from drastically reduced voltages. 3.3V logic is now common and devices are available down to total power supply voltages near 1V so that they can be operated from a single cell battery.

Here are some of the remaining SSI chips from the 74HCxx series of functions.

Table 12-9: Some 74HCxx SSI devices

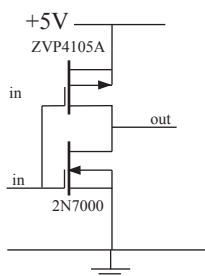
74HC00	Quad 2-input NAND	This is one of the commonest SSI chips.
74HC02	Quad 2-input NOR	
74HC03	Quad 2-input NAND open-drain	For use when you need to connect outputs or drive LEDs.
74HC04	Hex inverter	Six NOT gates in a package. Probably the most used SSI chip.
74HC08	Quad 2-input AND gate	
74HC32	Quad 2-input OR gate	
74HC125	Hex TRI-STATE buffer	Makes standard outputs into tri-state outputs.

Summary

Pairs of complementary FETs, that is p-type and n-type FETs can be connected to form a variety of simple switching circuits. These circuits offer extremely low static power dissipation and can operate at very high speeds. By adjusting the geometry of the switches we can build switches that form different logical combinations of their inputs.

We display the overall operation of logic switches using **Truth Tables** that show the logical output for each possible combination of logical inputs. The truth table shows logical values. The logical value 0 is usually represented by 0V and the value 1 by +5V.

The simplest CMOS logic switch is the inverter shown in the left figure below. It implements the logical **NOT** operation shown in the truth table in the middle and it is represented by the symbol on the right.



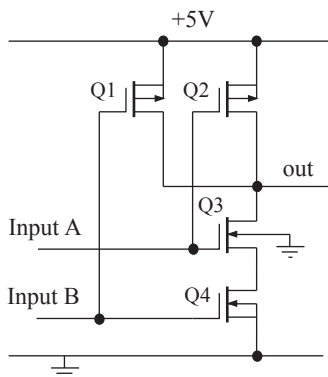
A	Out
0	1
1	0

NOT Truth Table



NOT Gate

Next is the CMOS NAND shown below left. It implements the $OUT=NOT(A \text{ AND } B)$ function described by the truth table in the middle and has the symbol on the right.



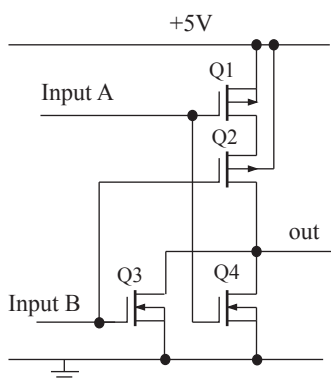
A	B	Out
0	0	1
0	1	1
1	0	1
1	1	0

NAND Truth Table



NAND Gate

Its complement is the CMOS NOR gate, which implements the $OUT=NOT(A \text{ OR } B)$ function described by the truth table on the upper right in the figure below. Its circuit is shown on the left and it is represented by the symbol on the right.



A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0

NOR Truth Table



NOR Gate

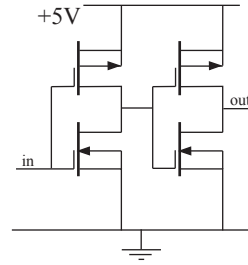
Integrated circuits pack several of these gates into a single package. They usually also include extra CMOS stages called **buffers** that improve the ability of the gates to drive loads.

Some integrated circuit logic gates have more complicated tri-state outputs. These have an extra output state in addition to 0V and +5V. In this third state the output is completely disconnected from the circuit. This state is called the **high impedance** state. If precautions are taken to make sure that only 1 tri-state output is active at a time then it safe to connect multiple tri-state outputs together.

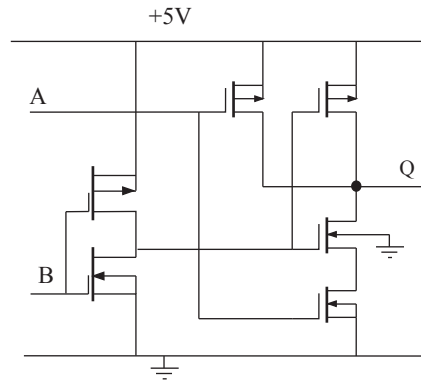
You must **NEVER** connect ordinary logic outputs together. If the outputs disagree on the output voltage then both chips will be destroyed.

Exercises

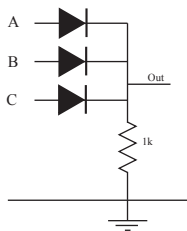
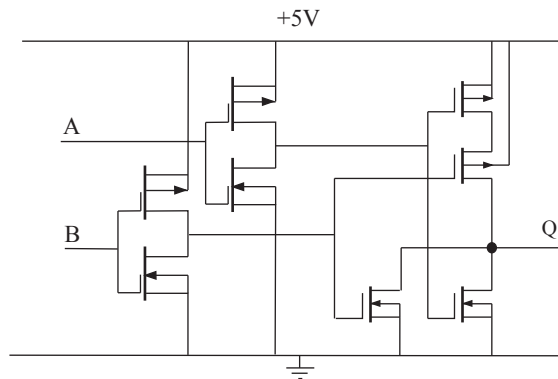
1. Find the truth table for the circuit on the right.



2. Draw the truth table for the following gate and write a valid logic expression for it.



3. Draw a truth table and give a logic equation for the circuit on the right



4. In the diode circuit on the left, each of the inputs, A, B, and C can be either 0V or +5V. Construct a truth table showing the output as a function of the 3 inputs and identify the logic function that this circuit performs. This is a Diode Logic circuit and it is very obsolete!

Chapter 13:Digital Logic Theory

13.1 Introduction

At its heart, every digital computer is made up of NAND, NOR, and NOT gates like those that we have just studied in Chapter 12. By themselves, these are extremely simple operations. However, just as we can combine simple arithmetic operations to produce all the complexities of calculus, so we can combine these simple logic operations into logic circuits that can perform much more complex tasks. Before we go on to look at some of the integrated circuit logic devices that are available it is important that we take some time to study the algebra of these logic operations.

There are two fundamentally different types of logic circuit. In the simpler kind, the output of the whole circuit depends only on the current inputs. For example a circuit to add two binary numbers produces the sum of the numbers for its output and needs no other information. This is called a **combinatorial logic** circuit because its output is a combination of its inputs.

The more complex kind produces an output that depends not only on the current input, but also on the previous state of the circuit; it has a memory. For example, a counting circuit adds one to its output when the input tells it to. That means that the new output depends not only on the input but also on the previous number. Such a circuit is called a **sequential logic** circuit because its outputs go through a sequence of values, the details depending on the inputs. Combinatorial circuits are simpler to understand than sequential ones so we shall start with a study of combinatorial logic.

We face two different kinds of challenge in combinatorial logic as in most electronics; we have to be able to figure out what an existing circuit does (**analysis**) and to be able to design a new circuit from an idea of what it should do (**synthesis**). In this chapter, we will study both techniques. Analysis is fairly straightforward, using bit-following to turn a circuit diagram into a truth table. Synthesis is a multi-step process where we first turn an idea into a truth table, then use formal methods to turn the truth table into a Boolean expression that we can then turn into a circuit diagram. If the problem is a large one, then we may need to manipulate the truth table or the logic expressions to minimize the complexity of the circuit. We shall look briefly at some ways to do this but will find that really complex cases are best handled by a computer. Any circuit simple enough to be minimized by hand is probably simple enough not to need much minimizing.

13.2 Truth Tables

When we studied the basic logic switches in Chapter 12, we described their behavior using truth tables, tables showing the output of the circuit for every possible combination of inputs. Every combinatorial circuit has its own truth table and every truth table can be implemented with a combinatorial logic circuit. Unfortunately, although each combinatorial circuit has only one possible truth table, each truth table can be implemented by many different logic circuits. One of the main tasks of the digital logic designer is to find the best way to implement a given truth table as a logic circuit. Most of this chapter is about ways to do that but we begin with the simpler task of finding the truth table for a given task.

To find the truth table you must list all the possible input states of the system and decide, by inspecting the desired function of the circuit, what the output should be. Since the only possible choices for each input and output are the two binary values 0 and 1 this is not too arduous a

task, though it can get tedious. If the circuit has n independent inputs then there are 2^n possible states for the system.

Remember A logic circuit with n inputs has 2^n possible states. A good way to be sure that you have considered all the possibilities is to arrange the input states in increasing numeric order. If all the numbers from $0 \rightarrow 2^n - 1$ are present then you have found all the states.

When the number of inputs gets large it can become tricky to keep track of all the states and to be sure that you have not missed any. The usual way to keep track of this is to treat the inputs as the bits of a binary number and to put the numbers down in order. If there are no numbers missing then you know that there are no missing input states.

Once all of the input states have been identified then you can work back through the table and put in the outputs. If one or more of the input states will not be found in practice then it will not matter what the output is in those states. You can mark this by putting an X in the output for such a state instead of a 1 or 0.

A simple example may help to make the process clearer.

Example

The inside courtesy light of a car has to turn on if any of the car doors is open and stay off when they are all closed. Let us try to produce the truth table for such a circuit. We will assume, for simplicity, that we have a two door car so the circuit has two inputs, the left door switch, L, and the right door switch, R. Each of the door switches is a binary input; the switch can be closed to show that the door is closed, or open to show that the door is open. Let us assume that the switch produces a 1 when it is closed and that an output, O, of 1 from the circuit will light the inside lamp. With two inputs there will be four different input states with bit patterns 00, 01, 10, and 11.

Let us list the cases and translate inputs and outputs into binary terms.

Left door open, right door open, light on $\rightarrow L=0, R=0, O=1$

Left door open, right door shut, light on $\rightarrow L=0, R=1, O=1$

Left door shut, right door open, light on $\rightarrow L=1, R=0, O=1$

Left door shut, right door shut, light off $\rightarrow L=1, R=1, O=0$

Now we have collected all the information that we need. We write the truth table by putting the input states into arithmetic order and writing the table out as in Table 13-1, which is the truth table for our door light. In this simple case we recognize a truth table from Chapter 12. It is just a NAND gate.

Note The input states are arranged in increasing numeric order.

Table 13-1: Door Table

L	R	O
0	0	1
0	1	1
1	0	1
1	1	0

Example

A slightly more complex example is the kind of two-way switch that is often found on stairway lights. There is one switch at the bottom of the stair and a second at the top. If the light is off then you can turn it on by flipping either switch. Similarly, if the light is on you can turn it off by flipping either switch. In practice this kind of switch is implemented with two-way, single-pole double-throw, switches and some extra wire in the wall. If you want to do the same for a double flight of stairs so that there are three switches then it gets far too complex to implement with multi-way switches but we can do it easily with logic.

Let us call the three switches A, B, and C and the output to the light bulb O. Then we can construct the truth table by following the rule that flipping any one switch, inverting any one input, will make the light change state. If the light is on, then it turns off and vice versa. There are two tricks to producing this table. The first one is that we have to pick a starting state. Let's assume that if all the inputs are 0 then the light is off. (This is quite arbitrary. If we make the other choice then we will end up with an equally valid but slightly different circuit.) The second trick is making sure that we generate all the possible inputs. We can't go through the states in numerical order this time because of the way our switch flipping rule works. However, we can write the $2^3 = 8$ numbers down in order and then search through the input states for ones that satisfy the rule. Here are the numbers 000, 001, 010, 011, 100, 101, 110, 111.

We start by assuming that the output, O, is off

A=0, B=0, C=0, O=0

Now we flip one of the switches and change the output state. The very next state in the list differs in only one bit so our next state is

A=0, B=0, C=1, O=1

There are three states that differ from 001 in only one bit. They are 000, 011, and 101. We have already used up 000 so we shall pick the next one

A=0, B=1, C=1, O=0

This time we can go to 001, 010, and 111. Again we shall pick the next highest unused state and have.

A = 0, B = 1, C = 0, O = 1

That has exhausted the states that have A = 0 so it is time to flip A.

A = 1, B = 1, C = 0, O = 0

The next set of possible states is 011, 100, and 111. Again we shall pick the next largest unused stated

A = 1, B = 0, C = 0, O = 1

That leaves only 2 unused states, 101 and 111. Only the first is one flip away so we have

A = 1, B = 0, C = 1, O = 0

and so the final state must be

A = 1, B = 1, C = 1, O = 1

There that is eight states, the whole lot. Now let us put them in numerical order and write them out as the truth table of Table 13-2.

Well, we obviously aren't going to recognize that table straight off because it has three inputs and we have only looked at 2-input gates. Still it gives us the idea of how to construct truth tables.

Table 13-2: Three-way Light Switch

A	B	C	O
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Info The sequence of states that we went through to generate this table is called a gray-code sequence. In the accompanying computer text we shall see that this sequence is useful for encoders that translate positions into binary numbers.

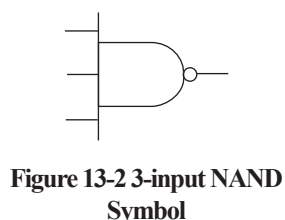
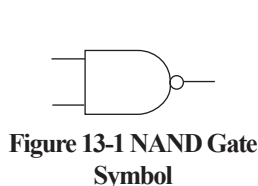
The next task is turning truth tables into logic circuits so that we can build them. We really need to collect a few more tools first so we will have a quick look at various simple gates and their truth tables. That way we shall have a little more to work with.

13.3 Some basic gates.

Because every 2-input truth table contains four rows, each different gate has a different 4-bit output column. There are $2^4 = 16$ different 4-bit numbers and thus 16 different 2-input gates. Only three of these gates, with their inverses, are interesting enough to warrant their own names.

13.3.1 NAND

Because of the associative property that we discuss below, multiple input NAND gates are possible and available. Here are the symbols and truth table for 2- and 3-input NAND gates. The extensions to four and more inputs are obvious.



A	B	NAND
0	0	1
0	1	1
1	0	1
1	1	0

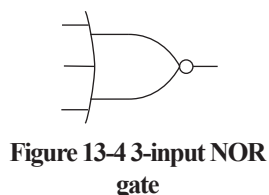
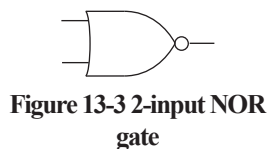
Table 13-4: NAND Truth table

A	B	C	O
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	0

Table 13-3: 3-Input NAND

13.3.2 NOR

Here are the symbols and truth tables for a 2-input NOR gate, taken from Chapter 12. Again, I have added the symbol and table for a 3-input NOR gate to show how it is done.



A	B	NOR
0	0	1
0	1	0
1	0	0
1	1	0

Table 13-5: NOR Truth table

13.3.3 AND, OR, and XOR

Each of the gates that we have met so far has a relative produced by adding a not gate to the output. If we negate the NAND gate, then we get the truth table of Table 13-7. This gate is called an AND gate because the output is true only when A is true AND B is true. It has the

A	B	AND
0	0	0
0	1	0
1	0	0
1	1	1

Table 13-6: AND Truth table

symbol of .Figure 13-5, which is like the NAND symbol but without the negation circle on the output.



Figure 13-5 2-input AND gate

We can see that the name and symbol of the NAND are derived from the AND gate. NAND is simply NOT AND and the symbol is made up from the AND symbol followed by an inverter. From a logical point of view the AND gate is more basic, but it takes fewer FETs to make a NAND gate. A real AND gate is made by following a NAND gate with an inverter!

Just as NAND is really NOT AND, so NOR is really NOT OR. Table 13-8 is the truth table for a 2-input OR gate, which has the symbol of Figure 13-5. As with NAND and NOR, we can just as easily extend these to 3 or more inputs.



Figure 13-6 2-input OR gate

A	B	OR
0	0	0
0	1	1
1	0	1
1	1	1

Table 13-7: OR Truth table

Now, this OR is a little different from the OR of common English. This is what is called an **inclusive or** because it is true when A is true, or B is true, or both are true. In common English speech, OR implies a choice: A or B but not both. This commoner English form is rarer in logic where it is called **exclusive or**. Exclusive OR, or XOR, has the symbol in Figure 13-6 and the truth table in Table 13-9.



Figure 13-7 2-input XOR gate

A	B	XOR
0	0	0
0	1	1
1	0	1
1	1	0

Table 13-8: AND Truth table

Just like AND and OR, XOR has its negative equivalent, the unpronounceable XNOR (it sounds almost like snore!). This interesting name stands for **exclusive NOR** but it really means **not exclusive OR** as you can see in the truth table (Table 13-) and symbol (Figure 13-7).



Figure 13-8 2-input XNOR gate

A	B	XNOR
0	0	1
0	1	0
1	0	0
1	1	1

Table 13-9: AND Truth table

13.4 Multi-gate circuits

Out of these 6 basic gates and our old friend the inverter we can build everything else. Let us look at how we combine gates to form new functions and how we find the truth tables for the new functions. Figure 13-8 shows a complex gate made of several simpler gates. You can see how we wire the input and output terminals of the individual symbols together to get more complicated logic functions just as we wire individual resistors, capacitors, and diodes together to get complex circuits. The final circuit has two inputs, A and B, and a single output, O.

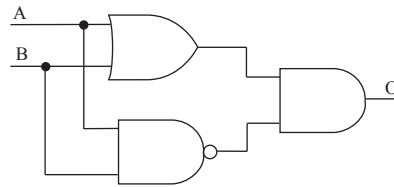


Figure 13-9

We can find out what this circuit does by the process of **bit following**. In this we build the truth table step-by-step. We set up a particular input state and then propagate the bits through the gates, using the truth table of each gate to work out its output as we go. Here is the bit following process applied to this circuit. We will go through the input states in numerical order so that it is easier to build the table. The first step is to set $A = 0$ and $B = 0$.

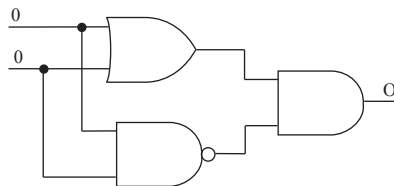


Figure 13-10

We propagate these bits along the wires to the inputs to the first layer of gates.

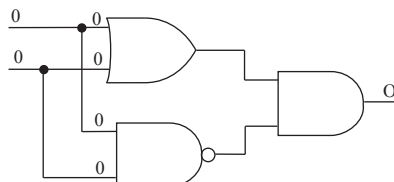


Figure 13-11

Now we use the truth tables for the OR and NAND gates to fill in the output bits from those two gates. $0 \text{ OR } 0 = 0$ and $0 \text{ NAND } 0 = 1$.

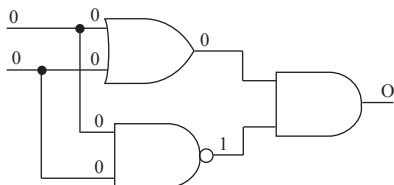


Figure 13-12

We propagate the outputs from the first layer forward to the inputs of the final AND gate and apply its truth table to get $0 \text{ AND } 1 = 0$.

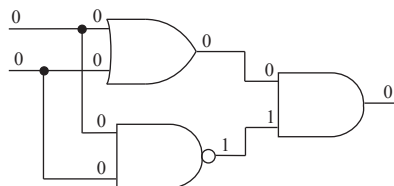


Figure 13-13

So we have enough information to construct the first line of our truth table (Table 13-11).

The next input state is $A = 0, B = 1$. Figure 13-13 shows what happens if we follow through this state.

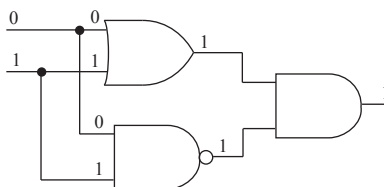


Figure 13-14

That seems to have given us two rows of the truth table. Actually we have three. Since the circuit is completely symmetric, the output for $A = 1, B = 0$ must be the same. That leaves only one state. Figure 13-15 shows the result of bit following through the state $A = 1, B = 1$.

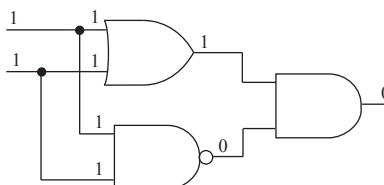


Figure 13-15

Table 13-10:

A	B	Out
0	0	0
0	1	1
1	0	1
1	1	0

We have now found the output for every possible input and built the truth table of Table 13-11, which we at once recognize as the truth table for an XOR gate.

So this circuit is one way (there are many others) of making an XOR gate out of simpler NAND, NOR, and NOT gates (viewing the OR as a NOR followed by a NOT etc.).

13.5 Boolean Algebra

Bit following can get tedious for large combinatorial circuits and we would like a shorter way to understand the working of a circuit. To find that way we have to learn a little about the algebra of these binary operations. This algebra was developed in the 19th century by George Boole and is called **Boolean Algebra**. In Boolean algebra we manipulate symbols and operators that stand for binary values and binary gates in a way that looks a lot like ordinary algebra. This algebra will allow us to analyze circuits without bit following in most cases and will lead us to methods for finding a circuit to implement any truth table.

13.5.1 Variables

Just as in ordinary algebra, we use names to represent binary values. The difference between algebraic variables and binary variables is that a binary variable can take on only the values 1 and 0. A, B, Closed, and Ready are all examples of possible binary variables.

13.5.2 Operators

The operators of Boolean algebra are the simple gates that we have already met, AND, OR, and NOT. Unlike conventional algebra, Boolean algebra does not have a single, uniform notation. Different authors have used many widely different notations and I will list some of them but I will use what I think is the most common and most useful notation.

NOT

The expression NOT A can be found written in at least the following ways in various different books.

$$\text{NOT } A, A^*, A', *A, 'A, /A, -A, !A, \bar{A}$$

I prefer, and will use, the last one. It is particularly easy to write but it used to cause problems for printers so older books tend to use one of the others. On rare occasions I will use the !A form.

OR

Here are some of the ways that OR can be written.

$$A \text{ OR } B, A | B, A \cup B, A + B$$

Again, I prefer the last symbol because it reminds us that OR operation in Boolean algebra behaves a lot like the addition operator in regular algebra. If we look at the standard arithmetic operator and make a sort of truth table for it, then we get Table 13-12.

**Table 13-11:
Plus table**

A	B	A+B
0	0	0
0	1	1
1	0	1
1	1	2

Here we see that the first three rows are exactly the same as the first three rows of the Boolean OR operator. The last one clearly isn't—you can't have a 2 in Boolean! There are only two things that could go in that last place, 0 or 1. The two different choices lead to the exclusive and inclusive OR operators of Boolean algebra so we use + symbols for them. Here are the usual forms for the XOR operation; again the preferred one is last.

$$A \text{ XOR } B, A \sim B, A \oplus B, A \oplus B$$

AND

Here are some of the ways to write AND.

$$A \text{ AND } B, A \& B, A \cap B, AB, A \cdot B, A \times B$$

Again, I prefer the last forms. Since I have been carefully putting the dot or a \times in for ordinary multiplication, I will use those for the AND operation as well. Once again a look at a little bit of the multiplication table (Table 13-13) will show you why we use a multiplication symbol for AND.

**Table 13-12:
Times table**

A	B	A·B
0	0	0
0	1	0
1	0	0
1	1	1

13.5.3 Expressions

Now we have the tools to write Boolean expressions. Let us begin by translating our three gate version of the XOR gate into an expression. Here is the circuit again.

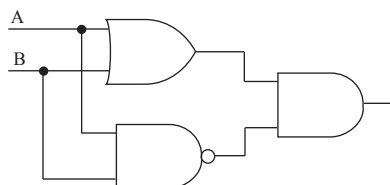


Figure 13-16

We convert this to a Boolean expression in much the same way that we produced the truth table by bit following. This time it is more like expression following. At the output of each gate we write an expression for the output in terms of its input. At the first stage this is easy and we get Figure 13-17.

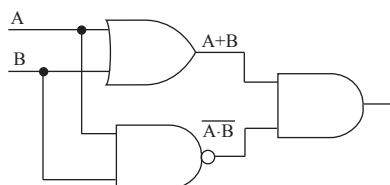


Figure 13-17

At the next stage we need to use brackets. One input to the final AND gate is the expressions $A+B$ so we put it in brackets. The other input is $\overline{A \cdot B}$ and we put it in brackets and use these as the arguments to the AND operator.

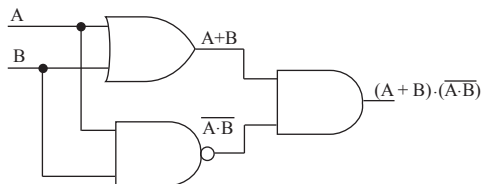


Figure 13-18

So the final expression is $(A + B) \cdot \overline{(A \cdot B)}$, which we know must be equal to $A \oplus B$.

Info For the mathematically inclined, Boolean algebra constitutes an Abelian field with OR playing the role of the addition operator and AND the role of the multiplication operator.

13.5.4 Algebraic Relations

We can obviously extend this to circuits that are as complicated as we want, though the resulting expressions may get very complicated. Boolean algebra allows us to manipulate expressions into simpler forms using a set of rules, some of which are similar to those for ordinary algebra.

Associativity

The binary operations of Boolean algebra obey an associative rule just like that of ordinary algebra.

$$\begin{aligned}A + (B + C) &= (A + B) + C \\A \oplus (B \oplus C) &= (A \oplus B) \oplus C \\A \cdot (B \cdot C) &= (A \cdot B) \cdot C\end{aligned}$$

This means that you can write strings of ANDs or ORs without parentheses because it doesn't matter what in order you do the operations. Thus we can write

$$A + (B + C) = A + B + C$$

and, as already mentioned, we can buy gates that have three or more inputs in addition to the more common 2-input ones.

Commutativity

The binary Boolean operators also obey a commutative law so that the order of terms in an expression does not matter.

$$\begin{aligned}A + B &= B + A \\A \oplus B &= B \oplus A \\A \cdot B &= B \cdot A\end{aligned}$$

Distributivity

Finally, the Boolean operations obey a distributive law in just the way that the arithmetic operations of multiplication and addition do. The only difference is that Boolean algebra has two different forms of addition so it has two distributive laws.

$$A \cdot (B + C) = (A \cdot B) + (A \cdot C)$$

and

$$A \cdot (B \oplus C) = (A \cdot B) \oplus (A \cdot C)$$

The Boolean operators obey a precedence rule in the same way that arithmetic operators do. AND has higher precedence than OR so that we can write an expression such as $(A \cdot B) + (A \cdot C)$ without parentheses as $A \cdot B + A \cdot C$ and the precedence tells us that we have to do the AND operations first and apply the OR to output of the AND operations. It is usually best not to rely too heavily on precedence. It is better to include a few unnecessary parentheses than to risk confusion.

De Morgan's theorems

Boolean algebra has two rules that have no counterpart in ordinary algebra. They are called De Morgan's theorems after their discoverer. These allow you to replace an AND term by an OR term and vice versa. Here are the two theorems.

$$A \cdot B = \overline{(\overline{A} + \overline{B})} \text{ which can also be written } \overline{A \cdot B} = \overline{A} + \overline{B}$$

and

$$A + B = \overline{(\overline{A} \cdot \overline{B})} \text{ which can also be written } \overline{A + B} = \overline{A} \cdot \overline{B}$$

These are a lot more powerful than the simple rules of algebra that we have previously seen. They are quite easy to prove by simply writing out the truth tables for the two sides and showing that they are the same. Since a Boolean expression is completely described by its truth table, *any two expressions that have the same truth table are equal to each other.*

Example

I shall prove the second De Morgan theorem. I like to use a lot of columns, one operation in each column. I start with columns for A and B and add columns for each of \bar{A} and \bar{B} , a column for $\bar{A} \cdot \bar{B}$ and, finally, a column for the final expression. That gives us 6 columns and so we have the large truth table shown as Table 13-13.

A	B	\bar{A}	\bar{B}	$\bar{A} \cdot \bar{B}$	$\bar{A} \cdot \bar{B}$
0	0	1	1	1	0
0	1	1	0	0	1
1	0	0	1	0	1
1	1	0	0	0	1

Since the final column of this expression truth table is the same as the final column for an OR gate, we have proved the theorem.

13.6 Logic simplification

Simplification is the process of taking a large expression and converting it into a smaller, but equivalent, one. The measure of what is a large and what a small expression is a little vague. Sometimes it means an expression that is shorter to write, but more often, it means an expression that takes fewer gates to construct. In particular cases, it may even mean an expression that uses more than the minimum number of gates but which makes the best use of the gates that are available. This is important when you realize that gates come several to a package. It is much better to use three gates rather than one if those three are already present as spares in other packages rather than adding another package to gain a single gate.

Simplification proceeds by applying the rules of Boolean algebra in some sequence. The trouble is there is no overall scheme to tell us what sequence to use. You just have to use trial and error. Try rules as they seem helpful and see if the expression you get is any better. If it is, keep going that way, if not go back and try again. Let's apply this method to our favorite expression

$$(A + B) \cdot \overline{(A \cdot B)}$$

Well, that NAND operation, $\overline{(A \cdot B)}$, is an obvious chance to apply a De Morgan theorem. De Morgan theorem 2 tells us that $\overline{(A \cdot B)} = \bar{A} + \bar{B}$ so we have

$$(A + B) \cdot (\bar{A} + \bar{B})$$

Next we can use the distributive rule to expand this just as we would expand a similar algebraic expression. That gives us

$$A \cdot \bar{A} + A \cdot \bar{B} + B \cdot \bar{A} + B \cdot \bar{B}$$

This is not looking particularly hopeful. We now have 4 terms instead of 2! However, some of those terms are peculiar because they only contain 1 variable instead of 2. These are candidates for simplification in much the same way that $(x) + (-x)$ is a candidate for simplification in arithmetic. Since both inputs to an AND operator have to be 1 for the output to be 1, the expression $A \cdot \bar{A}$ can never be 1. This means that the expression is exactly equivalent to 0. If we replace the occurrences of $A \cdot \bar{A}$ and $B \cdot \bar{B}$ by 0, we get

$$0 + A \cdot \bar{B} + B \cdot \bar{A} + 0$$

Now, we will appeal to standard arithmetic again. In arithmetic we know that $x + 0 = x$. Similarly, in Boolean algebra $A + 0 = A$. That reduces our expression to

$$A \cdot \bar{B} + B \cdot \bar{A}$$

This is a standard form of XOR that is easy to recognize and is well worth learning. It can often be used to simplify an expression. We can use this expression to go the other way and generate a new XOR circuit. It will take 2 NOT gates, two AND gates, and an OR gate, as in Figure 13-18.

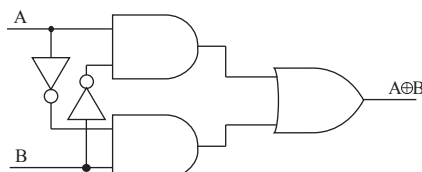


Figure 13-19

13.6.1 Some useful equivalents

Table 13-15 is a table of useful Boolean equations. Most of these are ones that we have already met. The rest are straightforward facts that are very useful for simplifying expressions.

Table 13-14: Boolean Facts

$A + B = B + A$	$(A \cdot B) \cdot C = A \cdot (B \cdot C)$
$A \oplus B = B \oplus A$	$A \cdot (B + C) = A \cdot B + A \cdot C$
$A \cdot B = B \cdot A$	$A \cdot (B \oplus C) = (A \cdot B) \oplus (A \cdot C)$
$(A + B) + C = A + (B + C)$	$(A \oplus B) \oplus C = A \oplus (B \oplus C)$
$A \cdot B = \overline{\overline{A \cdot B}}$	$\overline{A \cdot B} = \overline{A} + \overline{B}$
$A + B = \overline{\overline{A + B}}$	$\overline{A + B} = \overline{A} \cdot \overline{B}$
$A \cdot \overline{B} + \overline{A} \cdot B = A \oplus B$	$A \cdot B + \overline{A} \cdot \overline{B} = A \overline{\oplus} B$
$A + 0 = A$	$A + 1 = 1$
$A \oplus 0 = A$	$A \oplus 1 = \overline{A}$
$A \cdot 0 = 0$	$A \cdot 1 = A$
$A + A = A$	$A + \overline{A} = 1$
$A \oplus A = 0$	$A \oplus \overline{A} = 1$
$A \cdot A = A$	$A \cdot \overline{A} = 0$

13.6.2 Minimization

For logic systems with only a few inputs, up to 5 or 6, it is usually possible to find a good minimization by the trial-and-error application of the equations of Table 13-14. There are formal methods that are more reliable but they are too complicated to be worth using for smaller circuits. The two best known methods, Karnaugh mapping and the Quine-McClusky method, actually work directly on the truth tables rather than the logical expressions. There are computer programs available that implement these methods. It is well worth using such a program if you are designing a really complex circuit.

13.7 Logic design

We now have most of the tools that we need to pursue a design from a description of what the circuit should do to a complete circuit diagram. Indeed, so long as the truth table is not too complex or too random in appearance we can handle a lot of problems by inspection. For example, Table 13-15 is the truth table for the three-floor light-switch that we developed earlier.

A	B	C	O
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Although this is one of the first three input truth tables that we have looked at, some of the parts of it look very familiar. The first four rows look just like the truth table for B XOR C with an extra, useless, A input that is always zero. With a little more thought we can see that the lower four rows make up the truth table for B XNOR C, with an extra A input that is always one.

So, if A is zero then the output is B XOR C, if A is one then the output is B XNOR C. We can write that in logic form like this

$$\text{Out} = \overline{A} \cdot (B \oplus C) + A \cdot (\overline{B \oplus C})$$

If A is 0 then A is 1 and the first term becomes just C. But if A is 0 then the second term vanishes since $0 \oplus \text{anything} = 0$, so the expression works. Now we have to see patterns again. This time we have to recognize the pattern $X \cdot \overline{Y} + \overline{X} \cdot Y = X \oplus Y$ where $Y = A$ and $X = B \oplus C$ so that what we really have is

$$\text{Out} = A \oplus B \oplus C$$

and we can draw the circuit diagram very easily

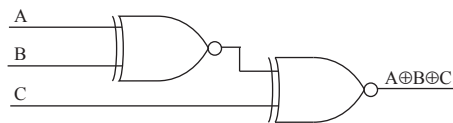


Figure 13-20

While relying on recognition works quite well for simpler circuits, for circuits with many inputs we need some more systematic tools. There are two methods that are guaranteed to convert any truth table into a valid logic equation, though they may not produce very efficient expressions. Once you have found an expression using one of these methods you can try to optimize it by hand or by computer. Here are the two methods.

13.7.1 Sum-of-Products

The sum-of-products method constructs an expression from a truth table one line at a time. In fact, one non-zero line at a time. The general form of the answer this method produces is

$$\text{Out} = \text{expr1} + \text{expr2} + \text{expr3} + \dots$$

where each expr is true for exactly one combination of input bits.

The expr's are chosen to put the 1 bits into the output column of the truth table and the or gate takes care of putting the 0's in. Each expr is an AND expression whose output is 1 only if all the inputs are 1.

To find the input expression we work along the row writing down the signal name everywhere there is a 1 and the complement of the signal everywhere there is a 0.

Example

Let's look again at our three floor light switch (Table 13-15).

The first row has a 0 in the output column so we don't do anything with it.

The second row has a 1 in the output column so we have to use this row. There are zeros in the A column and in the B column so these two inputs must be negated. The C entry is a 1 so it appears alone and the expression for this row is $\bar{A} \cdot \bar{B} \cdot C$

The third row also has a 1 in the output column and so contributes an expression to the answer. A is again negated but this time the 1 in the B column tells us that B is not. C now has a zero and so appears negated. The complete expression is $\bar{A} \cdot B \cdot \bar{C}$.

The fourth row has a 0 in the output and can be ignored.

The fifth row has a 1 and contributes $A \cdot \bar{B} \cdot \bar{C}$.

The sixth and seventh rows are both ignored.

The eighth row has a 1 and contributes $A \cdot B \cdot C$.

The complete expression that describes this truth table is then

$$\text{Out} = \bar{A} \cdot \bar{B} \cdot C + \bar{A} \cdot B \cdot \bar{C} + A \cdot \bar{B} \cdot \bar{C} + A \cdot B \cdot C$$

We can summarize the whole method with this algorithm—

Remember Sum-of-Products Algorithm

```

For each row in the truth table
  if the output is 0 do nothing
  if the output is 1 then
    for each input variable
      if there is a 1 in the current row, then write down
        the name of the variable
      if there is a 0 in the current row, then write down
        NOT the name of the variable
join all the signal names with AND's
join all the expressions with OR's.

```

The sum-of-products method is guaranteed to work but is most efficient for truth tables that have many 0's in the output and fewer 1's. If there are more 1's than 0's then the next method

is better. If there are equal numbers of 1's and 0's then use whichever one you prefer, they are of equal complexity.

13.7.2 Product-of-sums

Just as the sum-of-products makes use of the AND gate's property of having only 1 entry in its output column, so the product-of-sums makes use of the OR gate's property of having only one 0 entry.

Remember Product-of-Sums Algorithm

```

For each row in the table
if there is a 1 in the output column do nothing
if there is a 0 in the output column then work across the input terms
    if there is a 1 in the current row write down
        NOT the signal name
    if there is a 0 in the current row write down
        the signal name
connect all the terms with OR's
connect all the expressions with AND's
    
```

This method works down the truth table producing an expression for every row that has a 0 in the output column. Its final result looks like this—

$$\text{Out} = \text{expr1} \times \text{expr2} \times \text{expr3} \times \dots$$

Here is an example.

Example

We will work that same tired old example!

The first row is a zero so use it. The entry under A is a 0 so we write down A. The entries under both B and C are also 0's so we also write B and C. The whole expression becomes $A + B + C$.

The second and third rows are both 1's and are ignored.

The fourth row is a 0 and produces $A + \bar{B} + \bar{C}$.

The fifth row is 1 and is ignored.

The sixth row is a 0 and produces $\bar{A} + B + \bar{C}$.

The seventh row is a 0 and produces $\bar{A} + \bar{B} + C$.

The eighth row is a 1 and is ignored.

The complete expression is $\text{Out} = (A + B + C) \times (A + \bar{B} + \bar{C}) \times (\bar{A} + B + \bar{C}) \times (\bar{A} + \bar{B} + C)$

where I have used lots of parentheses to make sure that we get the order of the operations right because AND has a higher priority than OR.

Note This expression uses four 3-input OR gates and one 4-input AND gate. The previous method needed four 3-input AND gates and one 4-input OR gate so the two expressions are equally complex.

13.7.3 Multi-output circuits

So far, all the circuits that we have looked at have had a single output. It is perfectly possible for a circuit to produce two outputs from the same set of inputs. In fact, it is probably more common to have a multi-output circuit than a single output one. The good thing is that we already know all the techniques for dealing with multi-output circuits because we can treat a multi-output circuit as several different single-output circuits. When we analyze the circuit, we do so one output at a time and when we try to find a circuit to implement a given truth table we do so one output column at a time. Once we have analyzed the whole circuit as a set of separate one-output circuits we go back over the circuit to see if we can save some gates by sharing them between different parts of the circuit. We can use such sharing if there are common sub-expressions in the equations for different outputs.

An example of the kind of expression sharing that is possible would be a system that we shall meet again when we look at arithmetic circuits. The circuit has two inputs, A and B, and two outputs, X and Y. Here are the expressions that describe the circuit.

$$X = A \oplus B \quad \text{and} \quad Y = A \times B$$

If we are building this out of AND, OR, and NOT gates then we can use the XOR circuit we already found to write

$$X = (A + B) \times \overline{(A \times B)}$$

and then we see that there is a common term of $A \times B$ in the two expressions. This means we can save one AND gate and we get the circuit of Figure 13-21.

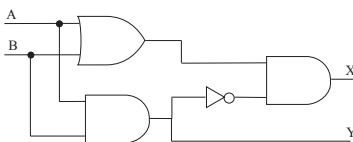


Figure 13-21 XOR Gate

13.8 Putting it all together

Let's look at a complete example, going from an idea or a description of what the final circuit should do to a full logic circuit. The circuit that I have picked is taken from a popular electronic toy that is often sold as a kit for people wanting to build their first logic circuit. Obviously they get the circuit all designed, for them but we are going to do it all ourselves. The circuit is part of a digital die kit (often incorrectly called a digital dice kit but "dice" is plural and we are only designing one). A part of the kit that we won't study until later is a circuit that has three outputs, which go through a counting sequence from 0-5, thus

000, 001, 010, 011, 100, 101, 000, 001,...

Another part of the kit has seven LEDs arranged as shown in Figure 13-22. Note that I added to the LEDs to make it easier to describe the output. The labels would not be part of a real device.

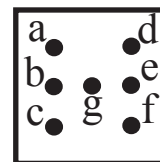


Figure 13-22 Die Face

Obviously, we want the LEDs to light up in the right sequence to show the values 1, 2, 3, 4, 5, 6. There are various ways to light LEDs from logic circuits but the rule we shall use is that an input of 0 to an LED makes it light up. We can construct the truth table our usual way, one line at a time.

Input 000 should light up only LED g so that we see a single central pip. The circuit needs to produce outputs a = 1, b = 1, c = 1, d = 1, e = 1, f = 1, g = 0. Let's write that output 111110 in alphabetic order to keep things short.

Input 001 needs to light up LEDs to show a 2. A quick glance at a conventional die shows me that we want to turn on LEDs c and d so the output is 1100111.

Input 010 needs to light up LEDs to show 3 so we will turn on c, g, and d with output 1100110.

Input 011 will give us a 4 by lighting a, c, d, and f using output 0100101.

Input 100 will give us a 5 by lighting a, c, d, f, and g with output 0100100.

Lastly, input 101 will give us a 6 by lighting a, b, c, d, e, and f with output 0000001.

We end up with the truth table of Table 13-16.

Table 13-16:

I2	I1	I0	a	b	c	d	e	f	g
0	0	0	1	1	1	1	1	1	0
0	0	1	1	1	0	0	1	1	1
0	1	0	1	1	0	0	1	1	0
0	1	1	0	1	0	0	1	0	1
1	0	0	0	1	0	0	1	0	0
1	0	1	0	0	0	0	0	0	1

The next stage is to work through the columns writing logic equations for each output.

Column a.

There are equal numbers of 1's and 0's in this column so we can use whichever method we want. Let's use sum-of-products. We need to get one outputs for 000, 001, and 010 states so the logic equation is

$$a = (I2 \cdot I1 \cdot I0) + (I2 \cdot I1 \cdot I0) + (I2 \cdot I1 \cdot I0)$$

Column b

This has only one zero and 5 ones so we will use product-of-sums to give

$$b = I2 + I1 + I0$$

Column c

This time we only have a single one so we will use sum-of-products again

$$c = I2 \cdot I1 \cdot I0$$

Note In this case we don't include all 8 possible input states since we are guaranteed by the conditions of the circuit that we will never encounter the remaining two input combinations. A die can never show the number 8 and the counting circuit should be designed so that it will never put its outputs into the states I10 or I11.

Column d

Cute, column d is just the same as column c so we can just connect them together and save one whole output!

$$d = c$$

Column e

Again, we have a repeater. Column e is just like column b so we can use

$$e = b.$$

Column f

Column f is the same as column a.

$$f = a$$

Column g

This time we have another unique column and need a new expression. Again it doesn't matter which method we use. Sum-of-products gives us

$$g = (I_2 \cdot I_1 \cdot I_0) + (I_2 \cdot I_1 \cdot \bar{I}_0) + (I_2 \cdot \bar{I}_1 \cdot I_0)$$

Thus we have four unique circuits. The circuits for a and g share the common term $(I_2 \cdot I_1 \cdot I_0)$ and those for a and c share the term $(I_2 \cdot I_1 \cdot \bar{I}_0)$. Figure 13-23 below shows the complete circuit, realized by writing the logic diagrams for each expression.

Note I have laid the circuit diagram in Figure 13-22 out in a special way to make it easier to follow, with the inputs coming in vertically and the outputs taken horizontally. This is only done for appearance sake and is in no sense a part of the diagram. However, it is really good way to structure complicated logic diagrams and I recommend that you get into the habit of drawing them in this general fashion, with the inputs coming in perpendicular to the outputs. I also often use the inputs on the right, outputs at the bottom organization.

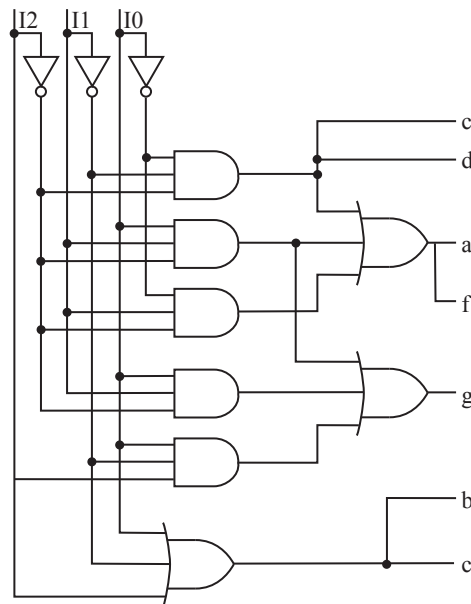


Figure 13-23 Logic Circuit for Digital Die

Summary

Combinatorial logic circuits form logical combinations of a set of input values to produce a set of output values that depend only on the current values of the inputs. They have no memory.

We describe combinatorial circuits by means of truth tables giving the output states in terms of the input states.

We can also describe combinatorial circuits by means of Boolean equations. These are equations that follow the laws of Boolean Algebra instead of the more familiar laws of arithmetic.

A single Boolean equation can describe one output of a combinatorial circuit and so a multi-output circuit may need several equations. The individual operators of Boolean Algebra correspond to the individual gates of a logic circuit and so the Boolean description is a particularly convenient one for designing a circuit.

Table 13-17: Boolean Facts

$A + B = B + A$	$(A \cdot B) \cdot C = A \cdot (B \cdot C)$
$A \oplus B = B \oplus A$	$A \cdot (B + C) = A \cdot B + A \cdot C$
$A \cdot B = B \cdot A$	$A \cdot (B \oplus C) = (A \cdot B) \oplus (A \cdot C)$
$(A + B) + C = A + (B + C)$	$(A \oplus B) \oplus C = A \oplus (B \oplus C)$
$A \cdot B = \overline{(\overline{A + B})}$	$\overline{A \cdot B} = \overline{A} + \overline{B}$
$A + B = \overline{(\overline{A \cdot B})}$	$\overline{A + B} = \overline{A} \cdot \overline{B}$
$A \cdot \overline{B} + \overline{A} \cdot B = A \oplus B$	$A \cdot B + \overline{A} \cdot \overline{B} = A \oplus B$
$A + 0 = A$	$A + 1 = 1$
$A \oplus 0 = A$	$A \oplus 1 = \overline{A}$
$A \cdot 0 = 0$	$A \cdot 1 = A$
$A + A = A$	$A + \overline{A} = 1$
$A \oplus A = 0$	$A \oplus \overline{A} = 1$
$A \cdot A = A$	$A \cdot \overline{A} = 0$

We can convert a truth table into a set of Boolean equations using the formal logical methods of Sum-of-Products and Product-of-Sums.

Sum-of-Products Algorithm

For each row in the truth table
 if the output is 0 do nothing
 if the output is 1 then for each input variable
 if there is a 1 in the current row, write down the name of the variable
 if there is a 0 in the current row, write down NOT the name of the variable
 join all the signal names with AND's
 join all the expressions with OR's.

This method is most useful if the output section of the truth table contains more 0's than 1's.

Product-of-Sums Algorithm

For each row in the table
 if the output is a 1 do nothing
 if the output is a 0 then for each input term
 if there is a 1 in the current row write down NOT the signal name
 if there is a 0 in the current row write down the signal name
 connect all the terms with OR's
 connect all the expressions with AND's

This method is most useful if there are more 1's than 0's.

Exercises

- Design a circuit to implement the logical function described by the truth table on the left. Suggest what this circuit might be used for.
- Design a circuit using only NAND and NOT gates to implement the logical function described by this truth table.

Table 13-19: 19

A	B	O3	O2	O1	O0
0	0	0	0	0	1
0	1	0	0	1	0
1	0	0	1	0	0
1	1	1	0	0	0

Suggest a name for this circuit.

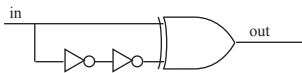
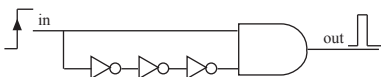
Table 13-18: 18

A	B	O1	O0
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

- There are several letters that can be displayed on a 7-segment LED display. In particular, the Hex digit letters A, b, C, d, E, and F (note the case, it is a hint) are all easy to handle. Remembering that a logic 0 is needed to light an LED segment write a truth table for a circuit to convert a 4-bit hex input between \$A and \$F into the seven logic signals needed

to light the LEDs corresponding to the value. Thus input \$A should display 'A', \$B display 'b', and so on.

4. Convert the truth table of problem 3 into a set of logic equations and draw the circuit diagram for such a device.
5. The central door locking system of my old Toyota exhibited the following behavior when you unlocked the driver's side door. The first turn of the key unlocked only the driver's door; the second turn unlocked all the rest of the doors. Design a logic circuit that takes two inputs, one from the key switch and the other from a switch in the door lock of the driver's door. The key switch produces a 1 when the key is turned and the door switch produces a 1 so long as the door is locked and a 0 so long as the door is unlocked. The logic circuit should produce two outputs, one that unlocks the driver's door when it is true and one that unlocks all the other doors.
6. A further complication to the locking system of my Toyota was that unlocking the passenger side door unlocks both the driver's door and all the other doors at once. Assuming that the lock in the passenger side door is similar to that in the driver's door, add a third input to your truth table to implement this behavior.
7. There are many ways to construct an exclusive OR gate from simpler gates. Use your knowledge of Boolean algebra to find the simplest such circuit that you can. (The simplest method I know of uses four NAND gates and no other gates. You may define simplest your own way.)
8. Write out a large truth table (like Table 13-14) to prove the first De Morgan theorem, $A \times B = \overline{\overline{A} + \overline{B}}$.
9. Remembering that it takes a short time (5-10nS) for a signal to propagate through a gate (that is, for the output of the gate to alter after the input has changed), explain how the following circuit produces a short output pulse every time the input goes from a 0 to a 1.
10. Explain why the circuit in problem 9 does NOT produce an output pulse when the input goes from 1 to 0.
11. Somewhat like the circuit of problem 9, the circuit below produces an output pulse from an input transition or edge. Under what conditions do you get a pulse from this circuit.



Chapter 14:Combinatorial Functions—Coders, Decoders and Arithmetic

14.1 Medium Scale Logic

In the last two chapters we have met some of the simplest logic building blocks and seen their implementation as SSI integrated circuits. Because a number of more complex logic functions are commonly found in many types of circuit, manufacturers found it profitable to produce integrated circuit versions of these common functions. This chapter describes some of the more complex functions available as Medium-Scale Integration (MSI) logic, devices with up to a few dozen individual gates inside them.

We shall concentrate on three classes of MSI circuit in this chapter. First there are coders/decoders, that translate one kind of binary code into another, or multiplexers/demultiplexers that allow more than one signal to share the same wire. We will look at the functions available, see what they do, and see how they are often used.

14.2 Coders/decoders

A coder or decoder translates one binary code into another. The terms are pretty much interchangeable but we typically use the term coder for a circuit that has more inputs than outputs and the term decoder for a circuit that has more outputs than inputs.

So far we have met only two binary codes, the standard binary counting sequence and the gray code used for positional encoder devices. Both of these have the same number of bits—a 3-bit gray code can be translated 1:1 into a 3-bit arithmetic code. Most other codes are much less tightly packed. The commonest is a one-of-N code where exactly one of a set of bits is 1 or exactly one bit is 0.

You would use a one-of-N code if you had an output device made of a set of lights, each with a number written on it, and then lit one light at a time to tell you which state the system was in.

Example

A 1-of-N decoder is ideal for controlling the floor display of an elevator. Unless you are in a very old elevator with the indicator dial over the door and a hand that moves from number to number, the floor display consists of a panel of numbered lights. Since the elevator is only on 1 floor at a time only one of the lights is lit at any moment. Internally, the elevator controller uses a standard binary code to keep track of which floor the elevator is on and then uses a 1-of-N decoder to produce a set of wires, only one of which is turned on at a time. These wires then control the lights in the floor display.

‘139 Dual 2-line to 4-line decoder

A 2-line to 4-line decoder converts between a 2-bit input and a one-of-4 line output where only one of the outputs is 0 at a time. There are 3 input lines. The first, G, is an enable line that can shut down the whole device. The other 2 inputs, A and B, form a 2 bit number and select which of the 4 output bits will be made active, that is, set to 0. It is quite common to find extra enable inputs like this. They make it easy to **expand** the device, that is to connect several devices together to form a larger circuit. In particular, we can use the G inputs and an inverter to make a single 3-line to 8-line decoder from the two 2-4 line decoders in a ‘139 (see below).

Table 14-1 (next column) shows the truth table for one of the decoders in this device (they are identical).

Note The Enable input should be called $\overline{\text{Enable}}$ since it is a negative logic signal. That is, it is active when it is *low*.

Table 14-1: 74HC139

Inputs			Outputs			
Enable	Select		Y0	Y1	Y2	Y3
G	B	A				
H	X	X	H	H	H	H
L	L	L	L	H	H	H
L	L	H	H	L	H	H
L	H	L	H	H	L	H
L	H	H	H	H	H	L

As is commonly the case, the truth table is specified in terms of high (H, 5V) and low (L, 0V) levels rather than 1's and 0's. As mentioned in Chapter 13 this avoids any uncertainty over the meaning of 0 and 1.

In addition to the usual H's and L's there are also some X's for inputs. These mark states where the value of that input is irrelevant. In this case they mean that all of the outputs will be forced high regardless of the settings of A and B so long as G is high. The device only functions as a decoder when G is low.

The actual circuit (Figure 14-1 below) is a little more complex than you might at first imagine. The logic function only requires the presence of the circuitry up to and including the NAND gates. The NOT gates which follow the NAND's do not alter the logical function of the device. Instead, they are used as **output buffers** to provide more power to drive the output. The FET's used inside most of the logic gates are very small, fast, low power devices. They are not suited to driving external circuitry. The FET's in the buffer NOT gates are rather larger devices that can drive significant external loads.

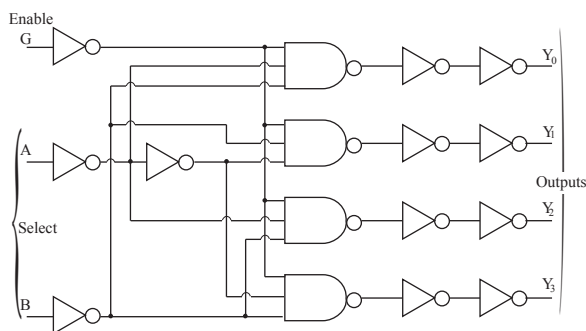


Figure 14-1 The 74HC139 Dual 2-4 Line Decoder

We are rarely interested in the internal logic of an integrated circuit. Instead we want to know the truth table, in order to understand its function, and the pin-out diagram that tells us what each pin does. We need this if we are to connect the device into a circuit. Here is the pin-out diagram for the '139 decoder.

This makes it clear that there are two separate 2-4 line decoders in the single package of the '139. It uses the same notation as the truth table so that we can relate the two and it shows the numbers of the pins to which the signals are connected. From this figure we can tell that if we want to use outputs 1Y0-1Y3 then we need to use inputs G1, A1, and B1. Similarly, if we want to use outputs 2Y0-2Y3 then we need inputs G2, A2, and B2.

Let us first see an example of how to use a decoder and then look at how to use the enable input to expand the decoder.

Note Logically we could have used AND gates followed by the NOT buffers but there is no direct way to build a CMOS AND gate. Instead we build a NAND gate and follow it by an inverter as shown here. It is only the final NOT gates that are built as buffers.

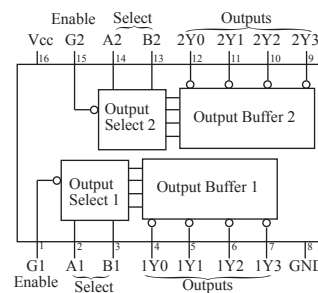


Figure 14-2 74HC139 Dual 2-4 line Decoder Pinout

Example

Memory select generator

Each location in the memory of a computer has a unique binary address. When the computer wants to read or write a single location, it puts the address on a set of wires called the address bus. Real computers have very large numbers of memory locations. For example, a microcomputer such as the MC9S08, which we shall study later, can address 65536 different memory locations. At most one of these locations can be accessed at any time (sometimes the computer is not using any memory locations). That means that we have to select one out of 65,536 different devices. Now, you can't have 65,536 wires going from the computer to memory; the package you would need would be enormous! Instead, the computer encodes the number of the desired memory location as a 16 bit binary number and the memory circuitry has to take care of decoding that number into a one-of-65,536 signal. Fortunately, memory circuits come with thousands of individual cells in a single package and the addressing of the individual cells is taken care of inside the package but we then need 2-to-4 or similar decoders to select which memory chip a particular address is in.

Several common types of memory come in 8k packages, that is 8192 bytes of memory in one package. If we have 4 such memory chips, then we can use these to provide the top 32k of memory for the computer. The only other thing that we need is a way to select which chip contains a particular address. We can use a '139 decoder to select which one is turned on for a given address.

Each memory chip has a Chip Enable (CE) input that has to be LOW for the chip to be turned on. When the memory chip is enabled, it will respond to addresses and commands to read and write memory locations. When it is disabled, the chip ignores these signals and draws very much less power. Each chip has 13 address wires, $2^{13} = 8192$, so that leaves three wires out of the 16 coming from the computer to select which chip is enabled for which address. The normal practice is to split the memory up into blocks of addresses and assign whole blocks to each chip as shown in Figure 14-3.

According to the map, chip 1 must respond to address from \$0000 to \$1FFF, chip 2 to addresses from \$2000 to \$3FFF, etc.

We can implement this memory scheme with a one half of a 74HC139 connected as in Figure 14-4. Let's look at what happens when the computer outputs an address, for example \$3A1B. The top three bits of the address, 001, go to the '139. The top bit is connected to the G input so that the '139 is enabled ($G = L$ in the truth table of 74HC139). The next two bits are the B and A bits respectively. We see from the truth table that, with $B = 0, A = 1$, the output Y1 will be L and all the others will be H. That means that chip 2, connected to Y1, will be enabled and all the others will be disabled. Thus chip 2 looks at the other 13 address bits and knows to access its bit \$1A1B. You can work through some other addresses to see what happens to them and to see how each address is directed to the correct chip.

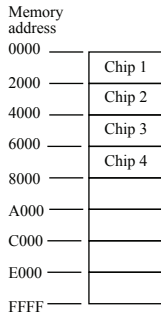


Figure 14-3 Memory Map

Note I have only put memory in half of the possible places. This is quite normal; very few computers have their whole possible memory space filled with real memory. If you write to the places that aren't there then the data are lost forever. If you read from those places, you will get some random value corresponding to the input lines being disconnected. This causes no problems because our programs tell the computer which memory locations to use and we know where the memory is. It is up to us, the programmers, to write programs that only use the memory locations that are actually populated.

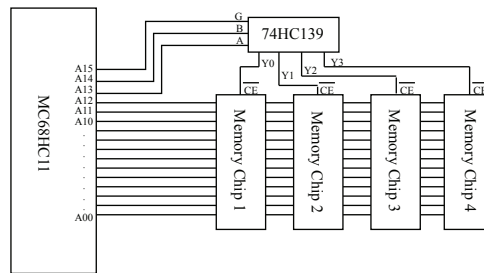


Figure 14-4 Memory Decoder

Expanding the 2-to-4 line decoder

With the aid of a single external NOT gate, we can combine the two separate 2-to-4 line decoders in one '139 to make a 3-to-8 line decoder. We use the third bit to select one of the two decoders as in Figure 14-5.

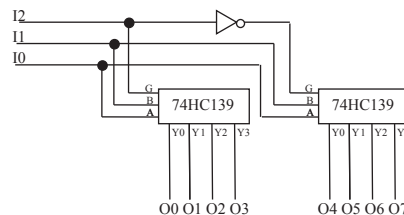


Figure 14-5 3-8 Line Decoder

You can verify for yourself that this circuit has the truth table of Table 14-2.

I2	I1	I0	O7	O6	O5	O4	O3	O2	O1	O0
0	0	0	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	1	0	1
0	1	0	1	1	1	1	1	0	1	1
0	1	1	1	1	1	1	0	1	1	1
1	0	0	1	1	1	0	1	1	1	1
1	0	1	1	1	0	1	1	1	1	1
1	1	0	1	0	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1	1

Table 14-2: 3-to-8 line truth table

‘137, ‘138 3-to-8 line decoder

These are very similar parts. Both implement the function of Table 14-2 with two extra enable lines for expanding the chip. The difference is that the ‘137 has a little memory added so that the inputs can change and the chip can go on remembering the saved value until it is told to store a new address value.

‘154 4-to-16 line decoder

This is pretty much the last word in decoders. It has 4 address inputs and a pair of enable inputs for expansion, should you be wild enough to want more than 16 select lines!

‘47 BCD-to-7 segment decoder/driver

This chip has a 4-bit BCD input but this time the output is designed to drive one of the seven segment displays that are the hallmark of our digital times, from wrist watches and calculators to the innumerable blankly flashing VCR clocks of the world. The ‘47 has extra driver transistors on the outputs to drive LED segments without any extra components. This chip produces a sensible output only for the 10 input states that represent the numbers from 0-9. It is possible to make readable approximations for the other 6 possible states of the input (the hex digits A, B, C, D, E, and F) but the extra logic to do this was not designed into these chips and they produce nonsense outputs for those states.

Note This function does not seem to be made in the 74HC family. A strong indication that multi-digit driver chips have won. One example is the 74C912, which can decode and drive 6 digits worth of 7-segment displays. It has a companion part with the extra decoder gates to display all 16 hex digits correctly on 7-segment display.

14.3 Encoders

There are not very many examples of encoders because there aren’t many many-wire codes to turn into few-wire codes. The best known is called a priority encoder. It takes a set of individual input bits each of which has a value, or priority, assigned to it and it outputs a binary number corresponding to the signal with the highest priority.

Let’s look at the simplest non-trivial example, the 4-line to 2-line encoder. As a 4-input circuit it must have 16 distinct input states (Table 14-3) but with only 2 output bits it can have only four output states. Thus an output of 10 (binary representation of 2) means that input 2 is high but input 3 is not. The highest priority active input is input 2.

NOTE that the output of 00 is ambiguous. It could mean that input I0 is active and no other or that no input is active. This is not usually a problem. These devices are normally used in situations where the output is only meaningful when at least one of the inputs is active.

At first sight it looks as though this should require two very large equations with many terms, even using the product of sums method. However, if we rewrite the table using ‘x’ entries to indicate “don’t care” states then we see the the problem is much smaller. With only 4 output states we really only need four lines in the truth table (Table 14-4).

Since the x entries mean that we don’t care what values those inputs take, we completely ignore them when building our equations. We only consider those inputs that have specified values.

I3	I2	I1	I0	O1	O0
0	0	0	0	0	0
0	0	0	1	0	0
0	0	1	0	0	1
0	0	1	1	0	1
0	1	0	0	1	0
0	1	0	1	1	0
0	1	1	0	1	0
0	1	1	1	1	0
1	0	0	0	1	1
1	0	0	1	1	1
1	0	1	0	1	1
1	0	1	1	1	1
1	1	0	0	1	1
1	1	0	1	1	1
1	1	1	0	1	1
1	1	1	1	1	1

Table 14-3: 4-line to 2-line priority encoder

I3	I2	I1	I0	O1	O0
0	0	0	1	0	0
0	0	1	x	0	1
0	1	x	x	1	0
1	x	x	x	1	1

Table 14-4: 4-2 priority encoder simplified

Thus the equations become

$$O0 = \bar{I3} \cdot \bar{I2} \cdot \bar{I1} \cdot I0 + I3$$

$$O1 = \bar{I3} \cdot I2 + I3$$

which are very simple.

The principle extends to larger numbers of inputs. The standard 74 series logic includes the ‘148, which is an 8-line to 3-line priority encoder, and the ‘147, which encodes a one-of-ten input onto four output lines with some missing codes. There does not seem to be a full 16-line to 4-line encoder, presumably because it would need such a large number of pins. However, the ‘148 has an extra input and two extra outputs that allow you to cascade two ‘148 chips to get a full 16-line to 4-line priority encoder in much the same way that you can cascade the sections of a ‘139 to make a full 3-line to 8-line decoder.

One use for priority encoders is found in the interrupt systems of many computers. An interrupt is a special kind of input that tells that computer that one of its peripheral devices needs attention. Complicated computers may have many peripherals and it may happen that more than one needs attention at one time. The computer can only attend to one at once. A priority encoder can be used to tell the computer which is the most important interrupt to service at any time.

Another use for priority encoders is in analog to digital converters. We will return to this subject in Chapter 29 but we can look at a simple example now. An analog-to-digital converter takes an analog voltage, a voltage that can take any value, and outputs a binary number that is an approximation to the value of the voltage. For example, such a converter might output the number 2 for all voltages between 2V and 3V. Consider the circuit below.

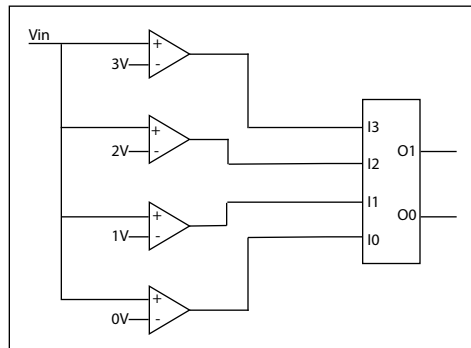


Figure 14-6 2-Bit Analog-to-Digital Converter

This takes a voltage on the line marked *V_{in}* and outputs a 2-bit binary number approximation to the voltage. The funny triangular things with two inputs are called comparators. They compare the voltages at their two inputs and output a 1 if the voltage at the positive input is greater than the voltage at the negative input, otherwise they output zero. For example, if we set *V_{in}* to 2.43 volts then the bottom comparator will output a 1 because 2.43 > 0. Similarly, the next two comparators will output 1 because 2.43 > 1 and 2.43 > 2 but the third will output a 0 since 2.43 < 3. We have seen above that if we put input 0111 into our 4-line to 2-line priority encoder then its output will be O1=1 and O0=0, the binary code for the number 2. You can try some other voltages and see that this circuit converts any voltage between 0V and 4V into a one-digit (2-bit) approximation to the voltage!

14.4 Multiplexers/demultiplexers.

A multiplexer allows several signals to share a single wire. We shall later see that computers are full of places where a single wire is shared by many signals, each getting to use the wire at its own time. Some of those places use multiplexers and some use a different technique based on the use of tri-state logic. A multiplexer is basically a digital version of the n-way switch, it can connect one of its inputs to the output and the use selects which output is connected at any time. Multiplexers are available in a variety of sizes, from 2-line to 1-line multiplexers that

Note A digital multiplexer is strictly a one-way device, like all digital circuits. The signal travels only from input to output; no information travels in the reverse direction.

come 4 to a package up to 8-line to 1-line multiplexers that use up all the pins of a package by themselves. A multiplexer has two sets of inputs, a set of n address inputs and a set of 2^n data inputs. The address inputs make up a binary number that selects which of the 2^n data inputs is connected to the output. Again, we will look first at an example of the devices available and then see how they might be used.

‘153 Dual 4-input multiplexer

This chip contains two sections, each of which can connect one of its 4 inputs to the single output. The address inputs, A & B, are common to the two sections so that the simplest use is to select one of four 2-bit input words and connect it to a 2-bit output word. Obviously, you can use several of these, side-by-side, to select larger word lengths. For example, with 4 of these chips we can connect one of four 8-bit input bytes to a single 8-bit output byte. Figure 14-7 shows the pin-out diagram of the ‘153 and Table 14-4 its truth table.

This truth table is the most complex that we have seen so far. There are three different kinds of input. First there are the two common address lines, B and A, that select which of the data inputs is connected to the output. Then there are 4 input lines for each half, one of which is connected to the output at any time. Finally, there is a strobe input, G, for each half that can enable or disable its half of the device. In simple uses the G line is tied high to allow the device to function but it can also be used with an external NOT gate to allow the device to function as a single 8-input multiplexer by tying the two outputs together with an AND gate as in below. Here we use the G input as an extra bit exactly as we did back in Figure 14-5.

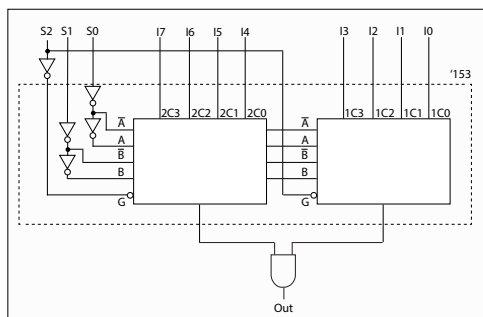


Figure 14-8 Cascaded 8-input multiplexer

The result is a device that uses a 3-bit input value to connect exactly one of the inputs to the output. So, for example, if we set the input to 101, the number 5, then the output will follow the state of input 5 and ignore the other 7 inputs.

This is an example of a device that is sufficiently complicated that it is very difficult to perceive the pattern in its operation from the truth table. It is easier to describe what it does in words and then to look at the truth table and see that it makes sense in terms of that description.

Example

The multiplexed display

This is a very simple example of a powerful technique that we shall meet again when we are studying computers. One of the problems with displaying large amounts of information is that each individual display element has to have its own wire to tell it whether to be on or off. If we are not careful, the wiring for even a moderate sized display can cost more than the display itself. For example, the kind of scrolling display often seen in supermarkets or airports uses hundreds of LED's to display letters; a 20 letter display may have 1600 LEDs arranged as 20 letters each of which is an 8 by 10 array of LEDs. If every LED in the display had to be connected directly to the computer that controlled it, it would need a cable as thick as your leg to carry all the signals. That is impractical. Instead, such displays rely on a phenomenon called the *persistence of vision* to reduce the number of wires enormously. Because of the persistence of vision, the eye perceives a light source as steady even if it is flashing on and off, so long as it flashes more than 20-30 times per second. Thus, we only need to send enough information from the computer to the display to light up one column of LEDs at a time if we send the data fast enough that each column gets lit up (refreshed) more than 30 times per second. This is called a **multiplexed** display. Using this technique, we only need 18 wires in the cable for our 1600 LED display.

To understand the principle, let us look at a rather simpler example. Instead of having 160 columns of 10 LEDs

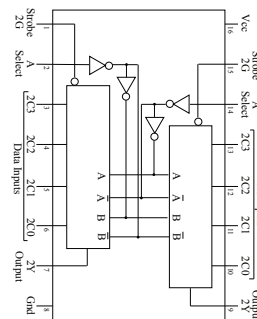


Figure 14-7 The 74HC153 Dual 4-input Multiplexer

Table 14-5: ‘153 4-input multiplexer

Inputs						Out put	
Select		Data					
B	A	C0	C1	C2	C3	G	Y
X	X	X	X	X	X	H	L
L	L	L	X	X	X	L	L
L	L	H	X	X	X	L	H
L	H	X	L	X	X	L	L
L	H	X	H	X	X	L	H
H	L	X	X	L	X	L	L
H	L	X	X	H	X	L	H
H	H	X	X	X	L	L	L
H	H	X	X	X	H	L	H

Note This time there are even more don't care (X) entries in the truth table. For example, when the strobe is high, the output is low regardless off the state of the other 6 inputs. If we did not use the X notation it would take $2^7 = 128$ rows to write out the truth table!

we will have only 8 columns with 1 LED in each. This will allow us to display 8 bits of information but, instead of using 8 lines to carry the information, we will use a multiplexer to compress the information onto 4 lines. The circuit is given in Figure 14-7 below.

The address generator sits and outputs the numbers 0 through 7 one thousand times per second. When the address lines are all zero then the multiplexer makes its output equal to the value of I0. At this time, the 3-to-8 line decoder, which is the active high kind, sets Y0 to 1 and all the other outputs to 0. That means that the FET for D0 is turned on and all the others are turned off. Thus, D0 follows I0, if I0 = 1 then D0 is on, if I0 = 0 then D0 is off. All the other diodes are turned off because their switches are turned off. One millisecond later the address lines are 001 and the multiplexer connects I1 to all the diodes. Now the decoder turns on the transistor for D1 and turns off all the others so D1 = I1. As time passes, each diode is connected to its input in turn while all the others are off. because of the persistence of vision effect, the eye sees the diodes following the inputs and it doesn't notice that each diode is only on one eighth of the time.

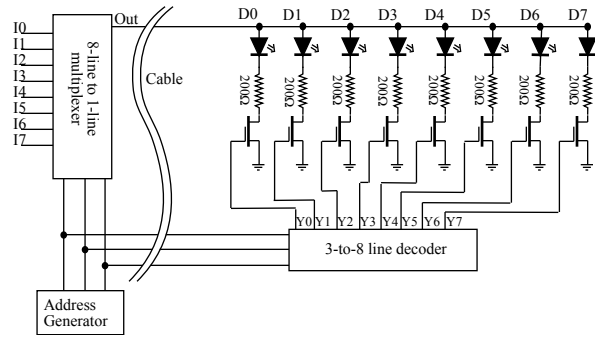


Figure 14-9 Multiplexed Display

14.4.1 Demultiplexers

The inverse of the multiplexer is the **demultiplexer**. It has one input and many outputs, only one of which is connected to the input at a time. The 3-to-8 line decoder and the switch FETs in the previous example form a simple 1-to-8 line demultiplexer and show the normal use of the demultiplexer, restoring a set of lines that have been previously multiplexed. Most demultiplexers are built from other components, just like the one in the example. The 74HCxx family contains one example of a demultiplexer.

‘155 Dual 2-4 line decoder/demultiplexer

A decoder and a demultiplexer have so much in common that a single device can serve either function as you can see in this extract from the data sheet for the 74HC155. A single wire can function as both the strobe for the decoder and the data input for the demultiplexer. When that line is high, all the outputs are high so the decoder is disabled. When that line is low, one selected output will be low so the decoder is enabled. For a given input address, the corresponding line will follow the data/strobe input!

14.5 Arithmetic Logic

The last class of circuits performs arithmetic and logic functions on whole sets of bits at once. For example, a single chip may add together two 4-bit binary numbers to form a 5-bit output. The most powerful of these chips are designed to form the hearts of computers. They have not only two 4-bit inputs and a 5-bit output (four bits plus a carry/borrow), but also a set of control inputs. These control inputs allow a single chip to add two numbers, subtract them, or perform various logic operations. We shall work our way to such complex chips starting from the simple arithmetic task of adding two one bit numbers.

We have already met the circuit that adds two single bits to give a 2-bit result (Figure 14-8 and Table 14-4). It is called a Half-Adder.

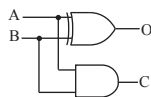


Figure 14-10 Half Adder

B	A	C	O
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

Table 14-6: Half-adder truth table

14.5.1 The Full-adder

Our goal is a circuit to add two multi-bit numbers. Let's look at the process of adding two 2-bit numbers and we shall see what tools are needed. For example, here is the process of adding 3 + 1.

$$\begin{array}{r} 11 \\ + 01 \\ \hline 101 \end{array} \quad \begin{array}{r} 11 \\ + 01 \\ \hline 1 \end{array} \quad \begin{array}{r} 11 \\ + 01 \\ \hline 01 \end{array} \quad \begin{array}{r} 11 \\ + 01 \\ \hline 101 \end{array}$$

We see two things. First, the result of adding two 2-bit numbers is a 3-bit number, not a 2-bit number. Second, because we have to deal with the "carries" from one column to the next, the fundamental operation that takes place in each column is not the addition of two bits but of three, two inputs and a carry from the previous column. That means that our basic Full-Adder circuit needs to have the truth table of Table 14-5.

We recognize that last column from the light switch problem.

$$O = I2 \oplus I1 \oplus I0.$$

The carry output, C, is a bit more complex. If we apply the sum-of-products method and simplify the resulting expression we get

$$C = I1 \cdot I0 + I2 \cdot (I1 \oplus I0).$$

Now we see that $I1 \oplus I0$ is a common sub-expression so we won't need to make that twice. It is also the main part of a half adder and we usually implement the full adder with two half adders as in Figure 14-9.

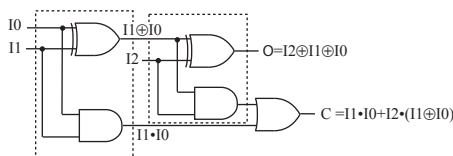


Figure 14-11 The Full-adder

Obviously, if we want to use that in other circuits, we need to have a shorthand symbol for it. Figure 14-12 shows a common one. Note that A and B are the number inputs, Co is the carry out to the next stage, and Ci is the carry in from the previous stage.

We can now design circuits to add numbers of any width, using a single full adder per bit.

Example

A computer represents numbers internally as patterns of binary digits (bits). The structure of binary codes is explained in detail in the Introduction to Computers book. Some of the simplest computers available use only 4 bits to represent a number and can thus count up to only $2^4 = 16$. Although they are limited in the range of numbers that they can represent they must be able to perform all the usual operations on those numbers and so need to be able to add two 4-bit numbers together.

When humans add two multi-digit numbers they do so serially, one bit at a time, starting with the rightmost bit, as in the example at the start of the section. While some of the earliest computers worked in this way, modern computers add all the bits at once, add them in parallel.

A 4-bit parallel adder has 8 inputs. Four bits from one number and 4 from the other. Let us call the bits from the first number A3-A0 (we number the bits from right to left) and those of the second number B3-B0. The output will be a 4-bit binary number, O3-O0, and a single carry bit, C. We use a full-adder to add each pair of corresponding bits along with the carry from the bits to the right. Here is the circuit for our 4-bit adder.

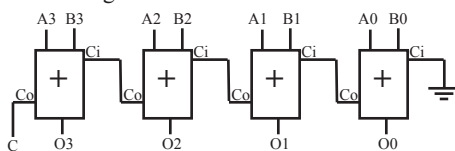


Figure 14-13 4-bit adder

Table 14-7: Full-adder truth table

I2	I1	I0	C	O
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

Note The dotted lines show the two half adders.

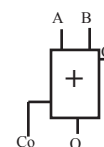


Figure 14-12 Full-adder Symbol

Info There is one small problem with this sort of circuit. If we change one or both of the input numbers, then it takes some time for the output to settle down to the right value because the carry information has to travel across the whole set of adders. For example, if it takes 50nS for one adder to respond to a change in its inputs, then 50nS after the change in the inputs only O0 is correct. At this time, the carry out from this bit settles down and so it will take another 50nS before O1 is correct. Once O1 is correct, its carry is correct so that O2 will be correct after another 50nS. Finally, O3 and the final carry are only correct 200nS after the initial change in the inputs. The carry information is said to **ripple** across the adder. Clearly, the problem gets worse the more bits we add at once. More modern addition circuits add extra gates to reduce the time it takes the carry information to propagate, making each output depend directly on more inputs instead of only through the carry chain. These are called **look-ahead carry** circuits. They are more complex but offer much better performance; a common trade off.

'283 4-bit Full Adder

Chip manufacturers do not seem to produce a chip with several independent full-adders but there is this complete 4-bit adder. Figure 14-12 shows the pin-out diagram. The wires A4-A1 are the four input bits from the first input word. The wires B4-B1 are the input bits for the second input word. C0 is the carry input from any previous stage and C4 the carry out to the next stage. Finally, Σ4-Σ1 are the four output bits. With a total of 9 inputs, the full truth table for the device requires 512 rows and is impractical to include. The chip manufacturers have produced a weird condensed truth table but it is so bizarrely incomprehensible that I have left it out. The function of the chip speaks for itself.

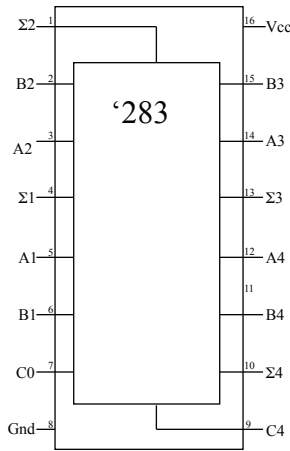


Figure 14-14 74HC283 4-Bit Binary Adder

The '283 is one of the more complex MSI chips. It has 36 separate gates inside it. Figure 14-15 shows the internal logic schematic. Note that it is *not* organized to show the half adders from which it is made. As I discussed above, the carry is not generated by rippling the information across the whole set of adders. Instead, the 6 gates ending up on pin 9 generate the carry directly from the input bits. This is called a **fast carry** circuit. It allows the circuit to claim almost the same **propagation delay** for the carry (49nS) as for a lowest order input bit (39nS). The **propagation delay** is length of time that elapses after you change one of the inputs before the output changes.

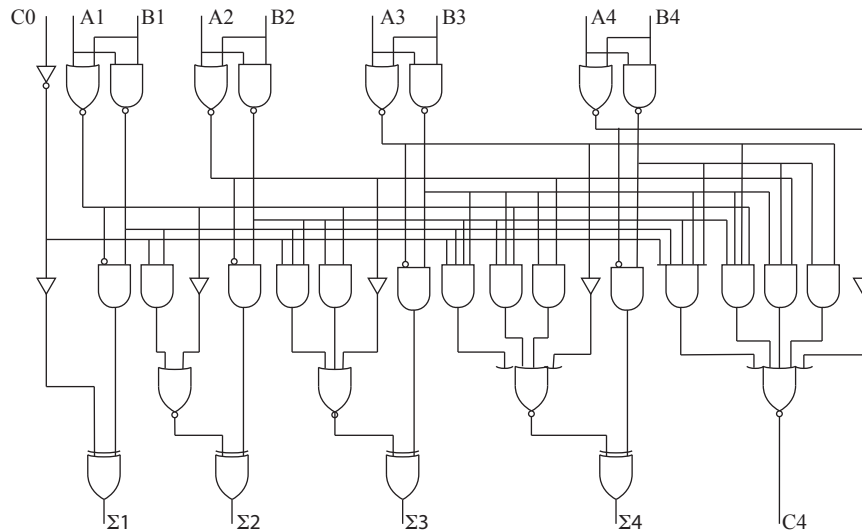


Figure 14-15 Logic circuit for 74HC283

14.5.2 Subtraction

Standalone binary subtractors are not as common as adders. They are usually found as part of more complex arithmetic/logic units. However, magnitude comparators are somewhat more common and they use exactly the same logic as subtractors so it is well worth our time to look at subtraction. Subtractors and comparators have extra outputs in addition to the output and carry bits. A comparator usually has >, =, and < outputs which provide information on the relative magnitude of the inputs. A subtractor has a zero output that is true if all the output bits are 0, a sign output to tell you if the answer is negative, and a borrow output. We will look at the subtractor form because it is the one found inside computers. Both forms, >, =, < and zero, carry, and sign outputs provide the same information; they just provide it in different ways.

We begin by looking at a simple example of subtraction and see what our tools are going to be.

$$\begin{array}{r} 100 \quad 100 \quad 100 \\ - \quad 1 \quad - \quad 11 \quad - \quad 1 \\ \hline \quad \quad \quad 1 \quad \quad 11 \quad \quad 1 \\ \hline \end{array}$$

As with addition we operate one bit position at a time and each position can involve three input bits, two inputs and a borrow. Unlike the adder, the inputs are not symmetric, A - B is not the same as B - A. Table 14-8 shows the truth table for inputs X, Y, and Bi, where we are computing X - Y and Bi is **Borrow in** from the previous stage. The outputs are O and Bo (**Borrow out**).

Table 14-8: Subtract truth table

X	Y	Bi	Bo	O
0	0	0	0	0
0	0	1	1	1
0	1	0	1	1
0	1	1	1	0
1	0	0	0	1
1	0	1	0	0
1	1	0	0	0
1	1	1	1	1

Interestingly, the output column is exactly the same as that for the adder, $X \oplus Y \oplus B$, but the borrow is somewhat different from the carry. The borrow is $X \cdot Y + (\overline{X \oplus Y}) \cdot B_i$. A little thought shows us that the only difference between addition and subtraction is in the carry circuit of the half-adder. While the half-carry was $A \cdot B$, the half-borrow is $\overline{X} \cdot Y$, an asymmetric formula as we would expect. So if we add one NOT gate to the half-adder we get the half-subtractor (Figure 14-14).

We can then build a full-subtractor from two of these in the same way that the full-adder was built from its two half-adders (Figure 14-15).

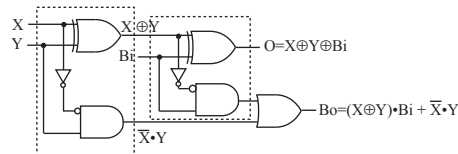


Figure 14-17 Full subtractor

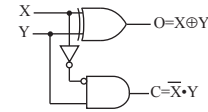


Figure 14-16 Half-subtractor

Now we have the full-subtractor we can build multi-bit subtractors and comparators in just the same way that we did with adders. As I mentioned, we usually add extra outputs to these multi-bit subtractors to tell us about the relative magnitudes of the X and Y inputs. One of these extra outputs is just a copy of the top output bit, called the **sign** bit or **negative** bit. Since, in 2's complement notation, all positive numbers have a zero in the top bit and all negative numbers have a 1 there, the top bit is 1 only for negative numbers. The other extra output is the NOR of all the output bits. It is 1 only if all of the output bits are zero and is used to test whether the two input numbers were equal.

There is no subtractor listed in current catalogue of 74HCxx devices but there is a 4-bit magnitude comparator. Here is a look at that circuit.

‘85 4-Bit Magnitude Comparator

Figure 14-16 shows the internal circuit diagram for a 74HC85 4-bit magnitude comparator. Once again, the circuit is not laid out to show the full-subtractors that live in the circuit for the magnitude comparator but the XOR gates in the first layer of half-subtractors at the inputs are clearly visible. The second layer of half-subtractors, the ones that compute the individual output bits, are hidden in the output logic. That layer is needed in the subtractor but is not really there in the magnitude comparator because it doesn't generate the difference output; it only generates the $>$, $=$, and $<$ outputs.

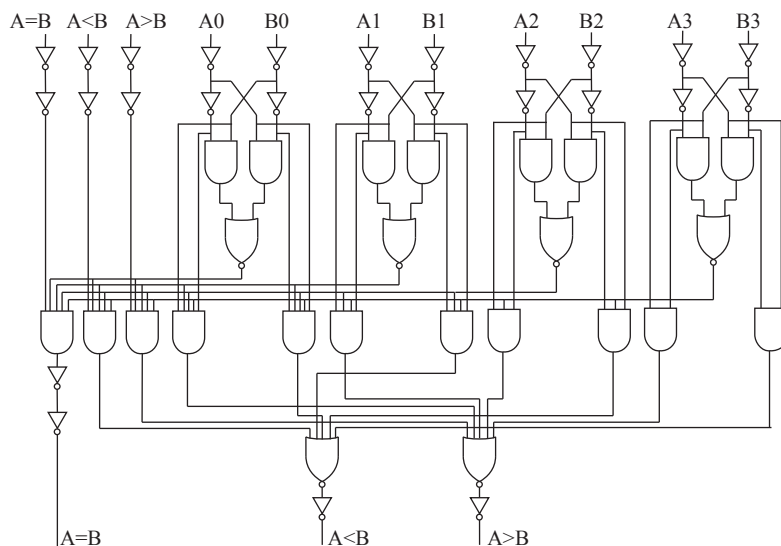


Figure 14-18 74HC85 4-Bit Magnitude Comparator

Note The $A>B$ in etc. are inputs provided for cascading these devices to longer word lengths. For example, you can compare two 8-bit words using two of these devices. One device compares the bottom 4-bits of A with the bottom 4-bits of B. The second compares the top 4-bits of A with the top 4-bit of B. You connect the $A>B$ out of the first chip to the $A>B$ in of the second chip and so on. That way the second chip knows how the lower order bits affect the results of its computation.

14.5.3 Arithmetic logic units

An arithmetic logic unit combines all of the arithmetic and logic functions into a single package. In addition to the inputs for two sets of data bits, for two input numbers, A & B, it has a set of function inputs. The function inputs select which of a wide variety of combinations of A and B is presented as the output of the circuit. For example, the '181 is a 4-bit arithmetic logic unit in a 24-pin package. It has five function select inputs and so can generate 32 different combinations of its two 4-bit input words. Many are rather peculiar but there are some really useful ones on the list including 0, 1, A, B, A OR B, A AND B, A plus B, A minus B, B minus A, and so on. Such a circuit is normally used only in the heart of a digital computer and I am not including the details because the chip is just too complex for normal use.

14.6 Building Digital Integrated Circuits

Because of the way in which modern FETs are built, by diffusing the various elements and layers into a single piece of silicon through a set of masks, it is easy to build several FETs on a single piece of silicon. These individual FETs can be interconnected using the metal layer of the device to form complete logic gates. The chip designer is free to tailor the characteristics of the individual FETs to suit their function without having to worry about the general utility of the FETs. This is done by adjusting the geometry of the individual devices. For example, small area FETs have high on-resistances and poor current handling abilities but low gate capacitances and so can switch very rapidly.

Let us look at the construction of a simple NOT output stage. Since this uses only 2 FETs it is fairly straightforward to show the process in 2-D figures. More complex gates use the same ideas but must spread into the third dimension. The actual devices are laid out in a very thin layer on the surface of the silicon wafer and can be seen if you sand away the top of the plastic IC package (extremely carefully, it is easy to sand away the whole silicon chip) and examine the device under a microscope. Note that this always ruins the device!

14.6.1 Metal-Gate CMOS Processing

Info The single-metal layer process described here was chosen for its comparative simplicity. Industry has moved on to more complex processes using both metal and poly-crystalline silicon interconnection layers, sometimes several of them. These processes are more complex in their details, using several more processing stages, but operate in basically the same fashion as the simple method described here. Further information can be found in, for example, application AN-310 from National Semiconductor Corp.

If you look back at Section 11.9, you will see that n-channel and p-channel FETs are constructed in very similar ways except that n-channel FETs must be built upon a piece of n-type silicon while p-channel FETs must be built upon p-type silicon. The first challenge in building a CMOS circuit is to fabricate both types of device on a single wafer of silicon. IC manufacturers do this by first creating large wells of p-type material for each of the p-channel FETs in the device.

They do this by the usual process of creating masks with holes in them and then adding p-type impurities to the system. The impurities can only pass through the holes and so the silicon wafer ends up with a set of p-type wells in exactly the places where the p-channel FETs will later be needed.

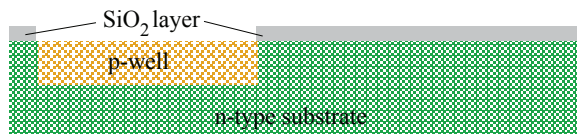


Figure 14-19 CMOS Stage1 Forming the p-well

Figure 12-17 shows the result of this processing. The p-well is formed by a process called ion-implantation where electric fields are used to force p-type impurities deep into the silicon wafer. These impurities can pass through the silicon layers but are blocked by the silicon dioxide mask layer. That layer can be added by heating the wafer in an oxygen-rich atmosphere and removed by etching in acids.

Info Silicon Dioxide is found in nature as the mineral quartz. The form produced in semiconductor processing is a form of glass.

Once the p-type well is in place, the silicon dioxide, or SiO_2 , is regrown over the whole surface and a new set of holes cut. High concentrations of p-type impurities are then diffused into the silicon. Again, the impurities can pass through the holes but are blocked by the SiO_2 layer.

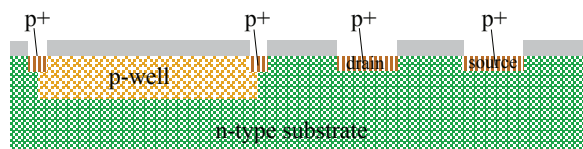


Figure 14-20 CMOS Stage 2 Forming p+ wells

As you can see in Figure 12-18, two of these regions will later become the source and drain of the n-channel FET. The other two regions, those surrounding the p-type well, are called **guard rings**. In the final device they will act as barriers to prevent current from leaking from one transistor to another through the n-type substrate.

The high concentrations of p-type impurities introduced in this step make the p+ regions quite highly conductive and so the right hand wells make good drain and source regions for the FET that will appear there. These layers are only very shallow because they are formed by diffusion instead of ion-implantation. In this process the silicon wafer is simply exposed to a gas containing the p-type impurities and the gas atoms creep into the very surface of the silicon wafer.

We are now finished with the p-type impurities. Once again the SiO_2 is regrown over the whole surface and a new set of holes cut. This time high concentrations of n-type impurities are diffused in, forming a set of n+ wells under the new holes in the SiO_2 .

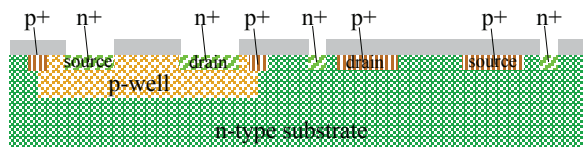


Figure 14-21 CMOS Stage 3 Creating n+ wells

Two of the new n+ wells will form the source and drain of the p-channel FET while the other two form guard rings for the n-channel FET (CMOS Stage 3 Creating n+ wells). All of the impurity wells have now been created. Next the SiO_2 is regrown and a new set of holes etched. A very thin layer of SiO_2 is grown in these holes to form the gate oxides of the new FETS. Further holes are etched in the SiO_2 layer to allow the upcoming metal layer to contact the silicon and a thin layer of aluminum is evaporated over the chip. Holes are cut in the Al layer, leaving only the gates and the metal interconnections, and a final layer of SiO_2 is added by a process called vapor deposition. The final device is shown in Figure 14-22.

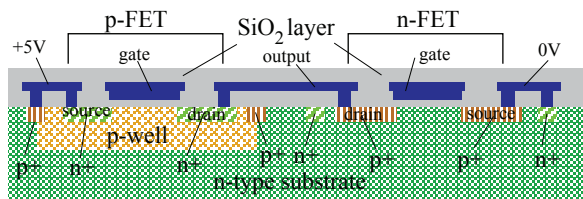


Figure 14-22 CMOS construction complete

On the left is the p-channel FET that is the upper transistor of the NOT gate. Its channel will be formed in the p-type well underneath the gate oxide between the source and drain in the well. Its drain is connected by the metal layer to the drain of the n-channel FET.

Several things are not apparent in this cross-section figure because they happen out of the plane of the figure. Most obvious is the connection from the metal layers, trapped under the top SiO_2 layer, to the outside world. The metal traces are brought to the surface in pads formed on the very edges of the chip of silicon and very fine wires welded to these pads to make the contacts.

Similarly, the gate of the n-FET and the gate of the p-FET, whose junction forms the input pin to the gate, are connected together by the metal of the gate layer in another part of the chip.

Figure 13-23 below shows an electron micrograph of one of the four NAND gates in a 74HC00 chip. The large blobs round the top and left edges of the picture are the pads to which external wires were attached. The serpentine pattern in the main part of the picture is the top

of the various FETs of the NAND gate. Since the gate is buffered the actual NAND portion is only a fraction of the complete gate. It is the more complex section forming the leftmost half of the FET region. To the right of that is a very narrow row of small fast FETs forming an inverter and then the large, outlined, region is the output buffer FETs.

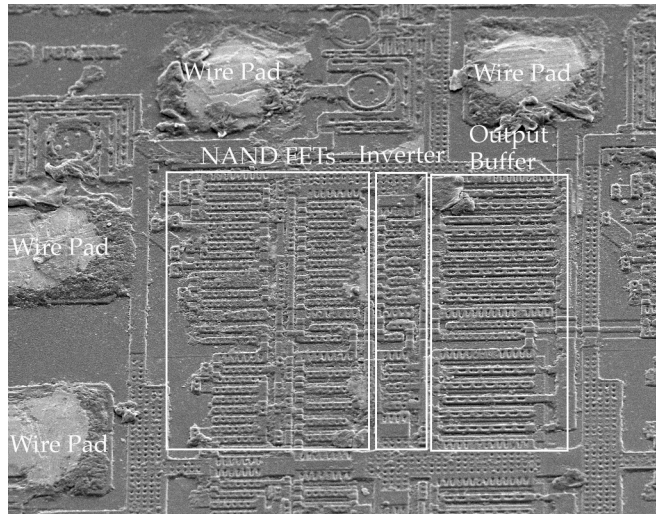


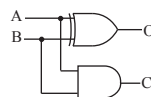
Figure 14-23 Electron Micrograph of a NAND gate

Summary

This chapter is full of individual examples of the rules from the previous chapter and so very little new to summarize.

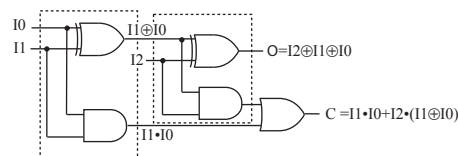
You should be able to recognize a multiplexer, a demultiplexer/decoder, and the half- and full-adder circuits.

Half-adder



B	A	C	O
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

Full-Adder



I_2	I_1	I_0	C	O
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

Exercises

1. Design a logic circuit to add together two 2-bit binary numbers to give a 3-bit binary result.
2. Give the truth table for a 2-bit by 2-bit multiplier. The inputs are a pair of 2-bit unsigned binary numbers and the output is a 4-bit binary product.
3. Find logic equations for each of the four outputs of the 2-bit by 2-bit binary multiplier.
4. Draw a circuit for 2-bit by 2-bit binary multiplier using as few gates as you can. You may use any mixture of AND, NAND, OR, NOR, XOR, and XNOR gates that you want. Similarly, you may use gates with any number of inputs.
5. Design the internal logic for a 7447 decimal display decoder. It should have four inputs, which are treated as a binary number in the range 0-9, and seven outputs. The outputs should be designed to drive the seven LEDs of a 7-segment display. This has 7 LEDs arranged in the pattern shown on the left. For example, to display the digit 3 you would turn on segments a, b, c, d, and g and leave e and f off. You should assume that a 1 on an output will turn on the LED and a 0 will turn it off. You may assume that the inputs will never take any any state outside the range 0-9.
6. In Q5 you were told to assume that the input would never be >9 . What outputs would your circuit actually display for each of the excluded input patterns? How would you alter your circuit to make it display correctly all 16 possible input patterns, from 0-15?

Chapter 15:Programmable Logic

15.1 Introduction

A comparison between an PC-AT style motherboard from 1986 and a modern Pentium class motherboard is a dramatic illustration of the trend in digital circuitry away from small, cheap, single-function chips towards large, expensive, extremely complex multi-function ones. The 1986 motherboard provides only the simplest of support functions for the CPU, memory controller, keyboard controller, interrupt support, and bus control logic. However, it takes about 50 chips, excluding the CPU and memory, to provide that support. By contrast, the 1996 motherboard has only ten chips apart from the CPU and memory but in addition to the resources of the earlier board it also provides controllers for two floppy disk, four hard disks, two serial and one parallel ports, and even a full super-VGA video system. These functions would have required at least four additional plug-in boards in 1986. Indeed, the video card alone for the 1986 computer had another 100 chips on it and it provided less than one tenth the capability of the modern board. What happened to all those chips and why?

What happened is that all the gates that were scattered throughout the dozens of chips in the mid-1980's became integrated into the small number of large chips in the 1990's. The trend began with the development of programmable logic arrays that allowed a single chip to replace several different simpler chips. This not only allowed manufacturers to buy and stock many fewer types of chip, since one programmable chip could emulate tens of different single-function chips, but also allowed them to make fixes or improvements to their products without having to rework the circuit boards. It used to be quite common to see circuit boards with a couple of extra wires or a resistor or two soldered on by hand after the board was built to fix some problem that was found after the board were manufactured. It was far less costly to fix the existing boards than to have new ones built. With the new programmable chips a fix could usually be made by altering the programming of the chip since much of the hard wiring was replaced by the programmability of the new chips. As chip manufacturers got better at putting thousands of devices on a chip, the complexity of the chips increased. At the same time, the price of the chips dropped so that a device that will sell thousands of units can now be custom designed with the gates laid out to suit one task, one very complex task. These are called ASICs or Application Specific Integrated Circuits. Such circuits now make up a large fraction of all the output of semiconductor manufacturers. Indeed, there are many companies which design and market such ASICs that are actually built by separate specialist fabrication companies called **chip foundries**.

The reason for this change is money. The cost of making a printed circuit board has remained fairly constant while the cost of the integrated circuits has fallen rapidly. The largest cost of a modern motherboard, apart from the CPU chip, is in building the board and assembling the components onto it. The fewer components you have to put on the cheaper the finished product. With current manufacturing costs, it is far cheaper to use one or two large ICs that each cost \$20 to manufacture than it would be to use a dozen simpler chips costing only a few cents a piece. Thus we have seen circuit boards shrink in size and we have seen the older DIP packages, which had to be inserted through the boards, replaced by modern surface mount packages that sit on the board. Drilling holes and plating them through, so that the top of the hole is electrically connected to the bottom of the hole, is quite expensive. A circuit board with no holes costs a lot less so we now see small boards with relatively few large ICs and a scattering of surface mounted discrete resistors and capacitors.

Neither the scientist producing a small number of circuits for his research, the amateur making a single homebrewed device, nor the development labs of even the biggest manufacturers can make use of the specially manufactured ASICs. Not only is the set-up cost very large, several thousand dollars for the first chips, but the designs require expensive software packages (many tens of thousands of dollars) and a lot of specialized expertise. However, programmable chips such as those we will discuss in this chapter are a valuable step in that direction.

15.2 The PAL

The simplest, and oldest, kind of programmable logic is called Programmable Array Logic, PAL. PAL chips implement the sum-of-products method of logic design in hardware. The very simple structure of the expressions in this design method leads to a simple hardware structure. Each output comes from an OR gate that takes its inputs from a set of AND gates. The inputs of the AND gates are either connected directly to device inputs or are connected to them through inverters. So we provide a set of inputs each with its own inverter and a set of AND gates and then we can realize many different truth tables by connecting up the AND gates to the inputs in different ways.

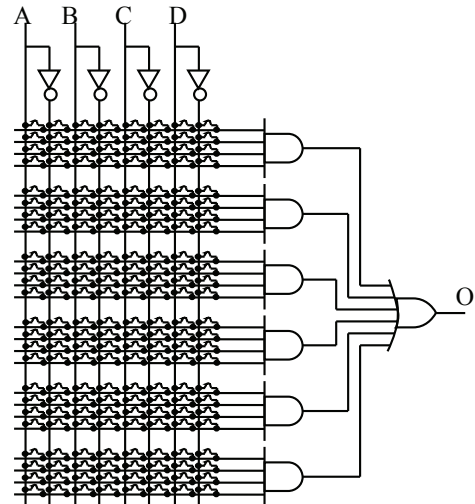


Figure 15-1 PAL Structure

Figure 15-1 is an example that can realize any four-input, one-output truth table that has no more than six 1's in its output column.

All we have to do to implement a particular truth table is to connect up the inputs of the AND gates to the correct collection of inputs and we have implemented the truth table. To build this from individual chips would take a minimum of 5 chips (1 74HC04 hex inverter, 3 74HC20 dual four-input NAND, and 1 74HC30 eight-input NAND). However, it only needs five external signal wires so that we could make a more complex system, for example with 6 inputs and 8 outputs, and still fit it inside one 16-pin chip. The trick would be wiring up the inputs to the AND gates. Programmable logic arrays do exactly that.

A	B	C	D	>
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	1
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	1
1	0	0	1	1
1	0	1	0	0
1	0	1	1	0
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	0

The oldest kind of PAL chip used tiny aluminum wires to make the connections. They were so thin that a current of several mA would vaporize the whole wire in exactly the way that a fuse burns out. The little wires are thus called fuses and are the little wiggly things in the figure.

To implement a particular truth table all you have to do is to blow all the fuses that you don't want and leave the ones that you do. For example, Table 15-1 is the truth table for a comparator. It treats the inputs as two two-bit signals, AB and CD, and the output is true if the two-bit number AB is greater than the two-bit number CD. For example, when AB = 01 and CD = 00 the output is true but when AB = 01 and CD = 01 the output is false.

We construct the logic expression using sum-of-products, as usual

$$O = \bar{A} \cdot B \cdot \bar{C} \cdot \bar{D} + A \cdot \bar{B} \cdot \bar{C} \cdot \bar{D} + A \cdot \bar{B} \cdot \bar{C} \cdot D + A \cdot B \cdot \bar{C} \cdot \bar{D} + A \cdot B \cdot \bar{C} \cdot D + A \cdot B \cdot C \cdot \bar{D}$$

Then we work our way through the fuse array. In the first row of fuses we blow all but the second one, leaving that AND input connected to A. In the second row we blow all but the third fuse, leaving that one connected to B and so on. Figure 15-3 shows the resulting circuit.

Clearly, you have to be very careful to get the right fuses. Once a fuse is blown there is no going back! A typical small PAL, the 16L8, has 2048 fuses and most of them have to be blown in each design. If you have to specify by hand which fuses to blow and which to leave, then very few of your devices will actually work. The manufacturers very quickly realized that these devices needed support and produced not only special programmer hardware to blow the fuses safely but also software to drive the programmer. The software takes its input in the form of logic equations just like the one we used and then derives the fuse map from that and sends it to the programmer. We will see more of this software later.

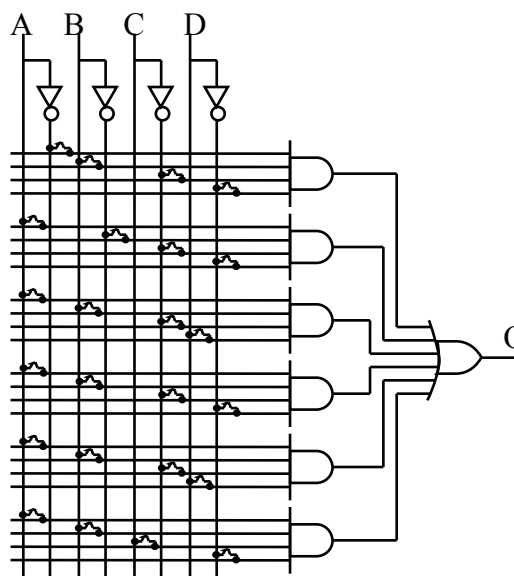


Figure 15-2 PAL Implementing 2-bit magnitude comparator

15.2.1 Some Real PALs

Manufacturers have created a large variety of different PALs that differ only in small details. Some logic functions need a lot of inputs and relatively few outputs while others need fewer inputs and more outputs. Accordingly, manufacturers supply a wide variety of different numbers of inputs and outputs.

The sum-of-products method is best suited to logic functions which have fewer 1's than 0's in their truth table output columns. We normally implement truth tables with many 1's and few 0's using the product-of-sums method but you can also use the sum-of-products method to create a complementary logic function and then add an inverter to make all the 1's into 0's and vice versa. In order to support this design method manufacturers make PALs with NOR gate outputs instead of OR gates. They even make ones where you can program each output to be either OR or NOR by blowing an extra set of fuses.

PALs are given part numbers that describe their internal structure. Each number is of the form nXm where

n is the number of input pins

X is a letter describing the outputs

H means OR gate outputs

L means NOR gate outputs

P means Programmable as either OR or NOR

m is the number of output pins.

Some of the devices available are

10L8, 10H8, 10P8, 12L6, 12H6, 12P6, 14L4, 14H4, 14P4, 16L2, 16H2, 16P2, 12L10, 12P10, 14L8, 14P8, 16L6, 16P6, 18L4, 18P4, 20L2, 20P2, 20L10, 16L8, and 20L8.

There are also more advanced versions that include "registered" outputs. We shall return to these in Chapter 16.

PAL chips became popular quite quickly. They were often used in memory decoder circuits, allowing one device to respond to only that range of addresses that belonged to it. Modern computers may have very large numbers of address lines. Current (2013) 64-bit processors from Intel and AMD have 40 physical address lines. This makes the task of allocating a particular block of addresses to a particular device much more complicated. The simple 3-to-8

line decoder circuit that we saw in Chapter 14 is no longer sufficient. PALs made it possible for one device to respond to a specific range of addresses without having to fully decode the memory as we did then. The main problem with them was that once you had blown the fuses, the device was committed and could not be altered. It was also a nuisance that there were so many different types, each one with its special architecture.

15.3 The GAL

It was not long before manufacturers found a way to replace the fuses with programmable switches. Each switch consists of an FET switch controlled by a capacitor. If the capacitor is charged then the switch is turned on and the connection is made. If the capacitor is discharged then the switch is turned off and there is no connection. The capacitors are made using the same technology as EEPROMs and will hold their state for years once programmed. The electronic switches can be programmed and erased by special circuitry built into the chip; circuitry which is not used at all during normal chip operation. Thanks to these electronic switches, a single device can be programmed and used for one job and then reprogrammed and used for another job at a later time. This is perfect for prototyping since it allows you to re-use your chips rather than throwing them out. It is also useful in commercial systems since an instrument can be altered or upgraded simply by reprogramming one or more GALs without having to throw out and replace chips.

A further benefit of these electrically alterable chips is that manufacturers have found ways to make their internal circuitry more flexible. Instead of the huge range of different architectures described in the preceding section, there are just three common GAL devices: the 16V8 in a 20-pin package and the 20V8 and 22V10 in 24-pin packages.

These devices have sophisticated programmable output circuitry that allows a single GAL to replace many different PALs by programming its architecture to mimic that of the PAL. Indeed, the acronym GAL stands for **Generic Array Logic** since a single generic GAL can be used in place of a range of specific PALs. For example, the 16V8 can be used in place of any of 21 different 20-pin PALs.

GALs became cheap enough to replace the less flexible PALs. They are ideal for the experimenter and scientist since they are reusable, easy to program, and you only need to stock 2 or 3 kinds. Single GAL chips cost \$1-3 depending on complexity and speed and can be programmed with a device costing a few hundred dollars. Unfortunately, the last manufacturer ceased to produce GALs in 2011, though there are still large quantities in the supply chain so they remain obtainable. Their replacements are still more flexible chips that, unfortunately, are usually only available in surface mount packages that are difficult for experimenters to play with.

15.3.1 The 16V8

The smallest and simplest of the GAL chips is the 16V8. These have been made by a large number of different manufacturers over the years and so you will find chips with a wide range of device number prefixes such as ATF16V8 and GAL16V8. Similarly there have been various speed, temperature, and package variants of each chip so that you will also find a range of suffixes. I have devices with markings that include GAL16V8D-25LP, PALCE16V8H-15, and ATF16V8B-10PC.

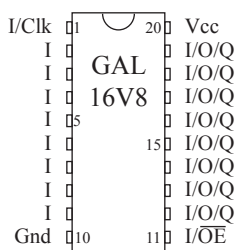


Figure 15-3 GAL16V8

While various details, including the programming times and voltage, vary from chip to chip, all of the variants share the same internal architecture and pinout and so can accept the same program and are completely plug compatible. That is, we can pull out a chip from one manufacturer and replace it with a chip from another and trust that the circuit will continue to work, so long as we put the same program in.

The 16V8 architecture specifies a 20-pin chip with 8 dedicated inputs, 8 pins that can be used as inputs or outputs, two pins that are either logic inputs or special inputs, and two pins for power and ground (Figure 15-3). The eight dedicated inputs are marked I. They are not as-

signed specific names in the fashion of a normal chip because they don't acquire specific functions until the chip is programmed. The eight pins marked I/O/Q can be used as either inputs (I), combinatorial outputs (O), or latched outputs (Q) depending on the program. Each pin is individually programmable and you can have any mixture of inputs, standard, and latched output. The other two logic pins, pins 1 and 11, have special uses that are only important when the output latches are in use so we shall leave those uses to a later chapter. When the chip is in use as a purely combinatorial chip these pins are also available as inputs.

Internally, the chip has a large array of AND gates with fuses to connect them to the 10 input pins and to the outputs from 6 of the I/O/Q pins. Each of the eight I/O/Q pins is driven by a chunk of logic called an output logic macrocell (OLMC) which has 7 AND-gate inputs. This means that each output can implement any logic equation that can be written with no more than 7 sum-of-products terms. If you have a function that needs more than 7 such terms then you have to use the output of one macrocell as the input to another, which means that a signal will take twice as long to propagate through the device since it will essentially have to go through twice.

The OLMCs allow each output to be configured as either an OR output, a NOR output, or a D-type latch output. In each case the output also passes through a tri-state gate so that GALs can be used in bus systems. There are some complicated restrictions on which outputs have which features that are described in detail in the device data sheets. However, the programming software (see below) understands these restrictions and will issue error messages if you try to compile a program that violates them.

15.4 Programming GALs

A typical GAL contains several thousand individual switches, each of which must be set to the correct state in order for the device to operate correctly. It would extremely tedious and nearly impossible for a human to translate a design into an error-free list of switches to be set and cleared. However, a computer program can handle such tedious tasks with ease and the chip manufacturers realised this from the start. The first such programs were made available as FORTRAN source code by the early PAL manufacturers. These have been improved over the years to support a wider and wider range of programmable chips.

Designing a working systems using these programs and a GAL is a three step process.

1. You design the internal function of the device. This results in a truth table, a set of Boolean equations, or a state diagram (see Chapter 17).
2. You input the design into a computer in a suitable form and run a compiler program to turn the design into a list of switches to set and clear. This list is created in a standard format called **JEDEC format**.
3. You take a suitable device and put it into a special programmer. This is a piece of hardware that connects to a desktop computer and that can control the signals on all the pins of the PAL device. The hardware is controlled by software running on the desktop computer that reads in the JEDEC file and uses the information to set and clear the switches inside the PAL.

There are a number of different device programmers available and each comes with its own specialized control software. However, the PAL compilers are usually separate programs written by different companies, who specialize in such software.

The first generation of PAL design software was a FORTRAN program called PALASM that was written and distributed by Monolithic Memories Inc., the designer of the first PALs. This software took a text description of the truth table that the device was to embody and converted it into a list of fuses to blow. This software was quite limited in the range of devices that it could handle and it required that the equations be organized according to the layout of the final chip. It was important in the early days of programmable logic (late 1970's through the 1980's) and was extended and developed into the mid 1990s when development was abandoned and the

Info JEDEC

JEDEC was originally an acronym for Joint Electron Devices Engineering Council, an industry group that is now called the JEDEC Solid-State Technology Association. This body sets standards for such things as the numbering of integrated circuits, the pin-out maps for computer memories, and the format of files to transfer information between various kinds of program involved in electronic design. These files include the fuse maps for programmable chips that we need here.

software released to public. The PALCMPL program is a version of PALASM that was sold by Timely Software in the early 1990s.

The main market for PAL design was then taken over by two products, ABEL (created in 1983 by Data I/O, now a part of XILINX) and CUPL (created by Logical Devices Inc. and now belonging to ATMEL). These are much more complex packages that can accept input in a wide range of formats and can handle not only GALs but also more complex devices such as PEELs and CPLDs. These programs not only convert the input specification into a list of switches but they can also perform optimisations to find the most efficient way to implement even extremely complex specifications. As these programs have in turn been superceded by still more complex systems they have also made their way into the public domain. The windows version of CUPL, WinCUPL, is freely available from ATMEL corporations web site.

FPGA

Field Programmable Gate Arrays mix large arrays of prgrammble cells, rather like PALs with arrays of memory and extremely large arrays of inter-connections that allow the logic to be wired very flexibly.

FPGAs can implement complete microprocessors, including internal RAM and PROM. They can also implement complex computations in hardware that can process signals much faster than is possible in software. For example, many of the complex operations involved in decoding MPEG encoded video can be implemented in FPGAs that offer far higher performance/Watt than is possible with conventional processors.

Such complex devices would require tens of thousands of idividual equations to specify their operation. Thus more abstract languages are used that allow simple circuits to become building blocks for more complex one in a heirarchical fashion. Then the complete device is described in a very high-level way that can be expanded into the thousands of equations automatically and used to program the millions of connections in the chip.

The latest generation of programs use more powerful description languages called Hardware Description Languages that can describe the hardware for complete CPUs in a few handreds lines of code. These languages target more the most complicated families of programmable logic chips, FPGAs that may have hundreds of pins and millions of gates and fuses. Whole processors are available as sub-programs that can be incorporated into larger systems. The two main languages are VHDL and Verilog and they are supported by (usually expensive) development systems from major chip manufacturers such as XILINX and Lattice.

15.4.1 PALASM, PALCMPL, and CUPL

These three languages offer very similar simple forms for describing simple circuits to implement in PALs and GALs. They begin with a header that includes various bits of information about the program and circuit. The header includes a command that tells the program which type of chip is to be used. Then follow sets of lines that associate logical signals, specified by names, with specific pins on the chip. The programs know which pins can be used as inputs and which as outputs and so can check that the design is using the pins correctly. Then there may be lines that configure the internal structure of the device. For example, while an output from a PAL is a simple OR or NOR gate, the output of a GAL go to a more elaborate circuit that can be programmed to operate as OR, as NOR, or as latched or registered outputs, as discussed in the next few chapters. The program may need to specify how these output circuits are to be configured. Finally, the program will have a set of equations, probably in sum-of-products form, relating the outputs to the inputs. These are the equations that determine the final function of the device. In addition to the actual program, these languages allow the user to insert comments using a syntax borrowed from the C language. The comment is a string that begins with the character pair ‘/*’ and ends with the pair ‘*/’. Anything at all can appear between these delimiters and it will be ignored by the compiler. This allows the user to add human-readable information about the design without the compiler caring.

**Table 15-2:
Gray-Binary Decoder**

Gray Code Input			Binary Output		
I2	I1	I0	O2	O1	O0
0	0	0	0	0	0
0	0	1	0	0	1
0	1	1	0	1	0
0	1	0	0	1	1
1	1	0	1	0	0
1	1	1	1	0	1
1	0	1	1	1	0
1	0	0	1	1	1

15.4.2 A real example: A Gray-code to Binary converter.

A gray code is commonly used by mechanical input devices such as rotary encoders. It encodes numbers in binary bit patterns in a way that is useless for arithmetic but very useful for input devices because it has the property that only one bit position changes value between any two adjacent numbers. This minimises the error from slight misregistration in the device.

The code we are looking at is an alternate encoding of the first 8 integers using 3 binary bits. We may need to translate the code coming from such a device before using it as a binary number. We can do that with a code converter. The truth table for this is very simple to write (Table 15-2 on the left)—we simply put the gray code on the input side and the binary code on the output side.

Now we can write a set of equations using one of the formal methods. It does not actually matter which one since there are the same number of 1’s and 0’s in the output side of the table. Since it is a more natural fit for the GAL structure I will use sum-of-products.

$$O0 = \overline{I2} \times \overline{I1} \times I0 + \overline{I2} \times I1 \times \overline{I0} + I2 \times I1 \times I0 + I2 \times \overline{I1} \times \overline{I0}$$

$$O1 = \overline{I2} \times I1 \times I0 + \overline{I2} \times I1 \times \overline{I0} + I2 \times \overline{I1} \times I0 + I2 \times \overline{I1} \times \overline{I0}$$

$$O2 = I2 \times I1 \times \overline{I0} + I2 \times I1 \times I0 + I2 \times \overline{I1} \times I0 + I2 \times \overline{I1} \times \overline{I0}$$

Note It would be possible to simplify some of these equations (especially the third) but it may not be worth doing unless we have other constraints, since the hardware can implement these directly.

Next we have to map the terms to the chip. Each of these equations is simple enough to implement with a single term in the GAL and so we can make our assignments quite freely. I will use pins 2, 3, and 4 for inputs I2, I1, I0 respectively, and the corresponding pins on the other side of the chip for outputs.

PALCMPL

PALCMPL requires the user to supply a header that contains text descriptions to maintain the authorship of the code and also to specify which kind of chip the program is intended for. It does this with the 'header=' and 'device=' commands.

Next PALCMPL requires that we specify which logical signals are connected to which pins. We only have to supply 'pin=' definitions for the pins that we use. Unused pins will be assumed to be unconnected. We have already decided to use pins 2-4 as input and 17-19 as outputs.

Because the outputs of a GAL terminate in tri-state buffers we must arrange for these buffers to be enabled otherwise the outputs will not be connected to the pins. There are two ways to do this. We can either ground pin 11 on the chip with a wire or we can connect an internal logic 1 signal to the enables using the syntax <pin name>.oe = true. I chose to use the .oe method.

PALCMPL uses the usual & and + symbols for AND and OR gates so we get the following PALCMPL program

```

/*
 * This file describes a 3-input Gray-code to Binary decoder
 * in a GAL16V8 chip.
 */
header={
    COMPANY    Physics 245;
    DESIGNER   Brian Collett;
};
device = gall16v8;
/*
 * Define inputs.
 */
pin 2 = I2;
pin 3 = I1;
pin 4 = I0;
/*
 * Then the outputs.
 */
pin 17 = O0;
pin 18 = O1;
pin 19 = O2;
/*
 * Enable tri-state outputs. These configure the output cells of the GAL
 * as active outputs, removing the need to ground to pin 11.
 */
O0.oe = true;
O1.oe = true;
O2.oe = true;
/*
 * And finally the equations. NOTE that the spaces are not required
 * but they are also not forbidden. I have included them to make the
 * equations easier to read.
 */
O0=!I2&!I1&I0 + !I2&I1&!I0 + I2&I1&I0 + I2&!I1&!I0
O1=!I2&I1&I0 + !I2&I1&!I0 + I2&!I1&I0 + I2&!I1&!I0
O2=I2&I1&!I0 + I2&I1&I0 + I2&!I1&I0 + I2&!I1&!I0

```


CUPL

The CUPL program is almost exactly the same. The main differences are in the header format and in the name used to specify the device. CUPL requires a much more detailed header, though it ignores almost all of the fields. The only ones we have to worry about are the Name field and the Device field. CUPL uses the value in the Name field as the basename of the output file. Because it is an aging program it allows no more than 8 characters in this field. The value in the Device field specifies the nature of the device into which the program is to be loaded. CUPL uses this to verify that the program will fit in the device and to create the correct fuse map.

The other noticeable difference between the CUPL file and the PALCMPL file is that CUPL does not predefine identifiers for true and false. We have to use explicit binary numbers which must be prefixed with the string 'b' to tell the program that this is in binary. Also, CUPL uses the # symbol instead of + for the OR operation. It also supports \$ for the XOR operation. With those changes in mind here is the CUPL file for the gray-binary decoder.

Note Like PALCMPL, CUPL is very forgiving of white space. You can put extra spaces pretty much where you want. You can also insert a /*...*/ delimited comment anywhere that white space would be allowed.

```
Name      Gry2Bin ;
PartNo    00 ;
Date      1/10/2013 ;
Revision  01 ;
Designer  Brian Collett ;
Company   Hamilton College ;
Assembly  None ;
Location  ;
Device    gl6v8 ;
```

```
/* ***** INPUT PINS *****/
PIN 2 = I2;
PIN 3 = I1;
PIN 4 = I0;
```

Note CUPL treats all unmarked numbers as hex values. We have to use a 'b' here to mark the 1 as a single bit number because the value on the left is only a single bit wide. CUPL won't let you assign a value that is too big for the destination.

```
/* ***** OUTPUT PINS *****/
PIN 17 = O0;
PIN 18 = O1;
PIN 19 = O2;
```

In addition to hex and binary numbers CUPL also supports decimal numbers using the 'd' prefix and octal ones with the 'o' prefix.

```
/* ***** OUTPUT ENABLES *****/
O0.oe='b'1;
O1.oe='b'1;
O2.oe='b'1;
```

Remember CUPL uses # for the inclusive OR operator and also supports the XOR operator, using the symbol \$.

```
/* ***** EQUATIONS *****/
O0=!I2&!I1&I0 # !I2&I1&!I0 # I2&I1&I0 # I2&!I1&!I0;
O1=!I2&I1&I0 # !I2&I1&!I0 # I2&!I1&I0 # I2&!I1&!I0;
O2=I2&I1&!I0 # I2&I1&I0 # I2&!I1&I0 # I2&!I1&!I0;
```

In addition to accepting a device description in terms of equations, CUPL offers the ability to specify the truth table directly and can infer the equations itself. The syntax for this is best given by example. The following code can simply replace the equation section of the previous code and will result in exactly the same final device.

```
/* ***** TRUTH TABLE *****/
TABLE I2,I1,I0 => O2,O1,O0 {
'b'000=>'b'000;
'b'001=>'b'001;
'b'011=>'b'010;
'b'010=>'b'011;
'b'110=>'b'100;
'b'111=>'b'101;
'b'101=>'b'110;
'b'100=>'b'111;
}
```

The first line tells CUPL that this table has three input columns and connects them to the inputs named I2, I1, and I0. It then says that the table has three output columns and connects them to the outputs named O2, O1, and O0. In each case the rightmost name will be assigned to the least significant bit, the leftmost to the most significant, and so on. Since there are three input columns and three output columns we will need to use 3-bit numbers for the input and output specifications. These could be given in binary or octal. I used binary as it is more familiar.

The actual body of the table is enclosed in curly brackets, and you will get errors if you use any other kind of bracket. It consists of a set of rows for the table. Each row has a three bit number on the left of the \Rightarrow and a 3-bit number on the right. This means that when the inputs equal the left number then the logic will set the output to the right number. I put each row of the table on a separate line to make it look more like the original table. CUPL does not care. Since each statement is terminated by its own semicolon, you can put several statements on one line if you want. I usually avoid it as it makes programs harder to read.

15.4.3 Device Testing

Many hardware design languages offer the additional ability to test the design and the resulting devices. They use the concept of **test vectors**, lists of inputs and their expected outputs, to specify a sequence of tests for the device programmer to perform on the device. For example, we know that the gray-binary converter of section 15.4.2 should generate the output 101 if it is fed the input 111 (see the truth table above). We could create a file of test vectors for our CUPL program and then CUPL could check that the program works as intended.

15.5 A tutorial example using PALCMPL.

In Chapter 14 we met the 3-line to 8-line decoder circuit that set one of its eight outputs low depending on the 3-bit number present on its inputs. Here is the truth table that we saw then.

I2	I1	I0	O7	O6	O5	O4	O3	O2	O1	O0
0	0	0	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	1	0	1
0	1	0	1	1	1	1	1	0	1	1
0	1	1	1	1	1	1	0	1	1	1
1	0	0	1	1	1	0	1	1	1	1
1	0	1	1	1	0	1	1	1	1	1
1	1	0	1	0	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1	1

Table 15-3: 3-to-8 line truth table

Since the output array has far more 1s than 0s, I will use product of sums to generate the equations below. With eight columns we will need eight equations and since each column has only one 0 each equation will have only one term. Here are the equations.

$$O0 = I2 + I1 + I0$$

$$O1 = I2 + I1 + I0$$

$$O2 = I2 + I1 + I0$$

$$O3 = I2 + I1 + I0$$

$$O4 = I2 + I1 + I0$$

$$O5 = I2 + I1 + I0$$

$$O6 = I2 + I1 + I0$$

$$O7 = I2 + I1 + I0$$

Now to prepare these for the PAL compiler we first have to decide which pin we will use for which function and then write a text file that puts the equations into the right format. As you remember from Figure 15-3, pins 1-9 are inputs (labeled I) and pins 12-19 can be selected as either inputs or outputs. We'll need all of the outputs and let us specify pins 1-4 as inputs. We have a choice of either using pin 11 to enable the tri-state output buffers or getting the program to do so. I have chosen to program the chip to enable the tri-state buffers so that the chip will just work when we plug it in. We prepare our data for the compiler using the following format.

- 1) Put in a header telling the compiler what chip we are using.
- 2) Define the names for all input pins.
- 3) Define the names for all output pins.
- 4) Tell the compiler to make the I/O pins into outputs.
- 5) Give the logic equations relating the outputs to the inputs.

The notation for the equations is a little different from the notation that we are used to but is pretty easy to learn. Instead of putting a bar over a variable to negate it, we put an exclamation mark in front of it. Thus we will write as !I0.

Logical OR is done with '+' just the way we do normally but AND is done with '&' and XOR with '\$'.

Using these rules we get the following input file. This needs to be put into a plain text file and given a name such as DECODE.SRC. The standard Windows text editor Notepad is ideal for creating such files.

```

/*
 * This is a comment.
 * This file describes a 3-line to 8-line decoder with negative true
 * outputs. Note that the semi-colons at the end of each line are required.
 */
header={
    COMPANY      Physics 245;
    DESIGNER     Brian Collett
};
device = gal16v8;      /* This is the device that we will be using */
/*
 * Define inputs.
 */
pin 1 = I2;
pin 2 = I1;
pin 3 = I0;
/*
 * Then the outputs
 */
pin 12 = O0;
pin 13 = O1;
pin 14 = O2;
pin 15 = O3;
pin 16 = O4;
pin 17 = O5;
pin 18 = O6;
pin 19 = O7;
/*
 * which have to be forced to be outputs and told to turn on
 * (they use tri-state outputs that need to be enabled).
 */
O0.oe = true;
O1.oe = true;
O2.oe = true;
O3.oe = true;
O4.oe = true;
O5.oe = true;
O6.oe = true;
O7.oe = true;
/*
 * And finally the equations. If you need more space then you can put a
 * backslash at the end of the line and continue onto the next line.
 */
O0 = I2 + I1 + I0;
O1 = I2 + I1 + !I0;
O2 = I2 + !I1 + I0;
O3 = I2 + !I1 + !I0;
O4 = !I2 + I1 + I0;
O5 = !I2 + I1 + !I0;
O6 = !I2 + !I1 + I0;
O7 = !I2 + !I1 + !I0;

```

Once we have created our .SRC we need to compile it into JEDEC format. Because PALC-MPL is a rather antique program we have to run it from the command line. This means that we need to start by opening a Command window. We do that by going to the Start menu and selecting All Programs. From the list that comes up we select the Accessories folder and then select Command. This will open a window showing white text on a black background.

The Command window is probably not open at the directory where we put our source file so we first have to navigate there. In my case I stored the program in a directory called Phy245. I can make that the current directory by typing

```
>cd Phy245
```

at the > prompt. The command cd stands for Change Directory. We can tell that we succeeded because the text before the '>' prompt now ends Phy245 but we can check that we are in the right place using the command dir to list the contents of the current directory.

Once we are in the right place we can run the program. We have to tell it what file to use. The result should look like this:

```
PALCMPL decode8.src
```

If all has gone well, we will have a file decode8.jed containing a description of the fuses to be blown. We can see that the file is there with the dir command again. The JEDEC file is actually a text file in a simple format so we can ask Command to type the program out so that we can see it. This should produce something like this.

The next stage is to program the information into the chip. We do this with a special piece of hardware called a Device Programmer, which is available on one PC reserved for this purpose. I should be there to help with this but here complete instructions if you need them. These instructions are a little more complex but persevere.

- 1) First we have to copy the .JED file to the programming PC. We can either use a thumb drive or a network for this.
- 2) Next we make sure that the programming device is turned on (LED on, switch on back).
- 3) Then we start the programmer software, SuperPro, by clicking on the icon on the desktop.
- 4) Next we make sure that the device number and manufacturer shown in the device window are the same as the actual device that we are using.
- 5) Then we use the Load button to load the .JED file into the program.
- 6) Put your chip into the ZIF socket on top of the programmer and close the socket. Make sure that the chip is the right way up and in the right place in the socket. (See raised image to the left of the socket.)
- 7) Before we can put in the new program we have to make sure the chip is blank. So we erase the chip using the buttons on the left side of the main panel and then run Blank Check to make sure that worked.
- 8) At last we can program the chip by pushing the Program button and then, finally,
- 9) take the chip out of the socket.

At the end we have a customized chip ready to be built into our circuit.

Exercises

1. Write a PALCMPL program to use a GAL16V8 as a 2-bit binary multiplier.

Chapter 16: Sequential Logic

16.1 Introduction

All the combinatorial circuits that we have seen so far share the property that the output pattern of bits depends only on the input pattern. You cannot make a computer from combinatorial circuits alone because a combinatorial circuit cannot modify its behavior based on previous events. A computer has to have some way of remembering previous inputs and modifying its current behavior based on things that happened earlier. So we need to add to our armory of logic circuits a memory circuit. Ideally, we want to be able to present a single bit to such a circuit and to say “remember this bit”. The circuit should then hold that piece of information at its output until we tell it to change. Once we have such a circuit, called a **latch**, we can collect latches together and build memories that can hold more than one bit. This chapter describes how we can combine our logic gates to build such a latch and then illustrates some of the things that we can do with latches.

16.2 The Flip-flop

The secret to building a memory circuit lies in making one simple change to the way we have assembled logic circuits. So far, circuits have had clearly different sets of inputs and outputs. Now we are going to mix things up by using the output of a circuit as one of its inputs. This is called **feedback**, because we feed the output back to the input. Figure 16-1 shows our first circuit with feedback. It is called a **flip-flop** or a **bistable**.

Obviously the first thing we try to do is to make a truth table. This circuit is small enough to make bit-following quite easy. Indeed a few seconds thought should convince you that it will be very hard to write down a logic expression for this circuit. Bit following is all we have.

As usual, I am going to work through the input states in numeric order and try to build a truth table for the whole device. First, I want to remind you of the truth table for a single NAND gate (Table 16-1). As we see, the output is 1 unless both inputs are 1.

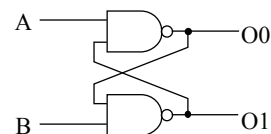


Figure 16-1 Flip-Flop

A	B	NAND
0	0	1
0	1	1
1	0	1
1	1	0

Table 16-1: NAND Truth table

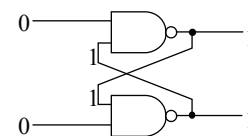


Figure 16-2

I will start our exploration of the flip-flop by setting both inputs to zero. Since the output of a NAND gate is 1 if either of its inputs is zero, both outputs must be 1 as we see in Figure 16-2. In this state all the gates have outputs that agree with their inputs. This is our first self-consistent state.

Now let input B go from 0 to 1 and follow what happens. This will take several stages because changing B alters output O1 and that, in turn affects output O0, which affects O1, and so on. We will follow what happens and show that we end up in a new self-consistent state.

Figure 16-3 shows the device in the instant after I changed input B but before the lower gate has had chance to change state. At this instant the lower gate has inputs that don't match its output. It will take some tiny amount of time, a few nanoSeconds, for the gate to make the change. For this tiny time the output of the system is 1,1 but the lower 1 will disappear as soon as the information propagates through the gate (about 23nS for a 74HC00 gate). I have marked this output with a '*' to emphasize that it is a temporary output.

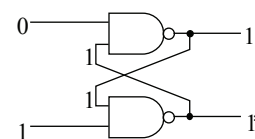


Figure 16-3

Since both inputs to the lower NAND gate are now at 1, that gate will switch to make O1 = 0 (Figure 16-4). Now the upper gate has two 0 inputs. Since the NAND output does the same thing whether it has one 0 input or two, the output does NOT change. We have reached a new self-consistent state.

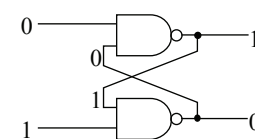


Figure 16-4

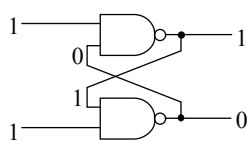


Figure 16-5

Normally, I would now go to the state $A = 1, B = 0$, but it will save time if we look at the state $A = 1, B = 1$ first (Figure 16-5). Interestingly, nothing changes when we go to this state. Output $O1$ is still at 0 so there is still a 0 input to the upper gate and $O1$ does not change. Clearly we can go back and forth between this state and the previous one as often as we like; nothing will happen.

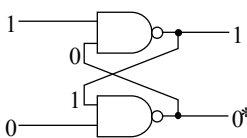


Figure 16-6

Now let us go to the last state, $A = 1, B = 0$. Immediately after the inputs change we are in another transient state (Figure 16-6).

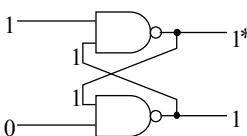


Figure 16-7

As the star suggests, that cannot last. The lower gate now changes state because it sees a zero at one of its inputs. That takes us to yet another transient state (Figure 16-7).

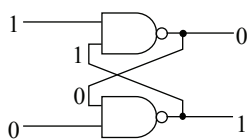


Figure 16-8

This time, the upper gate is unstable. It now has both inputs at 1 so its output switches to 0 and we have the state shown in Figure 16-8.

Now the system is stable again and the outputs are in a new state, $O0=0, O1=1$. So, we have looked at all four input states and found three different output states but no new behavior.

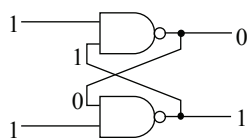


Figure 16-9

To see the new behavior we go back to the state $A = 1, B = 1$. This input state previously resulted in the output state $O0 = 1, O1 = 0$. This time we get to the state shown in Figure 16-9.

Interesting! We have a single input state, $A = 1, B = 1$, that can produce two *different* output states depending on how you got to the state.

If we go from input state $A, B = 1, 0$ to input state $A, B = 1, 1$ then the new output state is $O0, O1 = 0, 1$.

If we go from input state $A, B = 0, 1$ to input state $A, B = 1, 1$ then the new output state is $O0, O1 = 1, 0$.

We say that the output states $O0, O1 = 0, 1$ and $O0, O1 = 1, 0$ are **stable states** because once you have entered such a state it persists, even after the inputs have returned to $A, B = 1, 1$. Because the circuit has two stable states we call it a **bistable circuit**.

When the circuit sees inputs $A, B = 1, 1$ then the output circuit remembers the previous input state. Thus we call $A, B = 1, 1$ the **memory state**. This flip-flop is the basis of semiconductor memory.

Info In a physical circuit, one of the gates will switch slightly faster than the other and that difference will determine which state you end up in. You could build two apparently identical circuits that behave differently. This is clearly a bad idea and so we avoid this input state.

The third output state, $O0, O1 = 1, 1$, corresponding to the input $A, B = 0, 0$, is *not* stable and should never be used. If you go from input $A, B = 0$ to $A, B = 1, 1$, you have *no idea at all* what will happen to the output! The output will definitely go to one of the stable states but there is no way to predict which one it will be.

It is important to distinguish between two types of unstable state that we have encountered. The first type was a transient thing that went away by itself within a few nanoSeconds. It was fundamentally unstable; the output changed by itself while the inputs remained fixed. The second is the $O0, O1 = 1, 1$ state that we have just described. So long as its inputs remain fixed the output state will persist. The only problem with this state is that if we make a transition to the memory state ($A, B = 1, 1$) then you have no idea which of the two possible outputs will occur. We avoid this state as a matter of policy not because the state is in any fundamental sense invalid. A better name for this state would be the **undesirable** state.

16.2.1 Set, Reset, and the S-R Flip-Flop

If we omit the unstable state then the circuit has the property that its outputs are always complementary. To reflect that, we call the outputs not O0 and O1 but Q and \bar{Q} respectively. We then rename the inputs according to their effects on the Q output. When input A is a **zero**, it forces the Q output to a 1; it **sets** the Q output. Because of that, we call the A input \bar{S} ; S for Set and with the bar to indicate that it is negative true—a 0 sets the Q output. Similarly, a 0 on the B input forces the Q output to a 0; it **resets** the Q output. We therefore rename the B input \bar{R} . So our circuit, renamed the $\bar{S}\text{-}\bar{R}$ flip-flop or $\bar{S}\text{-}\bar{R}$ latch, now looks like Figure 16-10.

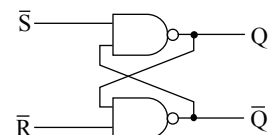


Figure 16-10 $\bar{S}\text{-}\bar{R}$ Flip-flop

We were able to describe our combinatorial circuits with truth tables showing the output for each allowed input but we can't do that for our bistable because a single input may have more than one output. Instead, we extend the truth table to the **state table**. A state table is more flexible. Table 16-2 is the state table for our S-R flip-flop.

Table 16-2:
 $\bar{S}\text{-}\bar{R}$ state table

\bar{S}	\bar{R}	Q	\bar{Q}
0	0	1*	1*
0	1	1	0
1	0	0	1
1	1	Q	\bar{Q}

Most of the table looks just like a truth table but the bottom line is different. Here we see that the Q output for input 1,1 is shown as 'Q'. That means that the output remains unchanged when we go to this state. If Q was 0 with the previous input, then it will stay 0 and similarly for the \bar{Q} output. With this notation, we can express the new output-state in terms of the previous state.

Note * This state is undesirable and should not be used.

Switch debouncing with the S-R Flip-Flop.

The S-R flip-flop is very useful for curing an annoying problem with mechanical switches. Switches usually consist of various springy bits of wire moved around by levers and the springiness leads to the problem of **contact bounce**. When a switch is moved from one position to the other, the connection is not made or broken smoothly. Instead, the bits of metal bounce around for a short time (usually milliseconds) before settling into the new state. So, if we connect up a simple single-pole single-throw switch and a resistor, it may produce the logic signal shown in Figure 16-11

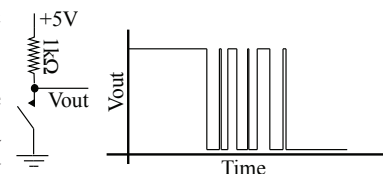


Figure 16-11 Noise from a switch

When the switch is closed to bring the output from +5V down to 0V, the contacts will bounce against each other for a few milliseconds. The output will not be a clean 1→0 transition but a messy thing like the one shown. It doesn't even take on only the values 0 and +5 because stray capacitance in the circuit forms an RC time constant with the 1k resistor. You may get all sorts of bumps and spikes until the metal stops bouncing and the output settles down to a clean 0.

Info These data were copied from the oscilloscope screen and are typical of the 10-20 switch throws that I tried. I used a small single-pole toggle switch rated at 125V and the bounces lasted anywhere from 2-15mS.

There are many situations where a noisy signal like this would be a major problem. Think, for example, of using this as the input to a circuit that counted pulses. You would have no idea how many counts you would get for a single flip of the switch. When a signal must make clean transitions, a noisy signal like this must be debounced. The usual way to do this is with a single-pole double-throw switch and an S-R flip-flop as in Figure 16-12.

When the switch is thrown, the contacts bounce as usual but the bouncing is only between the moving element and the contact to which it has moved. The moving contact does not bounce between the two fixed contacts because they are far too far apart.

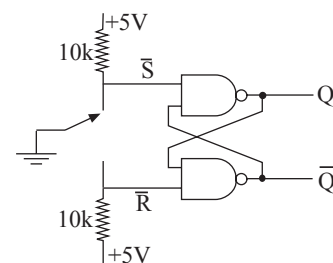


Figure 16-12 Switch De-bouncer

Let us follow what happens as we move the switch. At first, the switch is holding the S input of the flip-flop low so the Q output is high. When we move the switch, the S input goes high as soon as the contact is broken. Nothing happens to the output as the inputs are now in the memory state. Very soon, the moving contact hits the lower fixed contact and brings the R input low. This forces the flip-flop to change state and Q goes to 0. Now the contact bounces, allowing the R input to go high again. A high input on R takes the flip-flop back to the memory state and again nothing alters; the output stays at Q = 0. The contact may bounce for while, taking the R input between 0 and 1 but not altering the Q output. The output will not change until the switch is moved all the way back to the up state. Thus each throw of the switch alters the output exactly once, as we want.

16.3 The transparent latch

Now our bistable is not quite the memory circuit that we seek. It does have a simple sort of memory but we would like more control over when it remembers and when it learns. If we add two more NAND gates and a NOT gate then we have a real memory circuit as shown in Figure 16-13.

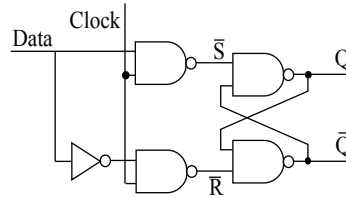


Figure 16-13 The transparent latch

Table 16-3:
State table for
Transparent Latch.

Clock	Data	Q	Q̄
0	0	Q	Q̄
0	1	Q	Q̄
1	0	0	1
1	1	1	0

Table 16-3 shows the state table for this circuit. When the clock input is high, the Q output follows the data input. When the clock is low, the latch remains in the state that it was in at the instant when the clock went low. Because the data flows straight through the latch when the clock is high, this circuit is called a **transparent latch**.

It is easy to understand how this circuit works. When the clock is low, both of the input NAND gates have high outputs so the S-R flip-flop is in its memory state. The Data line is logically disconnected from the circuit and cannot affect the outputs. When the clock is high, the R-bar input is equal to Data, while the S-bar input is equal to Data-bar. Thus when Data is high, it forces S-bar low and sets the Q output. When Data is low, it forces R-bar low and resets the Q output.

It is very common to have extra inputs to such a flip-flop. These inputs force the output to the set or reset state regardless of the clock and data inputs. Such inputs that ignore the clock are called **jam inputs**. They are usually provided so that the circuit can be forced to a known state when necessary, usually when the equipment is first turned on. We can add jam inputs to our circuit by adding more inputs to the flip-flop gates (Figure 16-14).

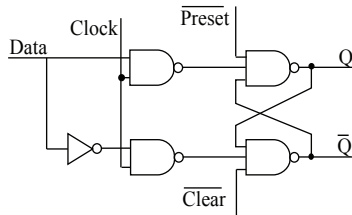


Figure 16-14 Latch with Preset and Clear

These extra inputs are given new names according to their functions.

Preset: a 0 on this line forces the Q output to 1 regardless of other inputs.

Clear: a 0 on this line forces Q-bar to 1 regardless of other inputs and usually forces Q to 0.

Obviously, you should never activate both Preset and Clear at the same time!

16.4 Logic diagrams for flip-flops

Circuit diagrams would get very complicated if every gate in every flip-flop had to be drawn out. Just as we adopted gate symbols to avoid having to draw the individual transistors from which each gate was built, so we adopt special symbols for complete flip-flops. The basic element is the rectangular box with a central clock input on the left and Q and Q-bar outputs on the right. The S-R flip-flop has inputs on the left opposite the outputs and the S-bar-R flip-flop adds little inverter circles to these inputs to indicate the inversion of the inputs.

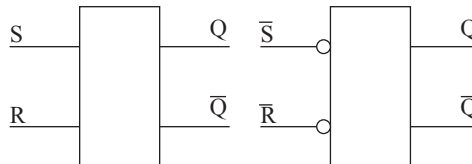


Figure 16-15 Flip-flop symbols

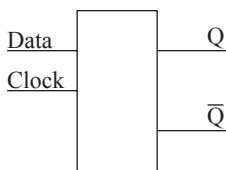


Figure 16-16 Clocked Latch

A clocked latch has its data input drawn opposite the Q output to show that it is the Q output that follows the data input (Figure 16-16). The clock is always drawn in the middle of the input side of the symbol to show that it affects both Q and Q outputs equally.

If the flip-flop has jam preset or clear inputs then they are put on the ends of the rectangle to show that they affect the outputs more directly than the ordinary inputs do.

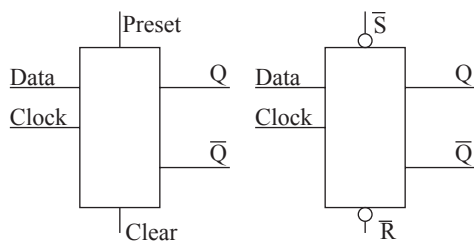


Figure 16-17 Flip-flops with jam inputs

The input on the end of the symbol closest to the Q output is always the set or preset input; a 1 here forces the Q output to 1.

The input on the other end is the clear input. A 1 on the clear input forces the Q input to 0.

If the inputs are inverted, as they were in the last example of the previous section, then that is shown with little inverter circles. So we get symbols such as those in Figure 16-17.

Note The terms **set** and **preset** are equivalent, as are the terms **reset** and **clear**.
 Set = preset = make $Q \rightarrow 1$.
 Reset = clear = make $Q \rightarrow 0$.

16.5 D-type flip-flops

There are many circuits in which it is important that the state of the latch depend only on the value of its input at a single instant determined by the clock. In this case, we need a special kind of latch called a **D-type** latch. The transparent latch is not a D-type latch; when the clock input is high the output follows the input. Let us look at the state table for a common kind of D-type latch (Table 16-4).

Table 16-4: D-type Edge-Triggered Flip-Flop

Clock	D	Q	Q
0	0	Q	Q
0	1	Q	Q
1	0	Q	Q
1	1	Q	Q
↓	0	0	1
↑	1	1	0

As you can see, the output stays fixed regardless of the level of the clock and data inputs. However, when the clock undergoes a transition from a low state to a high state, then the output changes state to follow the D input. We would call this a **positive edge-triggered** flip-flop. D-type flip-flops are available in both positive and negative edge triggered forms.

There are several ways to make such a latch by putting several S-R flip-flops together. We shall explore one method in detail below but there are other circuits that serve the same purpose inside modern integrated circuits.

16.5.1 Master-slave flip-flop

By driving two transparent latches from a single clock, we can make a latch that remembers the state of the input at the instant that the clock input goes from high to low. One latch, called the **master**, serves as the input for a second latch, called the **slave**, as in Figure 16-18.

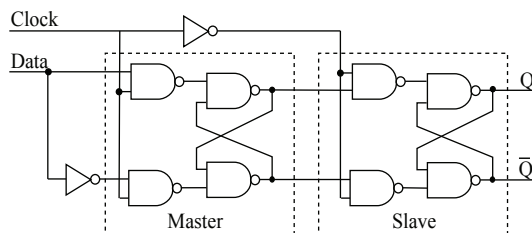


Figure 16-18 Master-Slave Flip-Flop

I have put individual Master and Slave flip-flops in boxes to make the construction clearer. Notice the NOT gate in the Clock line. It is the key to the behavior of the whole circuit. As usual, I shall demonstrate the behavior of the circuit by following bits through the gates. I shall start when $D = 0$, $Q = 0$, and the clock is low.

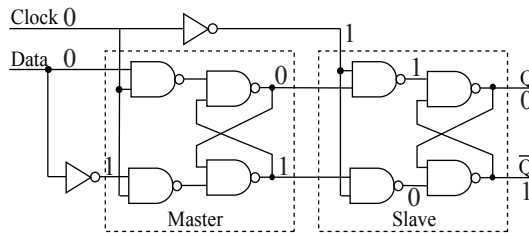


Figure 16-19

Because the clock input is 0, the master latch is in its memory state but the slave latch is in its transparent state; its clock is 1. While the Clock is low, the Data line can change as often as it wants and nothing can affect the output because the Master flip-flop is closed. Thus, when the Data line goes to 1 we have

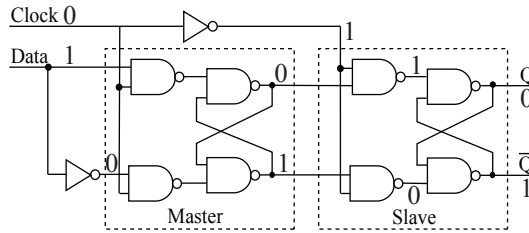


Figure 16-20

Now, while Data is still 1, I let the Clock input go high. Since the Master latch is transparent, the input to the slave changes from 01 to 10. The slave does not respond because it is now latched and there is still no change in the output. Again, the Data input can change as often as it likes without affecting the output because the slave latch is latched.

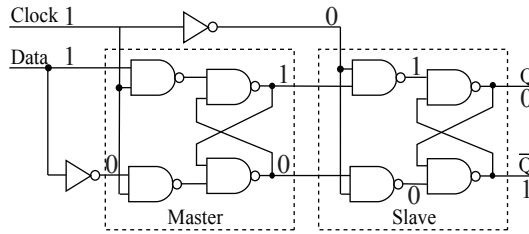


Figure 16-21

Lastly, I let the Clock return to 0 while Data is high.

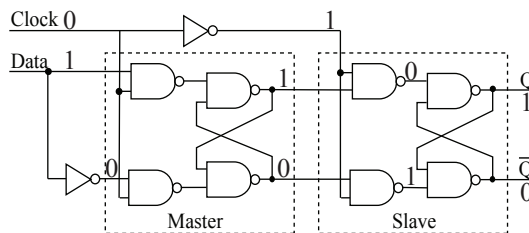


Figure 16-22

As the Clock falls, the Master latches the value on the Data input. Now the Slave clock goes to 1 so the slave opens and the output becomes equal to the value latched in the Master. Further changes are again impossible because the Master is now latched. So this flip-flop transfers the value of the Data input at the instant that the Clock goes from high to low to its output and then holds it there. Although the Data input is only sampled at the moment that the Clock falls, the Clock must make a whole 0-1-0 pulse in order to make the transfer.

We show this behavior in a state table by drawing a little picture of what the Clock must do to activate the flip-flop. Table 16-5 is the state table for our D-type master-slave flip-flop.

The first four rows tell us that nothing happens so long as the inputs are stable. The last two rows show us that the action takes place when the clock goes through a full pulse. The little

Table 16-5: State table for D-type Master-Slave Flip-Flop

Clock	D	Q	\bar{Q}
0	0	Q	\bar{Q}
0	1	Q	\bar{Q}
1	0	Q	\bar{Q}
1	1	Q	\bar{Q}
$\downarrow\uparrow$	0	0	1
$\downarrow\uparrow$	1	1	0

arrow on the trailing edge of the pulse tells us that that is the key edge. The value of the D input has at the instant of the trailing edge of the Clock pulse is transferred to the output.

16.5.2 D-Type Flip-Flop Symbols

All D-type flip-flops, no matter what their internal construction, share the same symbol (Figure 16-23). This is just like the clocked latch with the addition of the little wedge on the clock input. That wedge shows that the output changes only a rising edge. Obviously, if you have a flip-flop that operates on the falling edge you simply put a little inverter circle on the input.

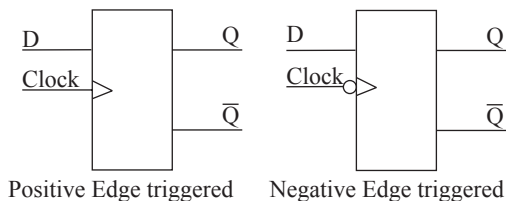


Figure 16-23 Edge-Triggered Flip-Flops

16.6 The J-K flip-flop

The most complicated and versatile flip-flop is the J-K flip-flop. It has two inputs, J and K, instead of only one. This extra complexity allows it to show more behaviors. J-K flip-flops are available as both level-triggered master-slave flip-flops and as true edge-triggered flip-flops. The older master-slave flip-flops have some nasty problems that make them tricky to use. I strongly advise you to avoid the older models and I shall discuss only the edge-triggered J-K flip-flops. Table 16-6 shows the interesting portion of the state table of an edge-triggered J-K flip-flop.

Table 16-6: Edge-Triggered J-K Flip-Flop

J	K	Q_{n+1}
0	0	Q_n
0	1	0
1	0	1
1	1	$\overline{Q_n}$

This is another new type of state table. Instead of having an output column named Q, it has a Q_{n+1} column. This shows what the value of the Q output will be after the next clock edge. There are some new symbols in that column.

An entry Q_n means that the Q value after the clock will be the same as it was before the clock.

An entry $\overline{Q_n}$ means that the Q value changes; if it was a 0 before the clock then it becomes a 1 and vice versa.

This is called **toggle** and you sometimes see the state table written with the word **toggle** in place of the Q_n entry.

The table shows that the J-K flip-flop can force its output to 1 or 0, in much the same way that the D-type flip-flop does. However it can also show other behaviors.

First, it can do nothing, ignoring the clock edge. Second, it can toggle as a result of a clock edge.

Commercial J-K flip-flops usually have more complicated state tables because they commonly have jam preset and clear inputs as well. For example, 16.7 is the symbol for the 74HC112 J-K master-slave flip-flop with Preset and Clear inputs.

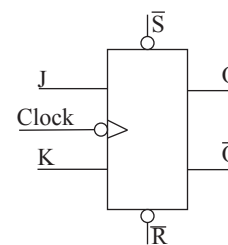


Figure 16-24 J-K Flip-flop with Preset and Clear

16.7 Simple counters

The simple memory circuits that I have described allow us to build whole new classes of logic circuits. One such new class is the **counter**; a circuit that responds to a clock input by stepping through a set of output states in numeric order.

The new toggle state of the J-K flip-flop allows to build our first simple counting circuit. Figure 16-25 shows a circuit that can up to four clock pulses before it has to start over.

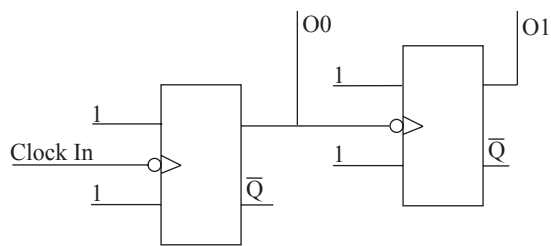


Figure 16-25 2-Bit Counter

Info We call the number of distinct states through which the counter steps the **modulus** of the counter. So our 2-bit counter is a modulo-four counter.

Because an n-bit binary number can have 2^n different states, it takes n bits to count up to 2^n and so our counter to four uses two flip-flops. The best way to see how this works is not to follow bits in our usual way but to draw a **timing diagram**; a diagram that shows the inputs and outputs as a function of time (Figure 16-26).

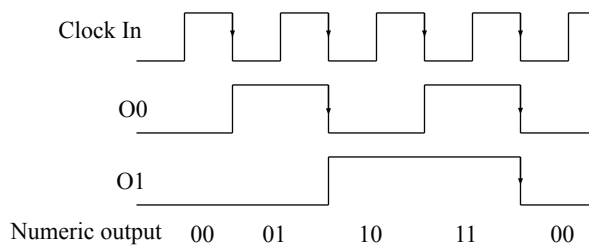


Figure 16-26 Timing Diagram for the 2-Bit Counter

The top line of the diagram shows the clock input. We create that so we can make it whatever we like. Here I have made it a series of identical pulses forming a 1:1 square wave. I assume that the flip-flops start in the $O0 = O1 = 0$ state and watch the outputs through five incoming pulses. Note that time proceeds from left to right as you would see it on an oscilloscope.

As expected, the output cycles through the four 2-bit numbers in the standard counting sequence. We can see how this happens. Each time the clock has a falling edge, O0 changes state because its J and K inputs are both 1. That makes O0 a square wave at one half the frequency of the input. Since O0 is the clock input for O1, O1 changes state every time O0 falls and the O1 signal is again a square wave at half the frequency of its clock input, O0. Because of the way the resulting square waves line up the two output bits march through the 4 possible states in numeric order.

This counting scheme can be extended to count up to any power of 2 that you like. Using three latches you can build a counter with 8 states that can count from 0 to 7. With 4 latches you have 16 states and can count from 0 to 15 and so on. This kind of counter, where the output of one stage is used as the clock input for the next stage, is called a **ripple counter** because of the way information ripples along from one stage to the next.

These counters have the property of producing a square wave output at an exact fraction of the input frequency. For example, the O0 output of the above circuit is at exactly half the frequency of the clock input and the O1 output is at exactly one quarter of the input frequency. Because of this, a counter is often called a **divide-by-n** circuit, where n is the modulus of the count. So, this circuit can be called a divide-by-four.

16.7.1 Other modulus counters

There are times when we don't want to count only powers of two. We would like to be able to build a counter that could count up to any limit, not just 2, 4, 8, 16, etc. For example, it is very convenient to be able to count in tens since then you can build a counter with a decimal output. This is done in frequency counters, devices that count the number of pulses arriving in one second and displaying the number as a frequency. Since the output is most convenient if it is in decimal, we need to be able to cascade a set of 0-9 counters rather than the binary 0-7 or 0-15 counters that we already know how to build.

We can alter a binary counter to any **lower** modulus by stopping the counter at just the right moment and starting again from 0. Here is a general algorithm for designing such counters.

Info This trick of resetting a counter to get a lower modulus is often called **short cycling** the counter.

Designing a Modulo-P Ripple Counter

- 1) Start with a generic n-bit ripple counter, choosing n so that 2^n is larger than the number of states that we want.
- 2) Add a reset circuit using a NAND gate to generate a 0 output when the counter reaches the value P.
- 3) Connect the output of this reset circuit to the Clear inputs of all the flip-flops.

Note There is a subtlety here. We want the counter to have P visible output states. They will be the states 0, 1, 2, ..., P-1. We want the next state to be 0 but we can't go to that state until we have briefly seen the state P. Thus, resetting the counter when it reaches P gives us a modulo-P counter whose largest visible output state is P-1.

Example

We can build a divide-by-three counter based on a divide-by-four counter. To get a divide by three we need to reset the counter as soon as its output reaches the fourth state so that the fourth state is forced to be the same as the first. Well, the fourth state would be 11 so if we reset the counter if the output gets to 11 then we will have a divide-by-three circuit. We use the Clear inputs of the J-K flip-flops to reset the counter to 00 as in Figure 16-27.

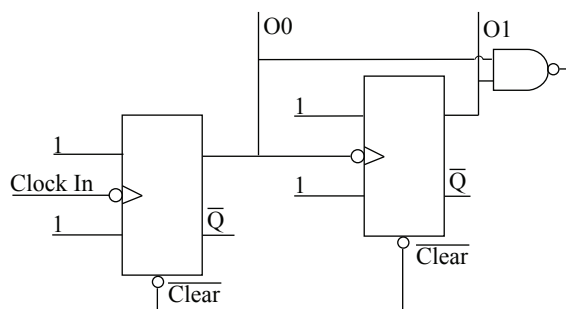


Figure 16-27 Divide-by-3

This circuit will now count 00, 01, 10, 00, 01, 10, 00 and so on. The output is not perfect, there is a little problem during the reset, but for many purposes it is quite adequate. The O1 output is a square wave at exactly one third of the input frequency. It is low two thirds of the time and high one third of the time.

16.7.2 Frequency Dividers

Short cycled ripple counters, like the one in Figure 16-27 above, go through their output states in strict numeric order. Often, this is exactly what is desired. Sometimes, however, we are not interested in the numbers on the n output lines. Instead we are interested only in the frequency of the slowest (most significant) output bit. In such a case we usually want the output to have a **duty cycle** that is as close to 1:1 as possible.

Info The **duty cycle** of a square wave signal is the ratio of the amount of time that it spends high to the amount of time that it spends low. A symmetric signal, a true square wave, spends exactly the same amount of time high as it does low. It has a 1:1 duty cycle. A duty cycle that is very far from 1:1 produces a signal that looks more like a set of short pulses separated by longer intervals.

For counters with even moduli, the output can always be arranged to have a 1:1 duty cycle while odd cases can usually come very close. The cost for this is that the outputs will not go through a counting sequence; instead they will go through some very different sequence of numbers. The general method of generating a 1:1 duty cycle with an even modulus $P = 2n$ is to build a divide-by-n and follow it by a divide-by-2. This way you get a total division of $2n$ and are certain that the output has a 1:1 duty cycle.

16.7.3 The decade counter or Divide-by-10

The obvious way to implement a decade counter is to short-cycle a 4-bit binary counter as shown in Figure 16-28.

The problem with this circuit is that the O3 output square wave does not have a 1:1 duty cycle. Look at the output from this circuit as shown in the timing diagram of Figure 16-29. The outputs go through a nice counting sequence, 0-9 and repeat, but the O3 output is low for 8 clocks and high for 2. That is hardly a 1:1 duty cycle.

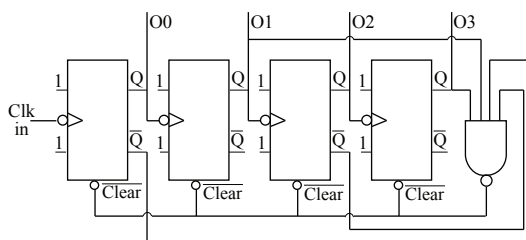


Figure 16-28 Simple Decade Counter

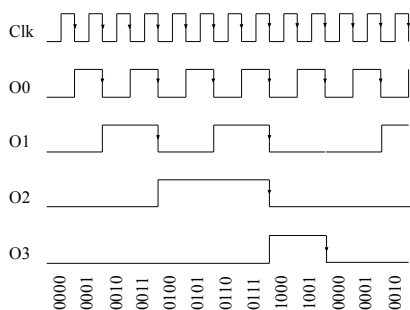


Figure 16-29 Timing Diagram for the Simple Decade Counter

Since 10 is an even number, we can fix the output by making sure that the last thing that happens is a divide-by-2. If you look at the output sequence above then you can see that the O0 output is just the input divided by 2. The next three bits then perform a divide-by-5 sequence. We can get a divide-by-10 with a 1:1 duty cycle output by doing the divide-by-5 and then the divide-by-2 instead of the other way round. Because of this, commercial divide-by-ten chips are often arranged with separate divide-by-2 and divide-by-5 circuits. If you want to go through a proper counting sequence then you connect them with the divide-by-2 first and follow it with the divide-by-5 (Figure 16-30).

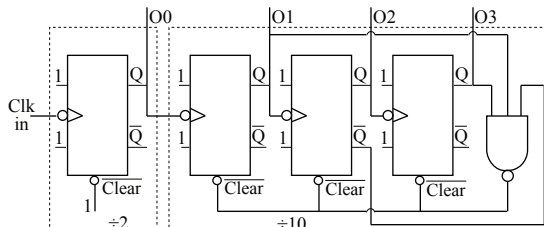


Figure 16-30 Improved Decade Counter

If you want a frequency divider, then you do the divide-by-5 first and then the divide-by-2 (Figure 16-31).

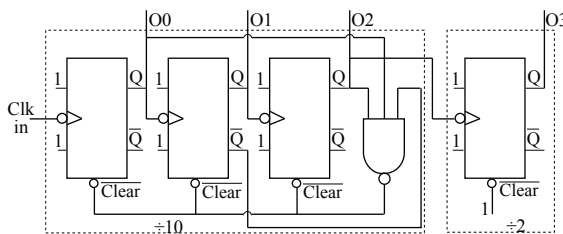


Figure 16-31 Improved Divide-by-10

The advantage of the Divide-by-10 configuration is shown in Figure 16-32. Clearly the output is nothing like a counting sequence but the O3 output is now a beautiful 1:1 square wave at 1/10th the input frequency.

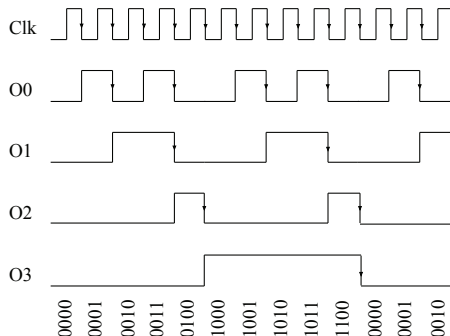


Figure 16-32 Timing diagram for Divide-by-10

16.8 Some other Chips

Here are some examples of the simple sequential logic functions available in our 74HCxx logic family. I am only going to show a few of the commonest and most useful devices. You should look at a CMOS databook to see the full range. Note, however, that many of the more interesting sequential circuits use synchronous logic and so are properly the topic of our next chapter.

16.8.1 Flip-flops

Today's CMOS logic family provides only a limited number of simple flip-flop circuits. The simplest is the 74HC75 4-bit latch, which puts four transparent latches into a single 16-pin package. There are not quite enough pins to make the latches independent so they are organized as two pairs of latches. Each pair of latches shares a single Clock input so that you cannot use one latch of a pair independently of the other.

The next most complex flip-flop is the D-type flip-flop. These come in several varieties. The 74HC74 contains a pair of D-type, positive edge-triggered flip-flops with $\overline{\text{Preset}}$ and $\overline{\text{Clear}}$ inputs. If you want more flip-flops in each package then there are its cousins, the 74HC174 Quad D-type flip-flop, the 74HC175 Hex D-type flip-flop, and the 20-pin 74HC273 Octal D-type flip-flop. Not surprisingly, the multiple flip-flop packages sacrifice the separate $\overline{\text{Preset}}$ and $\overline{\text{Clear}}$ inputs, opting instead for a single $\overline{\text{Clear}}$ input that clears all the flip-flops at once. There are some other forms that are called D-type latches but which are really multiple D-type flip-flops in a package but they are all similar to the '273.

When it comes to J-K flip-flops, there is a little more variety. There are dual negative edge-triggered J-K flip-flops available with $\overline{\text{Clear}}$ (74HC107), with $\overline{\text{Preset}}$ (74HC113), and with both $\overline{\text{Preset}}$ and $\overline{\text{Clear}}$ inputs (74HC112). However the dual positive edge-triggered J-K flip-flop is only made with $\overline{\text{Preset}}$ and $\overline{\text{Clear}}$ inputs (74HC109A). There are no multi-bit J-K flip-flops.

16.8.2 Counters

Most of the interesting counters use synchronous logic and so must wait till Chapter 17. Earlier logic families offered a wide selection of simple counters but the improved performance of synchronous counters comes at no cost in today's high performance logic families and so these devices are no longer available. The last simple counters are the 74HC390 and 74HC393, which pack two 4-bit counters into a single package. The '393 is a simple binary counter that can count from 0-15. The '390 is a decade counter of the kind described in the previous example. It has separate divide-by-2 and divide-by-5 counters that can be arranged to produce either a counting sequence or a frequency division.

There are a couple of more unusual counters in the extended range of 74HC series circuits. This is a small collection of functions with numbers of the form 74HC4xxx that are descended from a much older, and happily forgotten, 4000 family of CMOS logic. Most of these devices have vanished but there were a small number of rather unusual functions that have been retained and have been renumbered into the 74 family. In particular, there are two very large modulus counters, the 74HC4020 and 74HC4040. The first is a divide by $2^{12} = 4096$ that provides all powers of two from $\div 2$ to $\div 4096$. The 74HC4040 takes this even further, providing division up to $2^{14} = 16,384$. Because it has to fit in the same size package as the 12-bit '4020, the $\div 4$ and $\div 8$ outputs are omitted, but all other outputs are present.

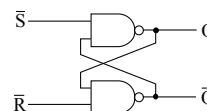
Summary

The output of a Sequential Logic circuit depends upon both the current values of its input and the previous history of those inputs. Sequential logic circuits have some form of memory that allows their current state to depend on their previous states.

The simplest sequential logic circuit is the latch with structure shown on the left and the state table on the next page.

Note the similarity in device numbers. This is not accidental. The first 100 device numbers (7400-7499) were allocated first and then, as more advanced ICs were developed, they were given numbers that resembled their low numbered cousins. So a number of the form 74x7y, where y is near 4, is likely to be some kind of D-type flip-flop.

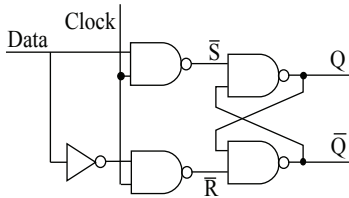
Note The $\div 4$ and $\div 8$ signals are present internally. They are simply not connected to output pins. They must be present since they are needed to generate the higher powers.



A logic 0 on either input will force the corresponding output to a 1. When both inputs are 1 the Q output remembers which input was last 0. This is the memory state.

$\bar{S}\bar{R}$ state table

\bar{S}	\bar{R}	Q	\bar{Q}
0	0	1*	1*
0	1	1	0
1	0	0	1
1	1	Q	\bar{Q}



Adding two NAND gates makes the simple memory cell or Latch on the left

When the clock is high Q = Data. As shown in the table below, when the clock is low the Q holds (remembers) the value of D at the moment Clock went low.

Transparent Latch State Table.

Clock	Data	Q	\bar{Q}
0	0	Q	\bar{Q}
0	1	Q	\bar{Q}
1	0	0	1
1	1	1	0

The transparent latch is sometimes inconvenient because the output can change freely when the clock is high. More sophisticated circuits allow the output to change only at closely controlled instants. The most usual is the edge-triggered flip-flop in which the output changes only at the moment the clock goes from a 0-1 (positive edge triggered) or from a 1-0 (negative edge triggered). The two most common edge triggered flip-flops are the D-type and J-K flip-flops. The D-type is quite straightforward.

D-type Latch State Table

Clock	D	Q	\bar{Q}
0	0	Q	\bar{Q}
0	1	Q	\bar{Q}
1	0	0	1
1	1	1	0
\downarrow	0	0	1
\downarrow	1	1	0

The J-K needs a new symbol in the state table. Q_{n+1} is output after clock edge given J and K before clock edge.

J-K Latch State Table

J	K	Q_{n+1}
0	0	Q_n
0	1	0
1	0	1
1	1	\bar{Q}_n

If we connect a series of flip-flops together then we can construct circuits that can count in binary. A circuit built with n flip-flops can count from 0 to 2^n-1 . With 2 flip-flops we can build a 2-bit counter.

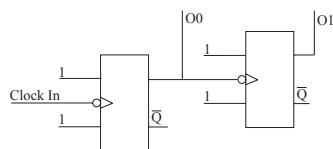


Figure 16-33 Divide-by-4

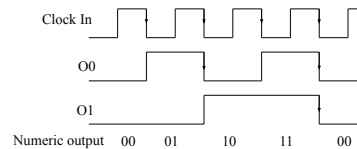


Figure 16-34 Divide-by-4 Timing

This counts through the states 0, 1, 2, and 3 and is often called a divide-by-4 since its O1 output is at 1/4 the frequency of its clock input.

By adding extra logic we can cut the count short at any lower number. One common counter is the divide-by-ten.

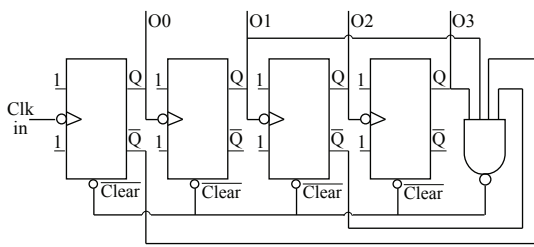


Figure 16-35 Divide-by-10

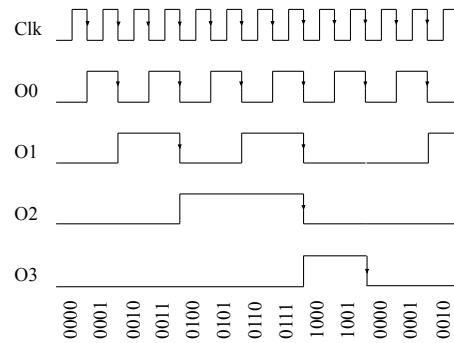
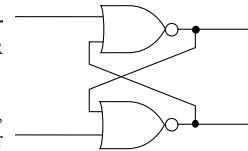


Figure 16-36 Divide-by-10 Timing

The NAND gate generates a reset pulse when the output reaches 0b1010 (decimal 10) and then the counter starts over from zero.

Exercises

1. Investigate the behavior of the latching circuit of the figure shown at the right. Explain its relationship to the S-R latch and suggest names for the circuit and its inputs/outputs (consult 16.2.1 for this last part).
2. Design a Modulo-7 ripple counter. This must count through the sequence of states 000, 001, 010, 011, 100, 101, 110, 000, 001, ... (Note that there will be a very tiny 111 state that we shall not learn how to remove until chapter 17.)
3. Why is it impossible to write down the logic equations for the circuit of Figure 16-1?



Chapter 17:Synchronous Logic

17.1 Introduction

All the larger sequential circuits we have seen so far suffer from a problem that we have not yet discussed. We have built circuits that can count to any number and in any base but if we examine the output of those circuits in detail we find that there are problems. There are small intervals of time when the output of the circuit is simply wrong. This chapter describes the problem and its source and develops techniques to cure the problem completely. It then shows how those techniques can be used to build sequential circuits to solve any possible kind of problem. They can be used to design a computer.

17.2 Glitches

We shall start by re-examining the simple divide-by-three counter that we developed in Chapter 16. Figure 17-1 shows the circuit diagram again.

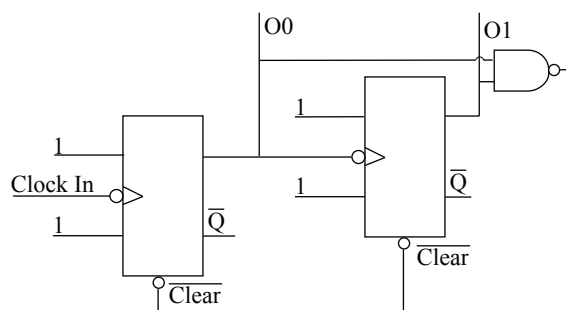


Figure 17-1 Divide-by-3 Ripple Counter

There are two problems with this circuit that only emerge if we look very carefully at what happens when a falling clock edge causes the O0 flip-flop to make a transition. The more obvious problem is seen when we encounter the third falling clock edge in a row (Figure 17-2). After this transition, the two flip-flops must go into the 1,1 state before the NAND gate can reset the counter. That means that the counter does not really count 0, 1, 2, 0, 1, 2, 0, ... as we wanted. Instead it counts 0, 1, 2, 3 woops, 0, 1, 2, 3 woops, 0, and so on. This very brief woops state is called a **glitch** and it is present in all short-cycle ripple counters, that is counters that rely on a reset to stop after a certain number of counts.

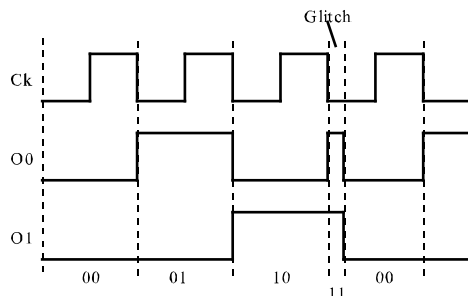


Figure 17-2 Glitch During Counter Reset

The second problem is more subtle and even more widespread. Look at Figure 17-3 on the next page, which is an enlargement of the events in the first few tens of nanoseconds following a falling clock edge.

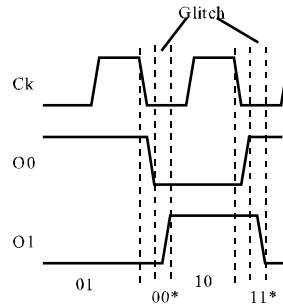


Figure 17-3 Glitches Due to Clock Ripple

After the clock edge has fallen, it takes some time for the output of the latch to change state, about 20nS for the 74HC112 J-K flip-flop. The output from that latch is used as the clock of the second latch so that it takes another 20nS or so for the second latch to change state. During the 20nS between the transition of the first latch and that of the second latch, the output state is incorrect!

Example

If we start in the state 01 (Figure 17-3) then, after the first latch has switched and before the second one has the output of the system is 00 not the 10 that it should be. In fact, if we include all the glitch states, marked with stars, then the counter does not go through the sequence 00, 01, 10, 00, 01, 10, etc. but through the very different sequence 00, 01, 00*, 10, 11*, 00, 01, 00*, etc.

Every circuit that uses the output of one stage to clock a later stage is subject to this problem.

17.2.1 A Divide-by-3 Without Glitches

The cure for these glitch problems is to clock all of the flip-flops with the same signal. Then we have to find some other way of making sure that the counter goes through the correct sequence since a collection of J-K flip-flops, all with J,K = 1,1 inputs and driven by the same clock, will all change state together. Such a circuit would generate outputs such as 000, 111, 000, 111, 000, etc. We can alter the sequence by manipulating the J and K inputs according to the current state of the whole counter. If we choose the J and K inputs correctly then we can make sure that outputs change when they are supposed to and stay put when they are supposed to. For example, look at the circuit of Figure 17-4.

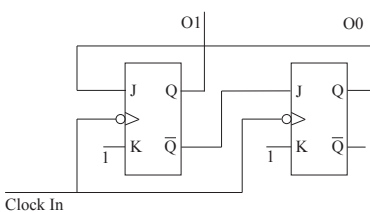


Figure 17-4 Synchronous Counter

As usual we will figure out what the circuit does by following it through a sequence of states. There are so many outputs and inputs here that bit following is easier than trying to go straight to a plot of the output. Let us start the system in the state 00 and watch what happens as the clock edges arrive.

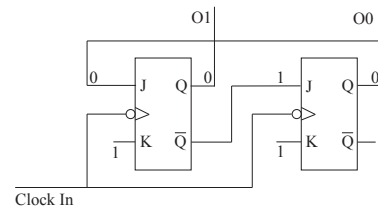


Figure 17-5

Before the first clock edge (Figure 17-5), flip-flop 0 has inputs J,K = 1,1 so that it will toggle when the clock arrives. Flip-flop 1 has inputs J,K = 0,1 so that it will be forced to the state O1 = 1 after the clock edge.

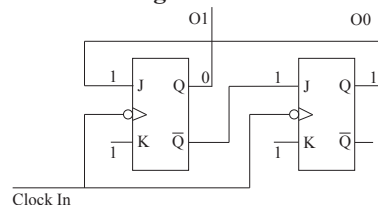


Figure 17-6

After the first clock edge arrives (Figure 17-6), we are in the state 01. Now the inputs to the flip-flops have changed. Flip-flop 0 now has J,K = 1,1 and is ready to toggle. Flip-flop 1 also has inputs J,K = 1,1 and is also ready to toggle

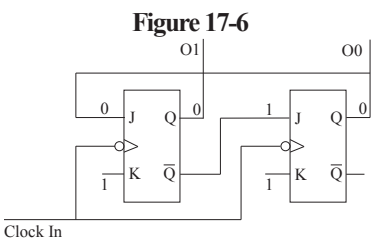


Figure 17-7

After the second clock edge (Figure 17-7), both flip-flops have changed state so we are now in the state 10. The new outputs create new input values. This time both flip-flops have inputs J,K = 0,1 so that both flip-flops will be forced to the 0 state after the next edge.

After the third edge, we are back where we started from (Figure 17-8). We have been through the cycle 00, 01, 10, 00 without meeting any glitch states on the way. This circuit is called a **synchronous** divide-by-three because both flip-flops are always clocked at the same time. That is, they are clocked **synchronously**.

Note There is still the possibility of extremely brief glitch states since we cannot guarantee that both flip-flops will take exactly the same amount of time to respond to the clock. Any difference in the switching time between two nominally identical flip-flops is called **clock skew**. Fortunately, while the glitch states in the ripple counter lasted for as long as it took one flip-flop to switch, 28nS for a 74HC74 D-type latch, those in the synchronous counter last only for the clock skew time, about 1-2nS. There are no design tricks to eliminate this kind of glitch but they are so short that they cause problems in only the fastest of circuits. The only cure is to use faster flip-flops and to make sure that the leads to different clocks are all the same length.

17.3 General Synchronous Systems

The general cure for the glitch problems of ripple counters is to clock each one of the flip-flops with the same clock signal. Once all of the flip-flops are driven by the same clock they will all change state at the same time and the glitches due to reset states and to delayed clocks will be eliminated. We call such a circuit a **synchronous circuit**.

The basic structure of all synchronous devices is the same. They consist of a set of flip-flops to store the current state of the system and collection of combinatorial logic that makes sure that the system goes through its states in the correct order (Figure 17-9).

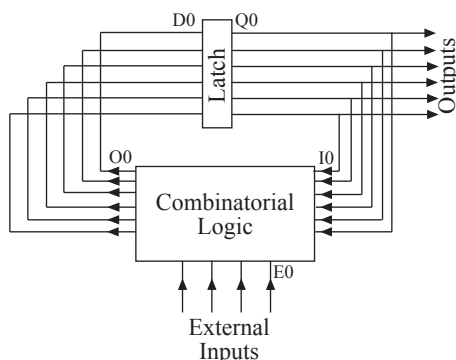


Figure 17-8 General Synchronous Circuit

The latch consists of a set of flip-flops (usually D-type) all driven by a common clock signal. The outputs of the latch are taken to the inputs of a combinatorial circuit that may also have external inputs. The combinatorial circuit uses the values of all its inputs to compute a set of outputs that are connected to the latch inputs. This feedback makes talking about the system a little difficult. I have tried to clarify things by labelling the different points in the circuit. I call the latches outputs Q , and its inputs D while I call the combinatorial circuits inputs I or E and its outputs O . In reality the Q 's and the I 's are the same signals as are the O 's and the D 's but we need to keep them logically separate.

The latch stores the instantaneous **state** of the system as a binary number on the individual Q outputs. When a clock pulse arrives at the latches the outputs all change at the same time and the system enters a new state. Which state it enters is completely determined by the values present on the inputs to the latch just before the clock pulse. Those inputs come from the combinatorial logic and their values are determined by the current state (the set of Q 's) and the external inputs. Each different combinatorial circuit produces a different sequence of states and so produces a different synchronous system.

Synchronous systems are far more general than the modulo- N counters that we have seen so far. In addition to various kinds of counter we can build synchronous circuits to control many kinds of sequential task. Some familiar examples include controllers for traffic lights and elevators. A traffic light controller must not only make sure that the lights work their way through the correct color sequence but may also take into account inputs from pedestrian buttons and from sensors buried in the road that can tell the system when cars are waiting at the light. The ultimate example of a synchronous system is a digital computer. Here the external inputs include the instructions stored in memory and so the memory instructions control the operations of the computer.

Info Synchronous means "occurring at the same time".

Note The figure shows a system with 6 latches, 6 outputs, and 4 external inputs. This is just an example. A general system may have any number of latches and outputs and may have several external inputs or none at all.

The figure also shows the outputs from the circuit coming from the Q outputs of the latches. This is by far the most common situation and is called a Moore machine. It is also possible to take the outputs from the combinatorial circuit and that is called a Mealy machine. The Mealy machine can handle more complex tasks with fewer gates than the Moore but is harder to design.

17.3.1 Describing Synchronous Systems

We describe combinatorial circuits with Truth Tables and sequential circuits with State Tables. Synchronous systems are special cases of sequential systems and so can be described by State Tables. However, the state tables can get very large and hard to understand. We often need a more convenient way of describing large synchronous systems. One common method is the **State Diagram**.

A state diagram is a graphical picture of the operation of a synchronous system that is somewhat similar to the flow charts often used to describe computer programs. In this, we draw each of the states of the system in a little bubble and then we draw arrows between states showing the transitions that are possible. We label the bubbles with the names of the states and we label the arrows with the conditions that cause the transitions to be taken.

It is normal practice to label each transition with the names of those external inputs that are true (1) and leave off those inputs that are false (0). Any transition that happens when all external inputs are false is left unlabelled. Obviously, if there are no external inputs then the transitions are left unlabelled.

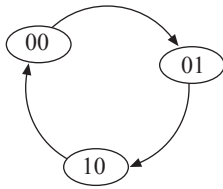


Figure 17-9 Divide-by-3 State Diagram

Example

Figure 17-9 is the state diagram for our synchronous divide-by-3 counter.

The arrows make it clear that the transitions are strictly one-way. The circuit always goes from state 00 to state 01 and never the reverse. A more elaborate synchronous system would have external inputs that control its behavior.

Example

We can design a more advanced 2-bit counter. We shall add an external input Up that controls the direction in which the counter counts and an input Reset that forces the counter to the 00 state. Here is the state diagram for this more advanced counter.

When the external inputs are both 0 the counter cycles round anti-clockwise, counting down. When Up is set the counter cycles round clockwise, counting up. When Reset is 1 the counter is forced into the 00 state. What is not clear from the diagram is what happens when both Reset and up are set. In that case Reset should win and so the lines labelled Up should be labelled Up-Reset but I left that off to make the diagram clearer.

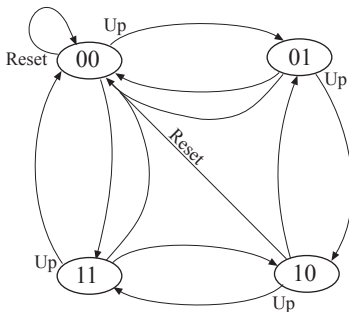


Figure 17-10 2-bit Up-Down counter with Synchronous Reset

Once you have created a state diagram it is fairly simple to design the combinatorial logic circuit that will implement a given synchronous machine.

17.3.2 Designing synchronous circuits with D-type flip-flops.

While it is possible to create a synchronous system using J-K flip-flops, as in Synchronous Counter, it is far more common to use a set of D-type flip-flops. We design a synchronous system based on D-type flip-flops in a series of stages.

1. Decide how many flip-flops make up the system. Since a set of n binary bits can represent any of 2^n unique states, a system that has anywhere between $2^{m-1}+1$ and 2^m states requires m flip-flops.
2. Draw the state diagram for the system.
3. Compute the values of the D inputs for each output state.
4. Write a truth table to generate the D values from the outputs.
5. Convert the truth table into a combinatorial circuit using standard methods.

I usually start step 4 by creating an empty table with headings like Table 17-1 on the next page.

Table 17-1:

Q2	Q1	Q0	D2	D1	D0
I2	I1	I0	O2	O1	O0

This emphasises the fact that each Q output becomes the corresponding I input and each O output is connected to the corresponding D input.

Next I fill in the Q values with the sequence through which I want the system to run. If there are any external inputs then I will need to write out a different sequence of Q's for each different set of external inputs.

Finally, I can complete the table by filling in the D/O side of the table. This is made very simple by the way in which a D-type flip-flop operates. When a clock pulse arrives the flip-flop copies the state of its D input and that becomes the new Q. Thus the D/O side of the table is filled in with the Q patterns in a different order. An example should clarify this.

Example

Divide-by-3 counter using D-type flip flops

We start by deciding how many flip-flops we need to implement the system. Since it has to have three unique states we need two flip-flops (capable of handling up to 4 unique states).

Next, we write down the state diagram (Figure 17-11).

If the circuit is to go through that sequence of states, then when the counter is in the state 0,0, the next state must be 0,1, and so the D-inputs must also be 0,1. Similarly, if we are in state 0,1 then the D-inputs must be 1,0, the outputs for the next state.

When we put all this together, we get Table 17-2.

We fill in the truth table using the top set of column headings, Q and D. Now we design the combinatorial circuit using the bottom set of headings, I and O. In terms of these headings we can write down logic expressions for the two output columns using our usual methods. In this case we find that

$$O1 = I0$$

$$O0 = I1 + I0 \text{ or } O0 = I1 \oplus I0$$

where we have a choice of expressions for O0 because we have not specified what will happen if the counter is in state 1,1. Either of these expressions will give us a counter that operates correctly, assuming that the counter starts off in a legal state. The two counters differ in their behavior if something goes wrong.

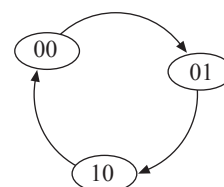


Figure 17-11 Divide-3 State diagram

Table 17-2:

Q1	Q0	D1	D0
I1	I0	O1	O0
0	0	0	1
0	1	1	0
1	0	0	0

Info When power is first applied you have no idea which state any flip-flop will be in. The state depends on which gate in the flip-flop happens to switch faster and force the whole latch into a defined state. This makes the starting state depend on random manufacturing differences between one gate and the next on the same chip. Two flip-flops with the same part number, even two flip-flops on the same chip, may start up in quite different states.

17.3.3 Excluded States in Synchronous Logic

The D-type divide-by-3 counter of the previous example illustrates a common problem with synchronous systems. In normal operation the counter should never be the state 1,1. We call this state an **excluded** state since it is excluded from the state diagram. In practice we cannot be sure that the state will never occur and that can lead to problems.

Let's look at what happens if the counter somehow gets into state 1,1.

1. If we choose the simpler expression, $O0 = I1 + I0$, then when the output state is 1,1 the inputs will be $I1 = 1, I0 = 0$. That means that the next state will be 10 and the counter is back into the normal cycle.
2. If we choose the more complex expression, $I1 \oplus I0$, then when the output state is 1,1 the inputs will be $I1 = 1, I0 = 1$. That means that the next state will also be 1,1. So, if the counter ever gets into this illegal state, then it is stuck there and will never emerge.

So the first version is much safer; it has a self-recovery mechanism. Therefore we will choose the first version and we can now design the circuit shown in Figure 17-12.

I have drawn this circuit to emphasize how the circuit fits into the general pattern of a synchronous device. I have enclosed the latch section and combinatorial section in dotted boxes to make them stand out. You would not normally draw the circuit this way but would lay it out however was convenient.

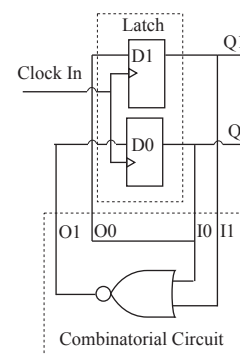


Figure 17-12 Synchronous Divide-by-3 Circuit

17.3.4 External Inputs

All of the synchronous systems that we have seen so far operate independently of the outside world, apart from their clock signal. In practice we often want the behavior of the system to depend on events in the real world and so must supply extra inputs, external inputs, to the combinatorial logic piece of the synchronous system.

Let us look at a more complex example to see how to include the effects of external inputs. We have already seen the state diagram for a 2-bit up/down counter with synchronous reset (2-bit

Up-Down counter with Synchronous Reset). Let us construct the truth table for this system. It will need 2 external input columns, two Q/I columns and two D/O columns. Let us fill in the simple case Up = 0, Reset = 0, first.

Table 17-3:

External		Q1	Q0	D1	D0
Reset	Up	I1	I0	O1	O0
0	0	0	0	1	1
0	0	1	1	1	0
0	0	1	0	0	1
0	0	0	1	0	0

When both Up and Reset are false, the counter is a normal 2-bit down counter.

Next we shall deal with the case Up = 1, Reset = 0. In this case the counter functions as a normal 2-bit up counter.

Table 17-4:

External		Q1	Q0	D1	D0
Reset	Up	I1	I0	O1	O0
0	1	0	0	0	1
0	1	0	1	1	0
0	1	1	0	1	1
0	1	1	1	0	0

Then we have to deal with Reset = 1. When Reset is true the counter is forced back to the 0,0 state regardless of the value of Up.

We can show that in the table in the usual way by putting in X's for values that are don't cares.

Table 17-5:

External		Q1	Q0	D1	D0
Reset	Up	I1	I0	O1	O0
1	X	X	X	0	0

Now we can put the whole thing together to get

Table 17-6:

External		Q1	Q0	D1	D0
Reset	Up	I1	I0	O1	O0
0	0	0	0	1	1
0	0	1	1	1	0
0	0	1	0	0	1
0	0	0	1	0	0
0	1	0	0	0	1
0	1	0	1	1	0
0	1	1	0	1	1
0	1	1	1	0	0
1	X	X	X	0	0

Now we are ready to write the equations. Since there are many more 0's than 1's in the output columns I shall use sum-of-products to find the equations

$$O0 = \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot \overline{I1} \cdot \overline{I0} + \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot I1 \cdot \overline{I0} + \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot \overline{I1} \cdot I0 + \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot I1 \cdot I0$$

$$O1 = \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot \overline{I1} \cdot \overline{I0} + \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot I1 \cdot \overline{I0} + \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot \overline{I1} \cdot I0 + \overline{\text{Reset}} \cdot \overline{\text{Up}} \cdot I1 \cdot I0$$

With a little Boolean algebra we can simplify these equations somewhat to find

$$O0 = \overline{\text{Reset}} \cdot \overline{I0}$$

$$O1 = \overline{\text{Reset}} \cdot \overline{\text{Up}} \oplus (I1 \oplus I0)$$

These give us the circuit of Figure 17-13.

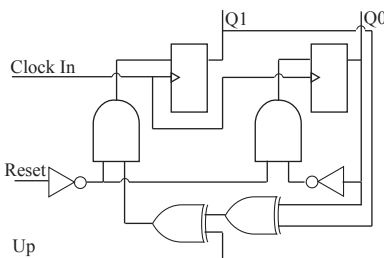


Figure 17-13 Synchronous Up-Down Counter

17.3.5 Resetting Synchronous Circuits

As we saw in section Excluded States in Synchronous Logic, there is a problem with most synchronous logic systems. States exist that are not normally encountered when the system is operating. They are called **forbidden states** or **excluded states** and the complete set of such states is called the **forbidden zone**. If the system gets into one of these states then one of two things can happen. The logic can push it back into a normal state or the system can get stuck in the forbidden zone. It often costs extra logic to make sure that the system gets itself out of the forbidden zone if an accident takes it there. The system designer has to decide whether the extra logic is worth it or not. Often the answer is not.

If the designer decides not to provide the extra logic to recover from the forbidden zone automatically, then she must make some provision for getting the system back into a standard state. This is usually done by providing a reset line that forces the system into some good starting state, often the all 0 state.

All flip-flops start up in random states and so you could be certain that your system would occasionally start up in a forbidden state and then you would be stuck. Indeed, the classic symptom of an unintended forbidden state is a circuit that works most of the time but sometimes won't work when it is first turned on. It can usually be made to work if you just turn it off and back on again but it is obviously better to design it correctly instead.

Remember All synchronous logic systems must either have a reset or be self-starting because you have no idea what state the latches will be when the power is turned on.

17.4 *J-K Synchronous

In practice almost all synchronous systems are built round D-type flip-flops but if you are trying to minimize the amount of external logic then you can sometimes save gates by using J-K flip-flops instead. Designing a synchronous circuit using J-K flip-flops follows a path that is similar in structure to that followed when designing with D-type flip-flops but is a little more complex in detail. The extra inputs and extra flexibility of the J-K flip-flop mean that the logic equations are a lot more open to choice as an example will show.

17.4.1 Divide-by-5 counter using J-K flip-flops

We will design a synchronous divide-by-5 circuit with J-K flip-flops. The counting sequence is 000, 001, 010, 011, 100, 000 etc. First we will remind ourselves of the J-K flip-flop's state table (Table 17-6).

In a J-K flip-flop, the output after the clock is not simply related to the input before it so we have a rather more complex design table. It needs three sets of columns instead of two. In the leftmost set we write down the desired values of the output bits in each state. In the rightmost set we write down the desired state of the output bits after the next clock edge. That gives us Table 17-7.

Table 17-7:

J	K	Q_{n+1}
0	0	Q_n
0	1	0
1	0	1
1	1	$\overline{Q_n}$

Table 17-8:

$O2_n$	$O1_n$	$O0_n$	J2	K2	J1	K1	J0	K0	$O2_{n+1}$	$O1_{n+1}$	$O0_{n+1}$
0	0	0							0	0	1
0	0	1							0	1	0
0	1	0							0	1	1
0	1	1							1	0	0
1	0	0							0	0	0

Next we work our way through the input entries one row at a time. In each case we use the J-K state table to decide what the values of J and K must be to produce the correct outputs in the next state. For example, flip-flop 2 starts in state 0 and goes to state 0. There are two ways this could happen. The inputs could be $J,K = 0,1$ to force the output to 0 or they could be $J,K = 0,0$ to allow the output to stay 0. Thus we write down $J,K = 0,x$, where the 'x' signifies that the K input can be either a 0 or a 1.

Table 17-9:

$O2_n$	$O1_n$	$O0_n$	J2	K2	J1	K1	J0	K0	$O2_{n+1}$	$O1_{n+1}$	$O0_{n+1}$
0	0	0	0	X					0	0	1
0	0	1							0	1	0
0	1	0							0	1	1
0	1	1							1	0	0
1	0	0							0	0	0

The next entries will be the same because flip-flop 1 also starts as 0 and stays 0. The third flip-flop however changes from 0 to 1. This could happen either because $J,K = 1,0$, in which case the output will be forced to 1, or because $J,K = 1,1$, in which case the output will toggle to 1. Thus we have $J,K = 1,x$ for this entry.

Table 17-10:

$O2_n$	$O1_n$	$O0_n$	J2	K2	J1	K1	J0	K0	$O2_{n+1}$	$O1_{n+1}$	$O0_{n+1}$
0	0	0	0	X	0	X	1	X	0	0	1
0	0	1							0	1	0
0	1	0							0	1	1
0	1	1							1	0	0
1	0	0							0	0	0

We can fill in the next two sets similarly, $O2$ stays 0 so $J,K = 0,x$, and $O1$ goes from 0 to 1, so $J,K = 1,x$. The transition is a new one. $O0$ is already 1 and will switch to 0 on the next clock. Thus the inputs will either be $J,K = 1,1$ to toggle the flip-flop or $J,K = 0,1$ to force the flip-flop to 0. Thus a $1@0$ transition gives us an input entry $J,K = x,1$.

Table 17-11:

$O2_n$	$O1_n$	$O0_n$	J2	K2	J1	K1	J0	K0	$O2_{n+1}$	$O1_{n+1}$	$O0_{n+1}$
0	0	0	0	X	0	X	1	X	0	0	1
0	0	1	0	X	1	X	X	1	0	1	0
0	1	0							0	1	1
0	1	1							1	0	0
1	0	0							0	0	0

Proceeding, we still have $O2$ staying 0 but now $O1$ stays 1. That means that either $J=0, K=0$ to keep the present value or $J=1, K=0$ to force the output to 1. So a $1@1$ transition requires an input of $J=x, K=0$. With that we have completed the possibilities and can fill in the whole table.

Table 17-12:

$O2_n$	$O1_n$	$O0_n$	J2	K2	J1	K1	J0	K0	$O2_{n+1}$	$O1_{n+1}$	$O0_{n+1}$
0	0	0	0	X	0	X	1	X	0	0	1
0	0	1	0	X	1	X	X	1	0	1	0
0	1	0	0	X	X	0	1	X	0	1	1
0	1	1	1	X	X	1	X	1	1	0	0
1	0	0	X	1	0	X	0	X	0	0	0

That completes the straightforward part of the problem. Now we have to try to write down expressions that treat the J 's and K 's as outputs and the Ox_n 's as inputs. There is an unusual flexibility to this because of the X entries. Any logic expression that will give the right 0's and 1's is acceptable regardless of what values it gives the X 's.

1. The $J2$ entry can be chosen to be $J2 = O1 \cdot O0$, which will give us a 0 for the single X .
2. The $K2$ entry could be almost anything but the simplest possibility is $K=1$ so we will take that, making all the X 's 1.
3. The $J1$ entry looks as though $J1=O0$ will work. All the non-x bits match and we don't care about the X 's.
4. The $K1$ entry can also be $K1=O0$.
5. The $J0$ entry is most easily realized as $J0 = O2$.
6. The $K0$ entry is another $K=1$ case.

That gives us six equations that can be realized with only one additional gate!

$$\begin{aligned} J_2 &= O_1 \cdot O_0 & J_1 &= O_0 & J_0 &= O_2 \\ K_2 &= 1 & K_1 &= O_0 & K_0 &= 1 \end{aligned}$$

Figure 17-14 shows the resulting synchronous divide-by-5 circuit.

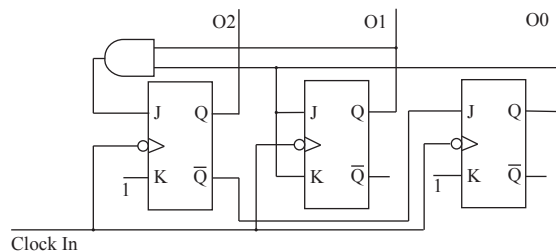


Figure 17-14 Synchronous divide-by-5

It requires no more gates than did the ripple-counter version.

Summary

Synchronous circuits are a special case of sequential logic in which all of the flip-flops are clocked at the same time, clocked **synchronously**.

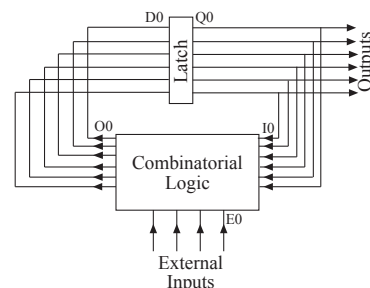
We normally construct synchronous circuits from D-type flip-flops and combinatorial logic. The flip-flops form the memory that stores the current state and the combinatorial logic steers the system from one state to the next in the correct order.

If the output of the system is taken from the latch (as above) then the system is called a **Moore** machine. If the output comes from the combinatorial circuit then it is called a **Mealy** machine.

We design a synchronous system based on D-type flip-flops in a series of stages.

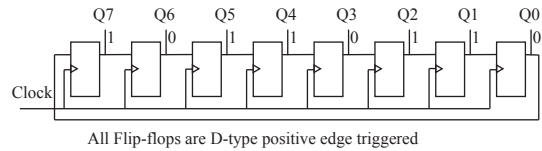
1. Decide how many flip-flops make up the system. Since a set of n binary bits can represent any of 2^n unique states, a system that has anywhere between $2^{m-1}+1$ and 2^m states requires m flip-flops.
2. Draw the state diagram for the system.
3. Compute the values of the D inputs for each output state.
4. Write a truth table to generate the D values from the outputs.
5. Convert the truth table into a combinatorial circuit using standard methods.

While it is possible to design synchronous machines with J-K flip-flops and less external logic it is very rarely worth the effort!



Exercises

1. Design and give a circuit for a 3-bit synchronous down counter, that is a counter that counts through the cycle 7, 6, 5, 4, 3, 2, 1, 0, 7, You can use either D-type or J-K flip-flops; whichever you choose.
2. Design a 3-bit synchronous up/down counter. This is an extension of the kind of synchronous system that we have met so far because it has an external input (in addition to the clock input). There is an external input line Up/!Down that controls the counting direction. When U/!D is true the counter counts up, 000, 001, 010, etc. When the line is low the counter counts down, 000, 111, 110, etc.
3. The circuit below is a rather special synchronous machine. The initial state of each bit is shown. Find the state of each output after four complete pulses have entered on the clock line. You should justify your answer and suggest a name for this circuit.



4. Draw a state diagram for an elevator program. The elevator travels between two floors and has three buttons inside it, Up, Down, and Door close. Each floor has a single Call button and an indicator that can show a '1' or a '2' to tell you where the elevator is. The program should presume the existence of four motor outputs, Door Open, Door Close, Go Up, and Go Down. Left to itself the elevator should sit on the ground floor with the door closed. I offer you the following states but warn you that these are only a sample of the states that you will need and that some of these situations will take several states in the machine because they are encountered under several different circumstances!

Waiting on ground floor with door closed

Waiting on ground floor with door open

Going up

Going down

Waiting on top floor with door closed

Waiting on top floor with door open

Chapter 18: Amplifiers

18.1 Introduction

One of the first great milestones in the electronics revolution was the 1906 invention of the vacuum triode by Lee De Forest. This was the first practical device to **amplify** an electrical signal. To **amplify** is to **increase the size of** and De Forest's Audion vacuum tube was the first device that could take an electronic signal and increase its size. This led directly to improvements in telephones and to the development of "wireless telephony"; what we now know as radio.

The field of analog electronics is dominated by amplifiers. The most common problem that we face is taking a small signal from some device and making it into a larger signal. Consider a radio receiver. The signal picked up by the antenna is typically only a few microVolts in size and carries an energy of pico-Watts. In order to drive a loudspeaker to useful levels we need a signal of the order of Watts and Volts. The signal must be amplified. The heart of every stereo system is an amplifier to take electrical signals and make them big enough to drive a loudspeaker. Amplifiers are found in practically all analog electronic circuits. They create signals, process signals, filter signals, and help transfer signals from one place to another.

Most practical amplifiers are complicated circuits using many active devices. The devices are arranged in **stages**. Each stage (which may itself contain several transistors) performs a portion of the complete amplification of the system. Different stages may play different roles in the overall amplifier. A typical structure, illustrated in Figure 18-1, has a specialized input stage that forms a signal from the difference of two inputs, then one or more gain stages, where the signal is increased in voltage, and finally an output stage that provides sufficient current to drive the final load.

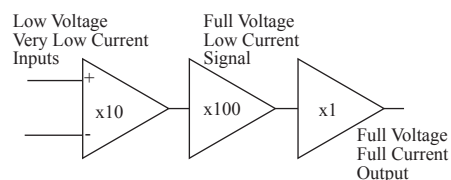


Figure 18-1 Typical Multistage Amplifier Structure

Info The multiple inputs to the amplifier probably seem rather strange at the moment. In later chapters we shall learn how these can be used to implement feedback. Feedback is an extremely powerful technique for improving most of the properties of an amplifier in exchange for reducing the overall amplification.

18.2 Gain

The fundamental property of an amplifier is its **gain**. This is the amount by which it increases the size of the signal. The gain can be expressed in different ways depending on the intended function of the amplifier.

18.2.1 Power Gain

All amplifiers act to increase the **power** of the signal. Thus the power gain is the most fundamental measure of the gain. It is given by

$$G_P = \frac{P_{out}}{P_{in}}$$

As we learned in Chapter 7, we commonly express power ratios using decibel units. We can convert between the two forms using the equations

$$\text{Gain in dB} = 10 \times \log G_P = 10 \times \log \frac{P_{out}}{P_{in}}$$

and

$$G_P = 10^{\frac{\text{Gain in dB}}{10}}$$

Example

One common commercial amplifier, the LM324, claims to have a gain of 100dB. We can find the power gain as a ratio using the above equation.

$$G_p = 10^{\frac{\text{Gain in dB}}{10}} = 10^{\frac{100}{10}} = 10^{10}$$

So the amplifier increases the signal power ten billion fold!

Although all amplifiers exhibit power gain, we reserve the term **power amplifier** for an amplifier that can deliver a large amount of power to a load. The LM324 in the previous example has a huge power gain but it can only deliver tens of milliwatts to its load so it would not be called a power amplifier. Probably the most familiar power amplifier is found in a stereo system. Such an amplifier might deliver anywhere from 10-100W to its load.

18.2.2 Voltage Gain

While the power gain is the most fundamental measure of the gain of a system the voltage gain is the most familiar. This is simply the ratio of the output signal to the input signal.

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

Once again the gain is commonly expressed as in terms of decibels. However, when we do that we always actually give the power gain. Since power is proportional to the square of voltage we have

$$G_p = G_V^2 = \frac{V_{out}^2}{V_{in}^2}$$

so that

$$\text{Gain in dB} = 10 \times \log G_p = 20 \times \log G_V = 20 \times \log \frac{V_{out}}{V_{in}}$$

Example

The amplifier in the previous example is really designed as a voltage amplifier. We can find its voltage gain like this.

$$G_V = 10^{\frac{\text{Gain in dB}}{20}} = 10^{\frac{100}{20}} = 10^5$$

So the amplifier increases the signal voltage by a rather more reasonable hundred thousand times.

18.2.3 Current Gain

Most amplifiers exhibit some current gain. That is, the output delivers more power to the load than the amplifier draws from the source. We express the current gain in the obvious way.

$$G_I = \frac{I_{out}}{I_{in}}$$

The strange thing about amplifiers that are designed specifically as current amplifiers is that they often do not change the voltage of the signal. The output signal looks just the same as the input signal on an oscilloscope but it can be applied to a much smaller load since it can deliver much more current. For example, the output stage of a power amplifier is always a current amplifier. You can make a current amplifier that also has voltage gain but it is quite common to separate the two functions. An amplifier that provides only current gain is sometimes called a **buffer**.

18.2.4 Linearity

A good amplifier should exhibit the property of **linearity**. A linear amplifier increases the size of a signal without altering its shape at all. The term comes from the shape of a graph showing the output voltage as a function of the input voltage. Here is the ideal shape for an amplifier with a voltage gain of 2 (6dB).

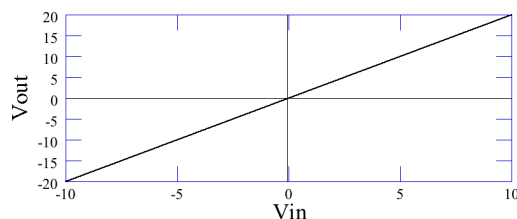


Figure 18-2 Linear Gain Plot

In the ideal case this graph could be extended to infinity in all directions. Real amplifiers cannot do this. There is some maximum output voltage, set by the power supplies, beyond which they will not operate. If you try to get more voltage out than the limit then the signal will get chopped off at the top or bottom. At that point the gain will cease to be linear.

No real amplifier can exactly reproduce an input signal. The output signal always differs from the input signal to some extent. Although, modern amplifiers are exceedingly good and can approach very close to the ideal. However all amplifiers do suffer from problems of two types.

18.2.5 Noise

Noise is a variation of the output signal that is not caused by the input signal. One unexpected consequence of the laws of thermodynamics, the laws that describe how heat energy works, is that any time current flows through a resistance noise is generated. A steady current flowing through a resistor should produce a steady voltage across the resistor. However, thermodynamics says that the voltage will actually vary in a totally random way. The size of the variation depends on temperature and is only tiny at room temperature (microvolts for a 10k resistor) but it is there in all electronic circuits. There are other, related, mechanisms that generate noise in FETs and other semiconductor devices. The end result is that a small amount of randomly varying signal is added to the real signal by every amplifier.

On an oscilloscope screen, noise looks like fur around the signal. To our ears, noise sounds like hissing or rumbling, depending on the source and on the details of the circuit. In a video signal noise most often appears as “snow” or as banding on the screen. All electronic devices add some noise to a signal but some designs are much worse than others. Semiconductor manufacturers tell us about the noise properties of their components and we must simply choose components whose noise is low enough for each particular task. In general, the lower the noise that you will tolerate, the more the components cost.

18.2.6 Distortion

Noise is unwanted information added to the signal by the electronics. It is independent of the signal. **Distortion** is an imperfection in the shape of the output signal. Ideally we should have $V_{out} = G_v \cdot V_{in}$ at all times. Any deviation from this relationship is distortion.

Distortion can take several forms. Two common forms are clipping and cross-over distortion but other more subtle forms exist. Anything that causes the gain, G_v , to vary in any way can cause distortion. Clipping occurs when an amplifier is asked to produce more voltage (or current) than it can deliver. The clipping takes place at the extreme maximum and minimum points in the signal. Cross-over distortion is an error seen where the signal passes through zero. The design of early transistorized amplifiers made them particularly susceptible to cross-over distortion but it has not been a significant problem in reasonable quality equipment for many years.

18.2.7 Clipping

Let us look first at some kinds of clipping. Distortion is most easily seen by looking at the effect on a sine wave. First we will remind ourselves what a sinewave should look like.

Info *Noise Specifications

A noisy signal is made up from a very large number of signals with all different frequencies. Because of this, the actual amount of noise that a device contributes to a signal depends on the bandwidth of the device. In fact the classic thermal noise has the property that the amount of power in a given frequency range is independent of the frequency. A device that lets through a wide range of frequencies will also let through a wide range of noise and so will have a larger noise contribution.

In order to remove the effect of the system bandwidth manufacturers specify the amount of noise power that a device introduces as so many watts per Hertz. This measure of power/Hz is called **noise density**. For a resistor the noise density is independent of frequency.

We are usually interested in the noise voltage rather than the noise power so that we need to be told the Volt²/Hz. In fact, the specification is normally given as the square root of this number and so noise is normally given as V / \sqrt{Hz} .

For example, the popular OP27 precision amplifier claims a noise density of $3nV / \sqrt{Hz}$. Thus if we build an amplifier with a bandwidth of 20kHz (good for audio work) then the total noise will be $3nV \cdot \sqrt{20000} = 424nV$. This noise is treated as if it is added to the signal at the **input** of the amplifier and so you can't use an OP27 to measure signals at the 1μV level without introducing nearly as much noise as signal in an audio bandwidth amplifier. However, if you want to measure brainwave signals with a 10Hz bandwidth then the noise is only $3nV \cdot \sqrt{10} = 10nV$ and so you can measure microvolt signals with ease.

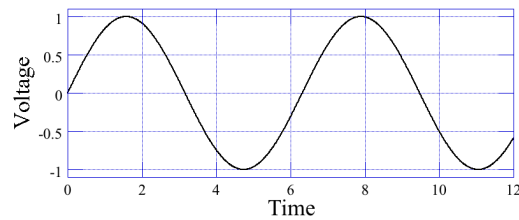


Figure 18-3 Pure Sine Wave

Here is a real sine wave. The time and voltage scales are purely arbitrary. Note the straight sides as the sine goes through zero and the smoothly rounded peaks and valleys. If you listen to such a signal at a frequency around 1kHz it will sound like a clear whistle.

If we pass that signal through an amplifier that suffers from clipping then the tops and bottoms of the wave will become cut off. In the worst cases we might see something like this

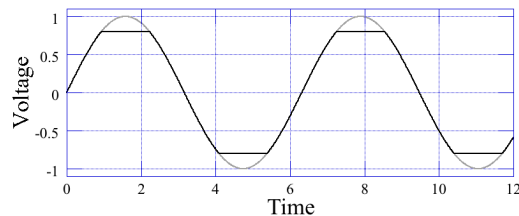


Figure 18-4 Hard Clipped Sine Wave

The hard clipped wave is shown in solid black with the original sinewave shown in gray for comparison. This is what usually happens when you try to force a transistor amplifier to produce an output signal that is larger than it can handle. The clipping could be due to trying to produce too much voltage or too much current. Either would look much the same. If you listen to this wave then it will have a nasty, harsh, whiny sound. Very unpleasant compared to the pure sine. If you do this to music then the resulting sound is extremely unpleasant.

Although any amplifier will clip the signal if you try to overdrive the amplifier, it does not have to do so in such a nasty fashion. Old vacuum tube circuits were famous for clipping in a much more graceful way, known as soft clipping. Here is a soft clipped sine wave.

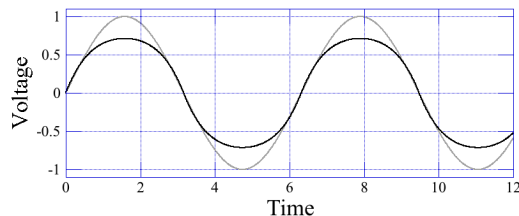


Figure 18-5 Soft Clipped Sine Wave

The output voltage still does not go above about 0.8 but now the peak is rounded and there are no hard shoulders to the wave. This sounds a lot less unpleasant than the hard clipped wave. The tone quality is altered but it does not have the harsh, whiny quality of the hard clipped wave. Electric guitarists are fond of vacuum tube amplifiers because they exhibit this soft clipping behavior. Guitarists commonly overdrive their amplifiers to get the extra bite that the soft clipped sound has. You can do the same thing with a fancy transistor amplifier but only at considerable expense. On the other hand, until it is overdriven, a transistor amplifier can offer much less distortion and much higher efficiency than a vacuum tube amplifier.

18.2.8 Crossover Distortion

Clipping is a flaw at the outer limits of an amplifier's gain curve. Crossover distortion happens instead where the signal level is near zero. In its clearest form it results in small flat areas in the output wave. Here is a sinewave that exhibits significant crossover distortion.

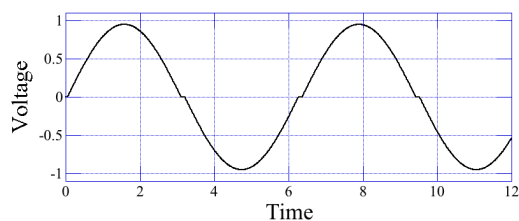


Figure 18-6 Sinewave with Crossover Distortion

The effects of crossover distortion are clearly visible as little flat areas in the voltage every time that it crosses through zero. Again, the effect on the sound is very unpleasant. It acquires a coarseness that is especially distressing when applied to music. This form of distortion occurs when an amplifier uses different transistors to generate the positive half of the wave from those used to generate the negative half. If the transition from one output to the other is not perfect then the result is crossover distortion. The usual method of eliminating this is to have a small region round 0V where both the positive and the negative halves of the circuit are operating and so to make the join invisible. Any decent amplifier made in the past 20 years should be free from audible crossover distortion.

18.2.9 Bandwidth

The bandwidth of an amplifier is the range of frequencies that the amplifier can pass without error. No amplifier can operate at all possible frequencies and so all amplifiers have a finite bandwidth. This is not a real problem because any particular signal only consists of a particular range of frequencies and so it is sufficient if the amplifier bandwidth includes that range.

Example

The human ear is only sensitive to frequencies between about 20Hz and 20kHz (and the two extremes are only audible to a tiny fraction of the population). Thus an audio amplifier need only amplify signals in this range. In practice there is little point in limiting the low frequency end and most commercial audio amplifiers operate from DC to 20kHz.

Example

A complete NTSC broadcast color television signal has to be transmitted in a 5MHz segment of one of the ranges of frequencies allocated to broadcast TV. For example, the UHF band lies between about 470MHz and 800MHz. Thus the first amplifier in a television receiver must have a bandwidth that covers 470-800MHz. However, the later stages of the amplifier chain need only process the 0-5MHz bandwidth of the actual TV signal and do not need to deal with the much higher frequencies of the radio wave used to carry the information.

It is usual to specify the bandwidth of an amplifier in terms of its 3dB points. These are the frequencies at which the power gain has fallen to one half of its maximum value. If we wish to be complete then we show the amplifier gain on a Bode plot (c.f. Section 7.5.1). This allows us to see the complete gain behavior of the amplifier.

In situations where it is of interest we often remind ourselves that the gain does depend on frequency by making the dependence explicit. Thus you will often the gain written as $G(f)$ or, more commonly, $G(\omega)$.

The finite bandwidth of an amplifier means that the amplifier's output will not be a perfect image of its input for any signal that contains information at frequencies outside the amplifier bandwidth. This effect is most easily seen with a square wave signal, because that contains frequencies that theoretically extend to infinity in order to produce the sharp corners characteristic of a square. Thus if we feed a square wave signal into a real amplifier then the frequency components above the amplifier's bandwidth will not be amplified as much and the edges of the square wave will be rounded off, exactly as we saw in Chapter 7. Here is a picture of an ideal 1kHz square wave and the result of passing that wave through an amplifier with a gain of 2 and a bandwidth of 20kHz.

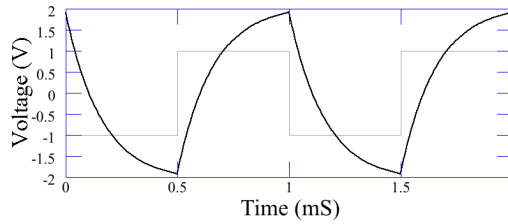


Figure 18-7 20kHz Bandwidth Limited 1kHz Square Wave

As you can see, the corners of the wave have been rounded off as the high-frequency information needed to describe the sharp corners is lost in the amplifier.

It must be stressed that although the output signal is not a perfect copy of the input signal we do NOT consider this to be a form of distortion. Distortion adds signal dependant information to the signal. The finite bandwidth simply removes some information from the signal. So long as the bandwidth is chosen large enough to pass the information of interests this is not a problem in the amplifier, merely a design factor.

Example

You can hear the effects of bandwidth limitation by playing with the tone controls on a stereo system. The treble and bass tone controls adjust the upper and lower 3dB points of the amplifier. You can hear how lowering the treble control makes the sound muffled while lowering the bass control (reducing the bass content) makes the sound seem thin and tinny.

18.3 Thévenin Model of an Amplifier

A good amplifier has a simple Thévenin model. The input acts as a simple resistor and the output consists of a voltage source in series with a resistor.

The input resistance of the amplifier is usually made very large to minimize the effect that the amplifier has on the signal source. We call that effect, whereby connecting a signal to another component reduces the level of the signal, **loading**. Thus we try to make amplifiers with high input resistance to minimize loading. Typical values range from 10kΩ on up.

The Thévenin voltage is shown as depending on the input voltage. This represents the ideal behavior of the amplifier. That voltage then passes through the output resistance, R_{out} , to become the actual voltage across the output terminals of the amplifier. In order to make the output of the amplifier insensitive to loading by further stages we try to make R_{out} small. Power amplifiers typically have values of about 0.05Ω or less.

If the output impedance of the amplifier is not small compared to the resistance of the load then the output voltage will be lower than the Thévenin voltage. The load and the output impedance form a voltage divider so that the load receives a smaller and smaller fraction of the output voltage as the output resistance increases.

Another way to look at this effect is to consider the current that flows to the load. As the load resistance decreases then the current needed to force the output voltage to its correct value will increase. As the output current increases so does the voltage drop across the output resistance and so the output voltage falls.

The fall in output voltage as the load current increases is seen as a fall in the gain of the amplifier. The actual output voltage, measured across the load resistance, is smaller than it should be and thus so is the gain. When the load resistance is the same size as the output resistance only one half of the Thévenin voltage appears across the load and so the gain is only one half of its open-circuit value.

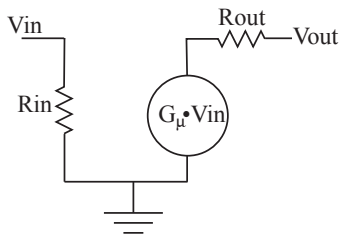


Figure 18-8 Thévenin Model of an Amplifier

Info Because the input and output resistances normally depend, at least somewhat, on frequency it is more correct, and more common, to refer to them as input and output impedances.

18.4 Common-Source FET Amplifier

Because an amplifier must increase the power in a signal it must use an active device. As we learned in Chapter 11, an active device is one that uses power from a separate external source to increase the power in a signal. We shall now explore the ways that we can use active devices, FETs, to build amplifiers.

The simplest FET amplifier is shown in Figure 18-9. The circuit is called a **common-source** amplifier because the source terminal is common to the input and output sides of the amplifier. The input is applied between the gate and the source and the output taken across the drain and the source.

As we shall see in the next few pages, this amplifier has too many problems for it to be much use in this form. However, almost all of the useful designs are variations on this one and so all that we learn about this design will carry over to the other designs. Accordingly we shall study this circuit in considerable detail.

The input signal is connected to the gate of an NFET and the drain current allowed to flow through a resistor, R_D . The output is taken from the junction between the drain and the resistor. We can express the output voltage, V_{out} , in terms of the resistance, the drain current I_{DS} , and the power supply voltage V .

$$V_{out} = V - I_{DS} \times R_D$$

Now the drain-source current is controlled by the gate-source voltage V_{in} . They are related by the transconductance curve of the FET. Here is an actual transconductance curve measured for a particular 2N7000 FET operating with a drain-source voltage of 15V.

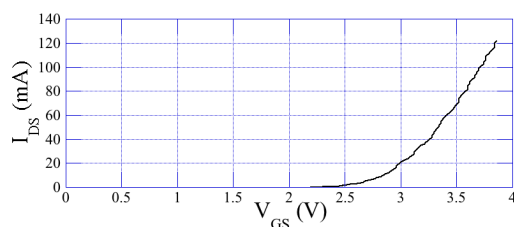


Figure 18-10 2N7000 Transconductance at $V_{DS}=15V$.

As usual, no current flows in the drain until the gate-source voltage reaches the threshold level and after that the current rises increasingly rapidly. We can use the transconductance curve to predict the output voltage for any given input voltage.

Example

Consider an amplifier using the circuit of Figure 18-9 with $R_D = 200\Omega$ and a power supply voltage $V = 15V$. At an instant when $V_{in} = 3V$ we know $V_{GS} = 3V$. From 2N7000 Transconductance at $V_{DS}=15V$, we see that when $V_{GS} = 3V$ the drain current is $I_{DS} = 20mA = 0.02A$.

Now we can substitute into the output equation to find

$$V_{out} = 15 - 200 \times 0.02 = 15 - 4 = 11V$$

If we do this for a 3V sinewave input voltage then we get the very strange output voltage shown below (Figure 18-11).

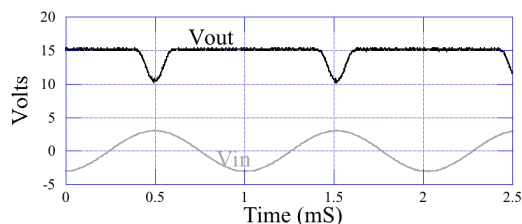


Figure 18-11 Simple 1-FET Amplifier Output

The output signal is extremely distorted. It spends almost all of its time stuck up at 15V and only dips down below it when the very tips of the input signal exceed the threshold voltage.

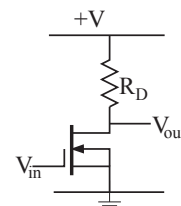


Figure 18-9 Common Source FET Amplifier

Obviously this is useless for amplifying signals that spend any of their time below the threshold voltage. If we add an offset to the input so that it stays above the threshold then we get better results (Figure 19-12).

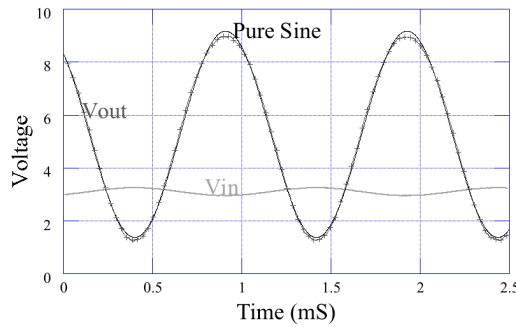


Figure 18-12 Simple 1-FET Amplifier with Biased Input

The light gray line shows the input voltage. Its average value is now about 3V and its amplitude has been reduced to only 0.3V peak-peak. The dark grey line with the crosses shows the output. It looks very much better. It now looks like a sinewave with the correct frequency and phase but with an offset of about 7.5V. The peak-peak amplitude is about 7.8V, corresponding to a voltage gain of 26. More careful inspection, however, shows that the output is not quite as perfect as it appears at first.

The thin black line running beside Vout is a perfect sinewave with the same amplitude, frequency, and offset as Vout. You can see that there are systematic differences. Vout always lies a little below the sinewave at the peaks and valleys. This is because the gain of this amplifier is not quite constant. Instead the gain is a little lower for low input voltages and a little higher for higher ones. Thus the output is slightly distorted. The distortion is greatest twice per cycle of the sinewave, once at the peak and once at the trough. Thus the distortion has a frequency twice that of the signal. We call this **second harmonic distortion**.

18.4.1 Gain of the common-source FET amplifier

The second harmonic distortion seen in Simple 1-FET Amplifier with Biased Input is related to the shape of the transconductance curve for the FET. Let us make a very simple model of the transconductance and consider how the model system would behave. Our model is going to replace the smooth curve of Figure 18-10 with a pair of straight lines. For input voltages below about 2.8V we will approximate $I_{DS} = 0$. For input voltages above 2.8V we will fit a straight line to the transconductance. This will give us the model transconductance of Figure 19-13 below.

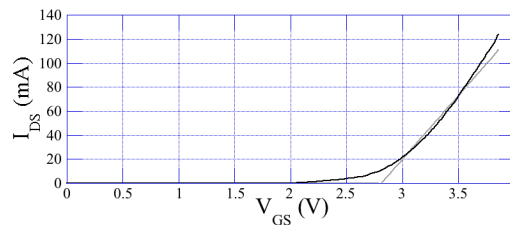


Figure 18-13 2N7000 Transconductance with Linear Model

Note In our model the transconductance, g_m , is constant but in the real FET g_m varies from an initial value of 0 to a steadily increasing value once V_{GS} is above the threshold voltage. At high values of V_{GS} it seems to be getting more constant but that is an artifact of the limited scale on which the graphs are drawn. In reality g_m increases continually as V_{GS} increases.

This is called a **linearized** model of the FET transconductance. It has the big advantage that we can write a simple algebraic expression for I_{DS} in terms of V_{GS} for input voltages above threshold. Since the line is straight we have

$$I_{DS} = g_m \times V_{GS} + b = 106 \times V_{GS} - 242$$

$$I_{DS} = g_m \times (V_{GS} - V_{Th}) = 106 \times (V_{GS} - 2.8)$$

where V_{GS} is in volts, I_{DS} is in mA and the slope factor, g_m , is in mA/V or mMho.

Once we have an explicit equation for I_{DS} we can substitute the model equation into the equation for V_{out} to find

$$V_{out} = V - (R_D \times I_{DS}) = V - R_D \times g_m \times V_{in} - R_D \times b$$

Thus V_{out} contains two parts, a constant term ($V - R_D b$) and the amplified signal. Now we can identify the voltage gain of the circuit with the term multiplying V_{in} and say

$$G_V = -R_D \times g_m$$

The gain equation tells us two things.

1. The common-source amplifier is an inverting amplifier. That is the output is 180° out of phase with the input.
2. The gain is determined by the transconductance of the FET and by the drain resistor. This means that we usually use a large drain resistance to get a high gain.

At first glance it seems as though we can get as much gain as we like out of the amplifier simply by making R_D large enough. Unfortunately, increasing R_D decreases the average current through the FET and so the FET's transconductance gets smaller, as we saw in Figure 18-13. So we have to compromise on a suitable value for R_D .

While it is possible to find the optimal resistance for a given FET, in practice we usually use trial and error to select suitable values. First we select an FET from whatever is available and select a convenient power supply voltage. Then we pick a value for R_D and calculate the approximate current flow in the FET. This allows us to read the g_m value from the transistor data sheet and so to estimate the gain. If we are close to or above the desired value then we can try some other resistance values to see how much better or worse they are. If we are a long way below the desired gain then we must choose a higher transconductance transistor or try a larger power supply voltage.

Example

I constructed an AC coupled common-source amplifier with a 2N7000 FET and measured the gain as a function of the drain resistance. In each case I computed the effective average g_m from the gain equation and obtained the results in Table 18-1. At first, the rise in g_m as R_D falls is not enough to counteract the fall in R_D and the gain rises as R_D rises. The gain reaches a maximum near $R_D = 100k$ and after that the decrease of R_D dominates and the gain falls steadily. These measurements were made with a 15V power supply and the output biased to 7.5V. The final column shows the average DC current flowing the FET.

Info There is a way to beat the fall in g_m as you increase R_D . What we want is a component that can pass a relatively large current and still have a very high resistance. Such a component is a **constant-current source**. In section 11.6.2 we saw how to make a constant-current source with a single FET. If we replace the drain resistor with a PFET constant current source then we can get a very large effective drain resistance while still letting several mA flow through the amplifier FET. For example I was able to get a gain of 240 from a single 2N7000 with a ZVP4105 current source load. This compares very favorably to the results in the example.

Table 18-1:

R_D (Ohms)	G_V	g_m (mMho)	I_{DS} (mA)
1M	71	0.071	0.0075
100k	80	0.8	0.075
10k	68	6.8	0.75
1k	45	45	7.5
100	16	160	75

18.4.2 Linearity of the Common Source Amplifier

If we were to put our offset sinewave into an amplifier constructed with the model FET then the V_{out} would be a perfect sinewave, as shown by the thin gray line in Figure 18-12. The distortion of the real output sinewave is caused by the difference between the real FET and the model. The real transconductance is not a straight line and so the output is not exactly proportional to the input since the gain changes from point-to-point along the signal. The distortion is caused by non-linearity in the transconductance of the FET.

Note that because the real transconductance follows a smooth curve the non-linearity gets worse and worse as the size of the input signal increases. If we had made measurements with a 0.03V p-p sinewave instead of a 3V p-p wave then it would have been very hard to see the distortion, though it would still be measurable with suitable equipment.

18.4.3 Biasing the common-source FET amplifier

In order to make our common source 1-FET amplifier produce any kind of reasonable output signal we had to add a large offset to the input signal. We call such an offset, added to put the amplifying transistor into a useful part of its operating range, a **bias voltage**. Sometimes it is possible to arrange the stages of an amplifier so that the output offset of one stage can become the input bias for the next stage. We shall see an example of this at the end of this chapter. However, in the many circuits this is not practical and we have to add extra components t

Info In commercial amplifier designs, especially integrated circuit amplifiers, it is not uncommon to find that most of the transistors in an amplifier are part of a bias network. The actual amplification is done by only a small fraction of the total number of transistors. For example, only 5 of the 20 transistors in the classic 741 op-amp are in the direct signal path. The rest are either in bias networks or in circuits to protect the amplifier from abuses such as shorting the output to ground.

form a bias network whose sole function is to push an amplifying FET into the right operating region.

The simplest form of a bias network is an ordinary resistor voltage divider operating off the supply voltage (Figure 19-14). This will allow us to generate any bias voltage between ground and the supply voltage.

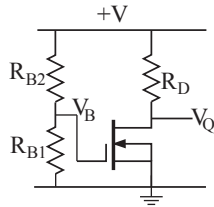


Figure 18-14 Common-source FET Amplifier with Bias Network

Choosing the resistors for a bias network is a multi-step process. First we must determine the operating point, the voltage on the output when there is no input. Because the output of the amplifier can go neither higher than the supply voltage nor lower than ground, we usually choose an operating point about half way between. That way there is the most room for the output to swing above and below its average value. The largest amplitude that we can theoretically get out of the amplifier is $V/2$ where V is the supply voltage. Real FETs cannot swing all the way down to ground so we may want to pick an operating point slightly above $V/2$.

Once we know the operating point, V_Q , we can find the bias voltage needed. The drain current at the operating point is $(V - V_Q)/R_D$ so we can use the transconductance curve to find the value of V_{GS} that will give us the correct V_Q . We will call this value V_B .

Knowing the desired bias voltage and the power supply voltage then we know that the bias resistors must be chosen so that

$$V_B = \frac{R_{B1} \times V}{R_{B1} + R_{B2}}$$

Since the gate of an FET draws no DC current, we are free to choose the sum of the bias resistors as large as we want. The larger they are the less power is wasted in simply heating up these resistors. In practice, it is usually easiest to pick a value for one of the resistors and let the formula determine the other. That way at least one of the resistors can be made a standard value. The other will almost certainly not be and we may have to make it up from a variable resistor or a variable resistor in series with a standard value. This has the added benefit that we can trim the bias voltage to account for differences in threshold between one transistor and another.

Example

Design a bias network for the amplifier in Figure 18-12.

The output signal in Figure 18-12 is nicely centred between ground and the 15V power supply so that the average level of the input is a good bias voltage. That average level is 3.1V. So we need to choose resistors to generate 3.1V from a 15V power supply. Since 3.1V is much less than 15V R_{B1} will be smaller than R_{B2} so I will choose $R_{B1} = 100k$ and let the formula find R_{B2} :

$$3.1V = \frac{R_{B1}V}{R_{B1} + R_{B2}} = \frac{100k \times 15}{100k + R_{B2}} \text{ so that } R_{B2} = \frac{100k \times 15}{3.1} - 100k = 384k$$

That is not a standard value so I will make it up from a 330k fixed resistor in series with a 100k variable resistor.

Remember

To design a common-source FET bias network for a given FET, R_D and supply voltage V :

1. Select an operating point, V_Q , approximately equal to $V/2$.
2. Find the quiescent current through the FET $I_Q = (V - V_Q)/R_D$.
3. Use the FET transconductance data to find the bias voltage, V_B .
4. Select one bias resistance and then find the other from the voltage divider equation.

Info Bias Stability

This sort of common-source FET amplifier suffers from a fairly serious problem with the bias level. The operating point is determined by the bias voltage and by the transconductance curve of the FET. Unfortunately, the transconductance curve is somewhat sensitive to changes in temperature, as we saw in section 11.5.4. As the transistor warms up because of its own power dissipation, the threshold voltage drops. The bias voltage stays constant so that the FET will turn on more strongly and the operating point will fall. At the least this will decrease the available output swing of the amplifier and in extreme cases it may force the operating point so far down that the amplifier ceases to function.

This sensitivity to temperature can be tempered by putting a resistor in series with the source, at the cost of decreased output swing. However, the effect is small unless you are prepared to severely limit the output swing and so this kind of amplifier is quite rare on its own. It is usually used as part of a more complex circuit that includes better methods of controlling the bias.

18.4.4 Coupling the Signal In and Out

We cannot simply connect most signal sources to the junction between the FET gate and the bias network because the bias voltage will interfere with the operation of whatever circuit generated the signal. Similarly, we cannot just connect the output of the circuit to most loads because of the large DC offset on the signal. For example, a loudspeaker is not designed to have any average DC current flowing through it and would be very unhappy connected to such an output.

We need a component that can pass a time varying signal while blocking passage of any DC current. Since this is exactly what a capacitor does, we use capacitors to couple the signal into and out of such an amplifier as in Figure 19-15 on the right.

An amplifier that has capacitors at its input and output is called an **AC coupled** amplifier. It has the fundamental limitation that it can only amplify signals that change fast enough to get through the input capacitor. In some situations this is an advantage as it makes the amplifier insensitive to very low frequency noise. In other situations it is a disadvantage because the signal of interest moves too slowly.

Example

One example of an AC coupled amplifier is the EEG amplifier used to record brainwaves. We are only interested in electrical activity faster than about 1Hz and so an AC amplifier is often used, especially as there are serious sources of very low frequency noise that can upset so delicate an instrument.

An example where a DC amplifier is needed is a thermocouple amplifier used to monitor temperature in a kiln. We don't want the temperature in our kiln varying rapidly up and down; we want it to be very stable over a period of many hours. Thus we cannot use an AC amplifier in such a system.

The AC coupling capacitors implicitly set a lower limit on the frequency that an amplifier can handle. The input capacitor forms a high-pass filter with the input bias resistors as we can see if we look at the Thévenin equivalent of our circuit shown in Figure 19-16.

There are several things to notice about this circuit.

1. There is no positive power supply and both bias resistors now go to ground.
2. The input resistance is made up only from the input resistors. The FET gate draws no current and so contributes nothing to the input resistance.
3. The Thévenin voltage now depends on V_G , the gate voltage, rather than on V_{in} .

The Thévenin equivalent describes the circuit as seen by the signal. From the signal's point of view there is a voltage generator, $G_V \cdot V_G$, that supplies the output signal and it does not care that the power for this generator comes from a positive power supply. That is why the positive power supply is not shown.

Similarly, from the signal's point of view the positive power supply and the ground wire are essentially the same place and so R_{B2} is shown connected to that place. There are two ways to look at this point of view.

- The mathematician notes that both the positive power supply and the ground rail are places whose voltage does not change. When we consider only the time varying part of a voltage (the signal) then both of these wires have no variation and so act as if they were connected together.
- The engineer looks back along the wires to the power supply. There she sees a very large capacitor going from the positive supply to ground (the filter capacitor used to suppress ripple). Since a large capacitor acts like a short circuit for signal frequencies she says that the power supply and the ground rail are actually connected together at all interesting frequencies.

The actual gain comes from the FET and so the output voltage is controlled by the gate-source voltage of the FET. Since the source is grounded that means that the output is controlled by the gate voltage, V_G . However, the input, V_{in} , is connected on the other side of the input capacitor from V_G and so current has to flow through that capacitor to alter V_G . This will happen more easily at high frequencies, where the capacitor acts like a low resistance, and not at all at extremely low frequencies, where the capacitor acts as an open circuit. Thus we have a low pass filter formed by C_{in} and the parallel resistance of the two input bias resistors. We recall from Chapter 8 that a low pass filter has a cut-off frequency at

$$f = \frac{1}{2\pi RC} = \frac{1}{2\pi R_{B1} C_{in}}$$

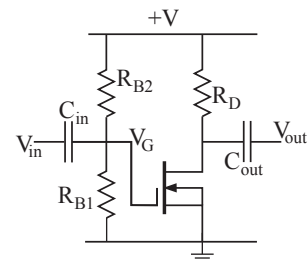


Figure 18-15 AC Coupled Common-Source FET Amplifier

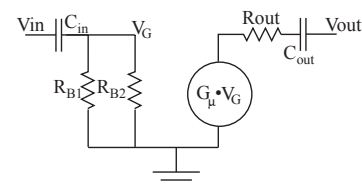


Figure 18-16 Thévenin Equivalent of AC Coupled Amplifier

where $R_{B//}$ is the resistance of the parallel combination of R_{B1} and R_{B2} .

This sets the low frequency cut-off of the whole amplifier and so we must choose the input capacitor large enough to make this frequency lower than the lowest frequency that we need to amplify. This is usually not difficult as the bias resistors can be made very large. Quite moderate values of capacitance, 0.01-0.1 μF , are usually sufficient.

Example

Choose a coupling capacitor for the amplifier in the previous example assuming that it must amplify audio signals.

We know that an audio amplifier needs to have a frequency response from 20Hz to 20kHz and so we need to set the low frequency roll-off of the amplifier at or below 20Hz. In practice setting it at 20Hz is sufficient.

The amplifier of the example had bias resistors of 100k and 384k so their parallel equivalent is

$$R_{B//} = \frac{R_{B1} \times R_{B2}}{R_{B1} + R_{B2}} = \frac{100 \times 384}{100 + 384} = 79.3 \text{ k}$$

So the input capacitor needs to be at least as large as

$$C_{in} = \frac{1}{2\pi R_{B//} f} = \frac{1}{40\pi \times 79300} = 10^{-7} \text{ F}$$

This happens to be a standard value so we can use a 0.1 μF coupling capacitor.

18.4.5 Frequency Response of the common-source FET amplifier

As we have just seen, the low frequency response is determined by the input coupling capacitor. If we DC couple the input then the amplifier will work perfectly down to DC, so long as we are careful to keep the FET turned on.

In order to understand the high frequency response we must turn to the AC model of an FET that we studied in Chapter 11. There we learnt that every FET has a set of capacitances built into it. When we add the parasitic capacitances to the Thévenin model, we get a model for the complete DC coupled amplifier that looks like Figure 19-17 below.

Note Resistor R_D is the real drain resistor in the amplifier. R_{DS} is part of the Thévenin model of the FET and accounts for the slight variation of FET output current with drain voltage. It is usually large compared to R_D but is shown for completeness.

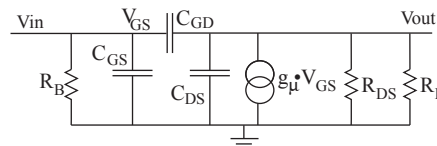


Figure 18-17 AC Thévenin Model of Common-Source FET Amplifier

The output capacitance, C_{DS} is in parallel with the output resistance. Because the voltage gain of the amplifier at DC is $g_m \cdot R_D$ the output capacitance affects the high frequency gain of the amplifier. As the frequency increases C_{DS} looks like a smaller and smaller resistance. At some point it will become comparable to R_D and the amplifier gain will begin to fall. From this point the gain falls steadily as the effective resistance of C_{DS} continues to fall.

In practice, the effect is made worse by the presence of C_{GD} , the reverse transfer capacitance. When an AC signal is applied to the input of the amplifier an enlarged, inverted signal appears at the output. Some of that signal leaks backwards through the reverse capacitance and reduces the gate voltage and so makes the gain appear to drop. Thus, the small reverse capacitance has a similar effect to the output capacitance.

Note The Miller effect is just as important in setting the effective output capacitance and input capacitance for switching circuits. It again acts to set a limit on the high frequency behavior of the FET.

A more advanced analysis shows that the effect of C_{DS} can be accurately modelled by replacing the feedback capacitance with a much larger capacitance, $C_M = (1+G_V) \cdot C_{DS}$, in parallel with the drain source capacitance and a similar C_M in parallel with the input capacitance. This is called the **Miller Effect**, after its discoverer, and is a major problem in simple amplifiers. Although the actual reverse capacitance is usually quite small compared to the output capacitance, the high voltage gain can make the Miller capacitance much larger than the intrinsic output capacitance.

Example

A 2N7000 FET has output capacitance of 25pF and reverse capacitance of only 5pF. In an amplifier with a fairly typical drain resistance of 10k we have a voltage gain of about 80 so that the Miller capacitance, $(1+G_v)C_{DG} = 81 \cdot 5 = 405\text{pF}$, about 20 times the intrinsic output capacitance! Similarly, we have to add C_M to the input capacitance of 60pF to make a total input capacitance of 465pF. Again, the Miller capacitance dominates the input capacitance.

We now have enough information to estimate the maximum frequency to which our simple amplifier will work. The gain will have fallen to half of its maximum when the effective resistance of the total output capacitance is equal to R_D (ignoring R_{DS}). That will happen at the frequency f for which

$$R_c = \frac{1}{2\pi f C} = \frac{1}{2\pi f (C_{DS} + (1+G_v) \times C_{DG})} = R_D$$

so that the upper cut-off frequency will be at about

$$f = \frac{1}{2\pi R_D (C_{DS} + (1+G_v) \times C_{DG})}$$

One immediate lesson from this is that the high frequency cut-off depends on R_D . That means that we can improve the high frequency response by reducing R_D . Of course, that will also reduce the gain of the amplifier so there is no free lunch here.

Example

In our 2N7000 based amplifier with a drain resistance of 10k and its combined output resistance of 25pF+405pF=430pF we will have a high frequency cut off at about

$$f = \frac{1}{2\pi \times 10\text{k} \times 430\text{pF}} = 37\text{kHz}$$

This value is in reasonable agreement with measurements on such an amplifier. The actual cut-off is somewhat higher corresponding to a lower feedback capacitance. However, the data sheet value is a maximum and it is quoted at $V_{GS} = 0\text{V}$. The actual capacitance alters somewhat as the bias voltage is applied and so exact agreement is not to be expected.

Putting together the low-frequency roll-off caused by the AC coupling and the high-frequency roll-off caused largely by the Miller capacitance we understand the complete frequency behavior of the common-source FET amplifier.

18.4.6 Common Source Output Drive

The common source amplifier is quite good at producing voltage gain but it has a very poor ability to drive current into an external load. We have seen that the way to get high amplification is to use very high values of R_D . Unfortunately, since all the output current must flow through R_D , this severely limits the output current that can be drawn before the output voltage suffers.

The usual way to quantify this is by the output Thévenin resistance of the amplifier. A simple calculation shows that this is just the parallel combination of R_{DS} and R_D , which is usually just R_D . So the higher you make R_D to increase the gain, the higher you make the output impedance of the amplifier. This means that to keep the gain high you cannot draw any significant amount of current from a common source amplifier. It is a good voltage amplifier but is not suitable for the output stage of a complete amplifier.

Chapter 19:*Multi-Stage Amplifiers

The common-source amplifier delivers significant voltage gain but has only a single input and cannot supply significant current to its output. If we are to build a multi-stage amplifier as described in section 19.1 then we also need input stages and output stages.

The input stage must take two input voltages and create from them a new signal equal to their difference. We call such an amplifier stage a **differential amplifier** or a **difference amplifier**. As will become clear in the next chapter, having a differential input stage will allow us to use the powerful technique of feedback to build much more linear amplifiers than we could otherwise.

The output stage must take a signal and apply it to a low resistance load, one that draws a lot of current. The common-source amplifier cannot do it. About 1 milliamp is the practical limit on its output current. Real world loads often take much more current and so we need specialized output circuits that can deliver that current. As we shall see, they do not have to provide voltage gain as well; we can leave that to common-source stages.

Example

Most loudspeakers have a resistance of only 8Ω . High power speakers, especially speakers designed for automobiles, have even lower resistance—often 4Ω and sometimes only 2Ω . A 100W audio amplifier must be able to deliver a peak current of about 5A to an 8Ω load or 7A to a 4Ω load.

One additional thing to notice about all the circuits in this chapter is that they use **bipolar** power supplies. A bipolar power supply has both positive and negative power supplies that share a common ground connection. Signals are all referenced to this common ground level and can thus go both positive and negative with respect to ground. This will be particularly important for the output stages but it is common practice for both input and output stages that need to deal with signals that spend time on both sides of ground.

19.1 The difference amplifier

A perfect difference amplifier should have the transfer function

$$V_{out} = G_V \times (V_{in+} - V_{in-})$$

The output then depends only upon the difference between the two input voltages and not at all upon their actual levels.

Example

Find the output from a perfect difference amplifier with a gain of 10 if the inputs are

- 1) $V_{in+} = 9.41\text{V}$, $V_{in-} = 9.38\text{V}$
- 2) $V_{in+} = -7.45\text{V}$, $V_{in-} = -7.48\text{V}$.

In the first case we have

$$V_{out} = G_V \times (V_{in+} - V_{in-}) = 10 \times (9.41 - 9.38) = 10 \times 0.03 = 0.3\text{V}$$

while in the second case we have

$$V_{out} = G_V \times (V_{in+} - V_{in-}) = 10 \times (-7.45 - (-7.48)) = 10 \times 0.03 = 0.3\text{V}$$

So that the output is exactly the same despite the large difference in the actual values of the two sets of inputs.

In practice this impossible to achieve. Real amplifiers all show some dependence upon the actual level of the inputs. We account for that by adding a second term to the transfer function like this.

$$V_{out} = G_V \times (V_{in+} - V_{in-}) + G_{CM} \times (V_{in+} + V_{in-})$$

The new term is called a **common-mode** term and it is characterized by its own **common-mode gain**, G_V . A good differential amplifier should have its common-mode gain very much smaller than its differential gain, G_V , so that the amplifier approaches the ideal as closely as possible. Instead of quoting the common-mode gain we usually specify the ratio of the two gains, the Common-Mode Rejection Ratio, in dB

$$CMRR = 20 \times \log_{10} \frac{G_V}{G_{CM}}$$

A decent discrete transistor differential amplifier should have a CMRR of at least 40dB, corresponding to $G_{CM} = G_V \times 100$. The main limitation on CMRR is the matching of the transistors so integrated circuit amplifiers can do better. The best integrated circuit amplifiers offer CMRRs in excess of 100dB

19.1.1 The Long-Tailed Pair

Although there are a number of circuit configurations that can perform the difference operation by far the most common circuit is the long-tailed pair shown, in its simplest form, in Figure 19-18 .

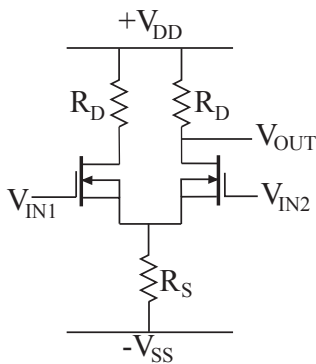


Figure 19-1 Long-tailed Pair Difference Amplifier

Info The name long-tailed pair comes from the appearance of the circuit diagram. The source resistor, R_S , is the long tail.

If we ignore the FET on the left for a little while then we are left with a slightly modified common-source amplifier. The extra resistor between the source and the negative power supply lowers the gain of the amplifier but does not really alter its behavior very much. V_{out} is inverted relative to V_{in2} so that raising V_{in2} lowers V_{out} .

Now consider what happens when we apply a second voltage to V_{in1} . If we increase the voltage on V_{in1} then more current flows in R_S and the voltage at the top of R_S rises. This decreases the gate-source voltage of the right-hand FET and so turns it off somewhat, causing V_{out} to rise. Thus, raising V_{in1} raises V_{out} . We have achieved the general behavior that we want.

Note Another way to look at this is to note that the total current in R_S is almost constant. As V_{in1} goes up so does the current in the left hand FET. Since the total current is constant that means that the right hand current gets smaller and V_{out} rises.

19.1.2 Gain of the Long-Tailed Pair

We can compute an approximate gain formula for the long-tailed pair using the linearized model of an FET. The general case is quite complicated and so we shall treat a slightly simplified case. According to the full solution the circuit will work best if the two FETs are perfectly matched so that their transconductance curves lie exactly on top of each other. In that case they have the same transconductance, g_m , and the same threshold voltage, V_{Th} . Under those circumstances we can derive the output voltage as a function of the two inputs

$$V_{out} = V_{off} + \frac{R_D \times R_S \times g_m}{1 + 2 \times R_S \times g_m} \times \left[V_{in2} - V_{in1} - \frac{V_{in1}}{R_S \times g_m} \right]$$

where V_{off} is a constant offset voltage that depends on the power supply voltages and the threshold voltage.

Apart from the removable, this has exactly the form of a difference amplifier. The output depends principally upon the difference between the two inputs with a small dependence upon the actual input level. Remembering that we want the common-mode term to be as small as possible we see immediately that we need to choose the FET transconductance and the source resistance as large as possible in order to minimize the common-mode term. As usual, there is a trade-off since increasing R_S decreases the current flowing through the FETs and so decreases g_m . In practice, values of 100k or more are common.

Once we have chosen $2 \times R_S \times g_m \gg 1$ we can simplify the differential gain

$$\frac{R_D \times g_m}{1 + 2 \times R_S \times g_m} \approx \frac{1}{2} \Rightarrow GV = \frac{R_S \times g_m \times R_D \times g_m}{1 + 2 \times R_S \times g_m} \times \frac{R_D \times g_m}{2}$$

Thus the differential gain is just half of that for a common-source amplifier made from the same FET and R_D .

The CMRR is just

$$CMRR = 20 \times \log_{10} \frac{1}{2 \times R_S \times g_m}$$

This looks as though it should be easy to make a high quality differential amplifier by making both R_D and R_S very large. The problem is that large resistors mean small currents flowing in the FETs and thus very small values of g_m and low gains.

Example

I built a long-tailed pair using 2N7000 FETs with $R_D = 10k$ and $R_S = 100k$ and $\pm 15V$ power supplies. With both inputs grounded the voltage on the FET sources was $-2.16V$ so that the total FET current was about $22\mu A$. The drain voltages were $14.4V$ and $14V$ so FETs were not perfectly matched but matched reasonably well. Note that with the quiescent drain voltage at $14V$ you can only have a maximum $1V$ output amplitude! I measured the differential gain to be 3.3 , a rather small value, corresponding to $g_m = 6.6mMho$. The common-mode gain was 0.03 so the $CMRR = 29dB$, not very good at all. Theory predicts the $CMRR = 42dB$, which would be just tolerable. The difference is due to the imperfect matching of the FETs.

The example illustrates the problems of the simple long-tailed pair. The output amplitude range is poor, the differential gain not very large, and the $CMRR$ unimpressive. All of these problems are caused by the low FET drain current. What we really need is a way to have a large R_S and still have a large drain current. One way is to use a very high negative power supply voltage, which is extremely inconvenient. A better way is to replace R_S by a constant current source.

19.1.3 Improved Long-Tailed Pair

Back in section 11.6.2 we met the FET constant current source. This little circuit keeps the current flowing in the drain of an FET almost constant despite large variations in the drain voltage. That means that it has a large slope resistance.

This is exactly what we want, a circuit to pass a fairly large current while exhibiting a large resistance. Figure 19-2 is the circuit of a long-tailed pair with a current source in place of R_S .

The drain current of Q_1 is set by the bias voltage V_B , which is held constant. We can thus tune the drain currents of the active FETs, Q_2 and Q_3 , to any value that we like. This allows us to operate the FETs at higher currents while keeping a large effective resistance. That way we can get higher values of g_m and make the quiescent value of V_{out} lower so that the output swing is larger.

The formulae for the gain and $CMRR$ are just the same for this new circuit except that R_S is now the slope-resistance of the current source. This is almost completely determined by the structure of the FET and by how steady we can hold V_B . Values of hundreds of $k\Omega$ are typical.

Example

I replaced the $100k$ source resistor in the previous example with a 2N7000 current source and adjusted the bias voltage to allow a total of $2mA$ to flow. With both inputs grounded the drain voltages were $7V$ and $2.4V$ so that the mismatch looks worse at this higher current. The new differential gain was about 40 corresponding to $g_m = 7.8mMho$. As we expected, the higher drain current has increased the g_m and thus the gain. The new common-mode gain was only 0.0082 so that the $CMRR$ was a very satisfactory $74dB$. Thus the current source has made a noticeable improvement in G_v and a huge improvement in $CMRR$.

Using a current source in the long-tailed pair makes a huge improvement in the $CMRR$. The only way to improve it further is to be extremely careful about matching the FETs.

Integrated circuit amplifiers offer the best common mode performance since the FETs in them can be made nearly identical. In addition, the FETs can be made extremely close together so that they both operate at the same temperature. Even very small differences in temperature can spoil the matching of two FETs because the threshold voltages are very sensitive to temperature.

19.1.4 Characteristics of the Long-Tailed Pair

Because the long-tailed pair is basically two common-source amplifiers placed back-to-back its characteristics are essentially those of the common-source amplifier.

Remember

$$R_{slope} = \frac{\text{Change in Voltage}}{\text{Change in Current}}$$

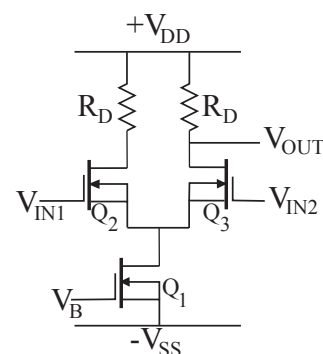


Figure 19-2 Long-Tailed Pair with Current Source

- Because the input signal is applied directly to the gate of an FET the input resistance is very high. At DC it is essentially infinite and it falls with frequency because of the input capacitance of the FET.
- It has a high output impedance that makes it useless for driving loads.
- Its linearity is that of the basic FET so that it is excellent for small signals but gets worse as the size of the output signal increases.
- Because the drain voltages vary, the high frequency response is dominated by the Miller effect and so is similar to that of a common-source amplifier with the same R_D .

Thus the long-tailed pair is an excellent input stage that offers a reasonable amount of gain, high input impedance, decent frequency response, and the all important ability to form the difference of two input signals.

19.2 The FET Source Follower

If we take a common-source amplifier and interchange the rôles of the drain and source then we obtain a 1-FET amplifier called a **source follower** or **common-drain** amplifier.

Again, the name common-drain comes from the way that the input is seen as applied between gate and drain while the output is taken between source and drain. Thus the drain is the terminal common to both input and output.

The operation of this circuit is quite different from the apparently similar common-source amplifier. The basic idea is a simple one. So long as the current that flows in the FET is not large, then the voltage between the gate and source will stay close to the threshold and we shall have, approximately, $V_S = V_G - V_{Th}$. Since the input is applied directly to the gate and the output taken from the source we have

$$V_{out} \approx V_{in} - V_{Th}$$

This gives us a voltage gain of only about 1 but the current gain can be large. At least at low frequencies, the gate current is almost exactly zero while the output current can be quite large. Thus the amplifier can exhibit a large current gain. We call a circuit such as this one, where the output voltage is equal to the input voltage a **follower** or a **buffer**. As we shall see, it is a reasonable first attempt at a power amplifier.

The FET follower has a rather large offset between input and output because the FET threshold voltage is quite large, typically a few volts. In addition, it will not operate for signals less than the threshold and so must be biased to operate with AC signals, much as the common-source amplifier was biased. The circuit is fairly obvious but for completeness here it is.

It is common practice to put a second capacitor, the output capacitor, between the output and the load. This isolates the output from the DC level of V_{out} . This is especially important for loads like loudspeakers that are not designed to have a DC voltage across them.

If the output load is happy to have a DC component to the current through it then it is possible to use the load as R_S and so deliver all the current to the load.

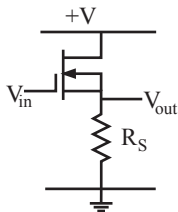


Figure 19-3 FET Source Follower

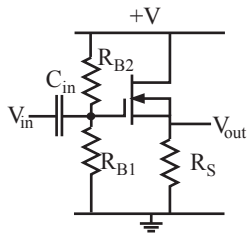


Figure 19-4 AC Coupled Source Follower

19.2.1 Gain of the Source Follower

If the FET gate-source voltage never varied from its threshold value then the simple analysis above would be correct. In reality, however, the gate voltage must increase as the current through the FET increases and so we must do a more careful analysis.

I will use the linearized form of the FET transfer function from earlier to approximate the real behavior. We can make the behavior more realistic by allowing g_m to vary as V_{GS} varies.

The linearized model tells us that

$$I_{DS} = I_{Th} + g_m \times (V_{GS} - V_{Th})$$

That drain-source current must flow through the source resistor, R_S , and so we can write the output voltage, which is the voltage on the source of the FET, as

$$V_{out} = V_S = R_S \times I_{DS} = R_S \times g_m \times (V_{GS} - V_{Th})$$

Now we must remember that V_{GS} in this circuit is NOT the input voltage. Instead, we have

$$V_{GS} = V_{in} - V_{out} .$$

We can substitute this value of V_{GS} into the previous equation to obtain

$$V_{out} = R_S \times g_m \times (V_{in} - V_{out} - V_{Th})$$

Now we have V_{out} on both sides of the equation so we have to rearrange the equation and then we can solve for V_{out}

$$V_{out} + R_S \times g_m \times V_{out} = (1 + R_S \times g_m) \times V_{out} = R_S \times g_m \times (V_{in} - V_{Th})$$

Thus the final gain equation becomes

$$V_{out} = \frac{R_S \times g_m}{1 + R_S \times g_m} \times (V_{in} - V_{Th})$$

If $R_S \times g_m \gg 1$ then we recover the simple equation. However, since the g_m of a typical FET is less than 100mMho = 0.1Mho, that means that we need $R_S > 100\Omega$ to get close to unity gain.

Because the load is connected in parallel with the source resistance, the overall source resistance seen by the FET is the parallel combination of R_S and R_{Load} . Thus the gain is really determined by the load resistance since the combined resistance can be no bigger than R_{Load} . So long as the load has a large resistance and draws only a small current from the follower then we can get a voltage gain as close to 1 as we want. However, if the load draws a large current then we have a small value of R_S and so the gain will be lower.

There is some benefit from the variation of g_m with I_{DS} . Since a smaller R_S will allow a larger DC current to flow in the FET, g_m will get larger as R_S gets smaller and so the gain will not fall as fast as it would otherwise. In practice gains of about 0.75-1 are attained.

Example

I built an emitter follower with a 2N7000 FET and measured the voltage gain for several different values of emitter resistor. In each case I calculated the g_m from the gain formula. Here is what I found.

Table 19-1:

R_S (Ohms)	G_V	g_m (mMho)
10k	0.98	5
1k	0.97	33
100	0.96	240
10	0.74	280

As we expect, the gain stays very close to one for the high values of R_S and only falls appreciably for the very low value of $R_S = 10\Omega$. For values of 100 Ω and more the variation in g_m cancels the effect of changing R_S and the gain is almost constant.

So far all I have discussed is the voltage gain. Since the FET draws no DC current, the DC current gain is infinite. As the frequency increases the gate starts to draw current in order to charge the gate-source capacitance. Because the output voltage follows the input voltage there is very little variation of the gate-source voltage and so very little charging current flows. The effective input capacitance is dramatically reduced by the amplifier's voltage gain behavior. However, there is still an input capacitance and so the input current does rise with frequency and so the current gain falls steadily from its infinite DC value. It remains high so long as the voltage gain remains near 1.

Example

The data sheet maximum input capacitance of the 2N7000 is 60pF. In an source follower with a 1k source resistor the voltage gain was found to be 0.97 and so the effective input capacitance is less than $(1 - G_V) \cdot C_{GS} = 0.03 \cdot 60\text{pF} = 1.8\text{pF}$.

If we drive this amplifier with a 1V sine wave at 1MHz then the output will be a 1V sine wave across 1k corresponding to an output current of 1mA. The input current will also be a sine wave with peak value $2\pi f C_{in} V_{in} = 2\pi \cdot 10^6 \cdot 1.8\text{p} \cdot 1\text{V} = 11\mu\text{A}$. So the current gain will be $1\text{mA}/11\mu\text{A} = 88$.

In practice, the data sheet capacitances seem to be on the high side and the real current gain will probably be even higher.

19.2.2 Linearity of the Source Follower

The source follower can be considerably more linear than the common-source amplifier. The gain equation has g_m in both numerator and denominator so that variations in g_m with the signal level tend to cancel out. However, the cancellation is not complete and the source follower does suffer from some distortion.

In the common-source amplifier V_{GS} was derived directly from V_{in} and so a large amplitude V_{in} would produce large variations in V_{GS} . In the common drain amplifier V_{GS} is the difference between the input and the output and so is nearly constant so long as the gain is very close to one. In the common-source amplifier the non-linearity increased with the voltage of the output signal. In the source-follower the non-linearity increases with the output current.

As the current through the FET increases, so must the V_{GS} increase. Thus the difference between V_{in} and V_{out} increases as the FET current increases and so the gain falls as the FET has supply more current.

The source-follower has excellent linearity so long as the total FET current is small or so long as the changes in FET current are small compared to the average current. This means that the linearity is best when the source resistance is high and very little current is fed to the load. When driving a low resistance load the linearity suffers.

19.2.3 Frequency Response of the Source Follower

If a source follower is AC coupled then the low frequency response is controlled by the high pass filter formed between the input coupling capacitor and the bias resistor network. The analysis is exactly the same as for the common source amplifier. If the amplifier is DC coupled then the gain extends all the way down to DC.

The high frequency response is dominated by the low pass filter formed by the output capacitance in parallel with the source resistance, R_s . That is, the cut-off frequency is given by

$$f = \frac{1}{2\pi \times R_s \times C_{DS}}$$

Unlike the common source amplifier, the source follower does not suffer from the Miller effect. Because the gate source voltage is almost constant, the feedback capacitance (C_{GS} in this case) does not conduct and so can be ignored. Thus the source follower operates happily to much higher frequencies than the common source amplifier. Of course, it also provides no voltage gain and so there is no way to use this configuration in place of the common source amplifier.

Example

The 2N7000 data sheet lists the maximum value of the drain-source capacitance as 25pF. Using that value we find that our best source follower, with source resistance of 10k, should have a high frequency cut-off no lower than

$$f = \frac{1}{2\pi \times R_s \times C_{DS}} = \frac{1}{2\pi \times 10k \times 25p} = 600kHz$$

My own measurements put the cut-off at least 6MHz, much higher than predicted. This suggests that the data sheet values for the output capacitance are rather generous maxima.

19.2.4 Source Follower Output Drive

It seems at first glance that the source follower should be much better at supplying output current than the common-source amplifier was. The output current now flows through the FET, which we know has a low resistance when turned on. However, if we look at the linearized model of our circuit we shall find that we are not that much better off. Here is the equivalent circuit of our linearized mode

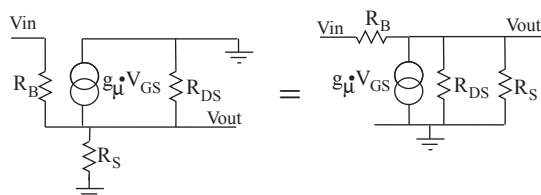


Figure 19-5 Linearized Model of Source Follower

Once again, the output consists of the current source in parallel with the combination of the internal drain-source effective resistance and the output resistor. Once again, the output resistance is equal to the parallel combination of R_{DS} and R_S and thus, essentially equal to R_S .

An amplifier will see a 50% drop in gain when the load resistance becomes equal to the amplifier output resistance. Thus we need to keep R_D significantly less than the minimum load resistance. We can do this at two costs.

1. The voltage gain will be somewhat reduced according to the gain formula. However, this is not a serious problem for any but the lowest resistances.
2. The FET and source resistor will both carry a large standing current even in the absence of any signal. This will lead to heating in the components and to a significant waste of energy.

The actual current delivered to the load is then easily estimated. The largest AC voltage that can be delivered is one half of the supply voltage because the output has to be biased midway between +V and ground to give a symmetric output. The quiescent current in the FET and resistor is then

$$I_{DS(Q)} = \frac{V_Q}{R_S} = \frac{V}{2 \times R_S}$$

This is also the maximum possible output current but a more conservative value would be to limit the output current to 1/20 of this value.

Example

In the 2N7000 amplifier of the previous example both the FET and the load resistor were biased to have 7.5V across them in the absence of any signal. For the low values of R_S the large I_{DS} currents lead to large power dissipations. In the 100Ω case we have

$$P = I \times V = 7.5V \times 0.075A = 0.56W$$

Thus both FET and resistor must dissipate half a watt of power. By contrast the power delivered to the load using our limit of 1/10 of the maximum current is

$$P_{LOAD} = I_{LOAD} \times V_{LOAD} = \frac{V}{20 \times R_S} \times \frac{V}{2} = \frac{15}{2000} \times 7.5 = 0.056W$$

So we deliver to the load only one tenth of the power lost in each of the FET and load resistor.

19.3 The Complementary Source Follower

As we have seen, the source follower is only somewhat better at driving low resistance loads than the common source amplifier. If we want to drive a really low resistance load, such as the 8Ω load of a loudspeaker, then we need to do something else.

The real limitation of the previous circuits is their output resistance. In each case the output resistance is equal to the source (drain) load resistance and we cannot decrease that too far without hurting the gain and dissipating excessive amounts of heat. What we want is a very low valued output resistor that will allow there to be a large voltage across it without much current flowing. That sounds impossible but it is actually a description of an FET. If we replace the source resistor with a second FET then we can improve the working of the circuit.

There are a large number of variations on this design but the most popular is the complementary follower shown in Figure 19-23 on the right. This circuit operates is complementary in two ways. First, it uses complementary FETs, both an NFET and a PFET. Second, it usually operates from complementary power supplies, equal and opposite power supplies ±V.

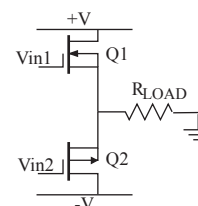


Figure 19-6 Complementary Source Follower

I have shown it with two input voltages, V_{in1} and V_{in2} . These can be derived in several different ways from a single input voltage. The methods differ in their complexity and in the quality of the output signal.

The simplest method of supplying the input is to tie them together so that $V_{in1}=V_{in2}$ (Figure 19-24). This is also the least satisfactory as we shall see.

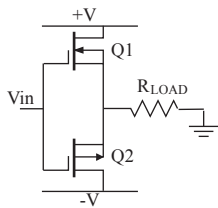


Figure 19-7 Unbiased Complementary Source Follower

Consider driving the circuit with a sine wave of several volts amplitude.

When the input voltage is greater than 0 then we expect that the output will also try to be greater than zero. In that case the PFET, Q2, will be turned OFF because it requires its gate to be negative with respect to its source in order to turn on. The NFET may or may not be turned on. If the input voltage is greater than the threshold voltage for the NFET then Q1 will turn and we shall have a perfectly ordinary NFET source follower with the load playing the part of the source resistance. The advantage here is that all of the FET current passes to the load and none is wasted. If the input is less than the threshold then Q1 will also be turned off and no current will flow to the output so $V_{out} = 0$.

Exactly the opposite happens when V_{in} is less than zero. Now Q1 is turned off because its gate is not more positive than its source. Again, if V_{in} is less than the (negative) threshold for Q2 then Q2 will also be off and $V_{out}=0$. But if V_{in} is negative and large enough then Q2 will turn on and again we shall have a source follower with the load as its source resistance.

Thus there are three ranges of operation

$$\begin{aligned} V_{in} &\geq V_{Th} & V_{out} &= V_{in} - V_{Th} \\ V_{Th} &> V_{in} > -V_{Th} & V_{out} &= 0 \\ V_{Th} &\geq V_{in} & V_{out} &= V_{in} + V_{Th} \end{aligned}$$

This produces a rather distorted output. For example here is the output recorded from a 2N7000/ZVP4105 FET pair driving a 100Ω load (Figure 19-25). The input was an 8V peak-peak sine wave at 100Hz.

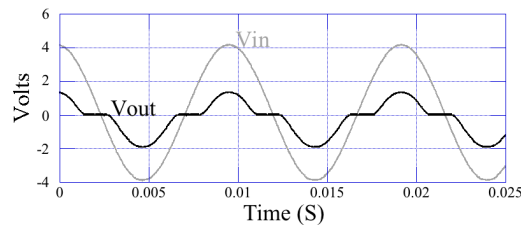


Figure 19-8 Unbiased Complementary Source Follower Output

This amplifier clearly exhibits a particularly severe form of crossover distortion. You can clearly see both the flat regions where neither FET is conducting as well as the peaks and valleys where one or the other FET is turned on. As usual for a source follower, the output is in each case lower than the input by the threshold voltage. In this case you can also see that the threshold voltages are not perfectly matched. The PFET has a somewhat lower threshold (about 2V) than the NFET (about 2.3V).

This amplifier has little to recommend it apart from its superior output drive capability. The output is very distorted and significantly smaller than the input.

19.3.1 Biasing the Complementary Source Follower

We can dramatically improve the behavior of this amplifier by biasing the inputs, much as we had to bias both the common drain and source follower amplifiers. This time we want to arrange that there is no time when both FETs are turned off. We can do this by putting a fixed voltage between the two gates. There are various ways to do this but we usually need only the simplest, a Zener diode. If we add a Zener diode and pair of resistors to set the current flowing the Zener then we get the circuit of Figure 19-26. First let us note a few things about the circuit.

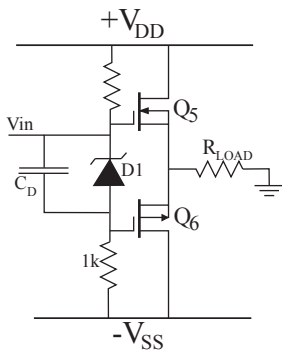


Figure 19-9 Zener Biased Complementary Source Follower

- The input is shown connected to the upper FET gate but it could just as well have been connected to the lower one. Indeed, you will sometimes see two Zeners used in series with the input connected to the midpoint.
- The capacitor, C_D , across the Zener is not essential but it improves the high frequency response of the circuit by making it very difficult for the Zener voltage to change rapidly.
- The Zener bias current comes through the resistors R1 and R2. It is not really constant as the voltages across the resistors vary with the signal. This will mean that the Zener voltage is not as constant as it could be. You can do better using constant current sources in place of R1 and R2.
- The Zener voltage must be chosen at least as large as the sum of the threshold voltages. I shall say more about this later.

So long as the Zener voltage is larger than the sum of the threshold voltages, there is no time in the cycle when both FETs are turned off. Now the upper FET is on any time $V_{in} > V_{Th}$ but the lower FET does not turn off until $V_{in} - V_D > -V_{Th}$. This means that, while the output voltage is always less than V_{in} by the threshold voltage of the NFET, there are no breaks in the output and we always have $V_{out} = V_{in} - V_{Th}$ to reasonable accuracy. You can see this in the next figure (Figure 19-27), which was taken from the same amplifier as Unbiased Complementary Source Follower Output but with a Zener added as in Zener Biased Complementary Source Follower. Here the amplifier is driving a 2V amplitude sine wave into a 100Ω load at 100Hz.

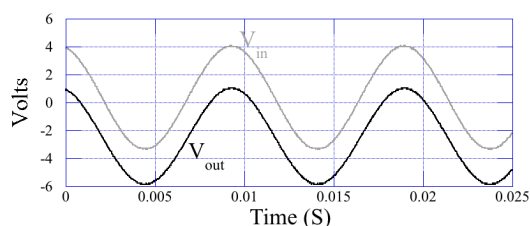


Figure 19-10 Biased Complementary Source Follower Output

The improvement is dramatic. The output is still offset from the input by the NFET threshold voltage but it now follows the input perfectly. Subtracting the output from the input reveals a very small difference signal that looks like a nearly pure sine wave. This is because the gain is a little less than 1 (I measured it to be 0.95, very consistent with the source follower).

19.3.2 Linearity of the Complementary Source Follower

The complementary source follower is rather less linear than the simple source follower. Now you not only have the non-linearity of each FET to contend with but also any mismatch between the two FETs. The two NFET and PFET share a common source resistance, the load, and so any difference in transconductance between the two FETs will appear as a difference in gain for +ve going outputs compared to -ve going.

A perfectly linear unity gain buffer amplifier would have a plot of V_{out} vs. V_{in} as a straight line with a slope of 1. The variations in FET gain with output current distort that line for a real amplifier. Moreover, as with the simple source follower, the problem gets worse and worse as the output current increases. The higher the output current, the higher the V_{GS} needed to turn the FET on sufficiently and so the lower the gain. The worst problems appear as the output goes from +ve to -ve. Imperfections here produce crossover distortion as they did in the unbiased case but the effect is very much smaller.

We saw in Biased Complementary Source Follower Output the output of a biased complementary source follower driving a fairly high resistance of 100Ω. The distortion was not visible to the naked eye. If we decrease the load to 10Ω then the peak output current rises to 100mA at 1V amplitude. Under these conditions we can see the distortion quite clearly (Figure 19-28).

Info Seeing no distortion is not really enough for some purposes. High-quality audio power amplifiers need to have distortion levels that are too small to see. Unfortunately they are also too small to measure without specialized, expensive, apparatus. The best audio power amplifiers offer less than 0.001% distortion across the whole audio band while delivering tens of watts to a loudspeaker. They use similar output circuits to this one but with special care taken to minimize distortion.

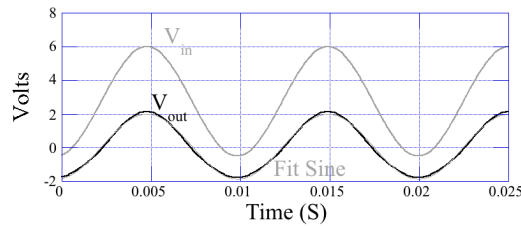


Figure 19-11 Complementary Source Follower Driving 10 Ω Load

The output voltage, the black line, lies about 3V below the input voltage, which is reasonable for these FETs when delivering 200mA to the load. It is quite a bit smaller than the input voltage. The gain has fallen to about 0.62 as the output current has risen.

The output is a rather good copy of the input but the distortion is clearly visible. The thin gray line is a pure sine wave fit to the output. It shows what the undistorted output should like. The real output is too high at the peaks and valleys and too low near the zero crossings. This is because the gain is not constant throughout a cycle. Indeed, if we plot V_{out} vs. V_{in} then we see the non-linearity quite clearly (Figure 19-29).

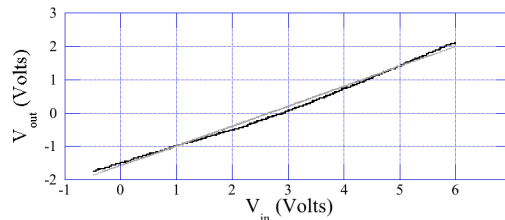


Figure 19-12 Gain Plot for Complementary Source Follower w/ 10 Ω Load

Again, the heavy black line shows the measured data while the thin gray line is a straight line fit to the data. The input is offset by 3V (the FET threshold voltage) from the output because the input was applied to the upper end of the bias Zener. It means that the real crossover point is where V_{out} changes sign and $V_{in} = 3V$.

The cause of the distortion is clearly visible. When V_{out} is positive the NFET is conducting and the slope of the line is significantly greater than when the PFET is conducting (V_{out} is negative).

There is not much we can do about this non-linearity except try to choose two FETs that match better than these two. There are very considerable differences in characteristics between two FETs of the same type and so careful selection of the FETs can improve the matching between the two. This is very time consuming and is not normally practical. The only other way to improve the linearity is through the use of feedback as we shall see in the next chapter.

19.3.3 Complementary Source Follower Output Drive

Because all the FET current now flows to the load this circuit is as efficient as possible at delivering a high current. The output is limited by three things:

1. the amount of current available from the power supplies,
2. the maximum power and current that the FET can handle without damage, and
3. the drive to the FET.

The first two are fairly obvious. We simply need to choose the power supplies and output FETs large enough to handle the load that we want to control. The third is a little more subtle. Because the transconductance of the FET is not infinite, more gate-source voltage is needed to produce more output current. Thus the input signal needs to be enough bigger than the output to supply the extra drive needed.

Example

Consider the transconductance curve for the 2N7000 given in Figure 11-7. From that we see that gate-source voltage of about 2.5V is sufficient for very small drain currents but that V_{GS} needs to be about 3.75V to produce a drain-source current of 100mA. Thus, if we desire to produce a 1V amplitude sinewave output then the input must rise to 3.5V if the load current is only about 1mA but must rise to 4.75V if the load current is to reach 100mA.

Thus we must choose power supplies that are not only sufficient to drive the desired current into the load but also large enough to provide the extra drive needed for the FETs.

Consider trying to drive 1W RMS into an 8Ω loudspeaker. The peak power needed is 2W and that allows us to compute the power supply current needed.

$$P=I^2 \times R \rightarrow I=\sqrt{(P/R)}=\sqrt{(2/8)}=0.5A.$$

So we shall need to supply 0.5A into 8Ω which will need a voltage of $V = I \times R = 0.5 \times 8 = 4V$ at minimum. Now, we need an FET that can handle a steady current of 0.5A which is more than our favorite 2N7000/ZVP4105 pair can tolerate. The IRFD014/IRFD9014 FETs can handle currents up to 1.7A and dissipate 1W each and so make a good choice. According to its data sheet, it will take a gate-source voltage of about 4.3V to allow 0.5A to flow through the FET so that our minimum power supply is really 8.3V and we would probably use 9-10V for safety.

This system can deliver 1W RMS (2W peak) to an 8Ω load. The cost is the power dissipated in the FETs. The maximum current through either FET is 0.5A and it occurs when the voltage across the FET is $9V-4V = 5V$. Thus the peak power lost in the FET is 2.5W, slightly more power than was delivered to the load. The RMS power in the FET is about 1.25W, more than the bare FET can dissipate and so heatsinks will be needed on the FETs. A higher power amplifier would be rather more efficient because the voltage needed to drive the load will increase much faster than the extra voltage (V_{GS}) needed to drive the FET.

Note We can get clever and use a higher voltage power supply for the signal that drives the FETs and a lower supply for the FETs. In that case we could lower the FET power supply to about 4.5V instead of 9V and then the power lost in the FETs would fall to 0.25W.

In summary, the complementary source follower is an excellent choice for a power amplifier. With suitable choice of FETs and power supplies it can drive large currents into small load resistances.

19.3.4 Frequency Response of the Complementary Source Follower

Like the ordinary source follower, the complementary source follower does not suffer from the Miller effect and so has very good high frequency response. The frequency response is set by the load resistance in parallel with the output capacitance of one of the FETs. The low values of both output resistance and capacitance mean that the circuit can operate to very high frequency. Because the amplifier has a voltage gain less than 1, the circuit does not have a tendency to oscillate even though it has a very high bandwidth.

Example

A 2N7000/ZVP4105 complementary source follower driving a 1V amplitude sinewave into a 10Ω load, peak current 100mA, showed no decrease in gain up to the 3MHz limit of a signal generator.

19.4 A Complete Multi-Stage Amplifier

We now have all the building blocks needed to construct a realistic amplifier. We will use a long-tailed pair with a current source for the input, a single common-source gain stage, and a complementary source-follower output stage. The circuit looks very much the way we would expect (Figure 19-30).

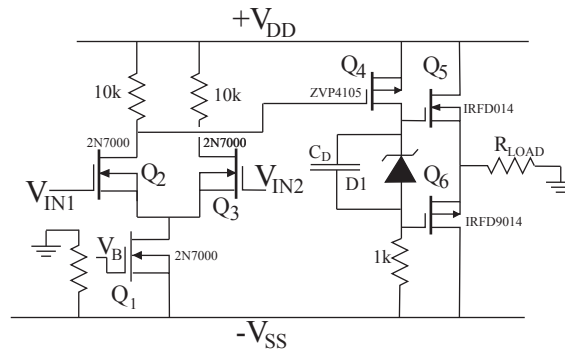


Figure 19-13 3-Stage Amplifier

Transistors Q2 and Q3 form the input long-tailed pair with the constant current source Q1. The bias voltage, V_{B1} , for Q1 is set to allow a total of 1mA of current to flow in the long-tailed pair.

The output is taken across the 10k drain resistor of Q3 so that V_{in2} will be the inverting input and V_{in1} the non-inverting input. This stage has a gain of 15 and a CMRR of 70dB.

The main gain stage is made up of PFET Q4 with its 1k drain resistor. This stage has a gain of 33. This is not very high because 1k is a small value of load resistor. If we increase the 1k resistor, however, the voltage drop across the resistor gets rather large and starts to limit the output voltage swing.

With about 5mA flowing down the bias chain there is 5V across the 1k resistor. That would limit the negative output voltage to $-15V + 5V + 2.5V$ for the FET threshold = $-7.5V$. With a $-15V$ supply this is not a very good output swing. You can make a major improvement to both the gain of this stage and to the output swing by replacing the 1k resistor with a constant current source. That way the Zener bias stays more constant and the gate of the output PFET can swing almost all the way to $-15V$, greatly increasing the output swing.

Q5 and Q6 form a complementary source follower output stage whose bias voltage comes from a 6.8V Zener diode, D1. The IRFD014/IRFD9014 FET pair was chosen because it can handle rather more current than the 2N7000/ZVP4105 pair. This output stage has a gain of 1 at low currents and can easily drive a 10V sine wave into a 10Ω load resistor.

It is a little hard to recognize the simple common-source gain stage because it has been mixed into the bias chain for the output pair. The Zener diode bias generator, D1 and C_D , insert a fixed voltage drop between the drain of the FET and its 1k load resistor but do not interfere with its operation. Voltage changes at the drain of Q4 are perfectly echoed at the top of the load resistor and so, from the point of view of the signal, the two points act as if they were connected together.

I used a PFET for the main gain stage so that the offset voltage of the long-tailed pair can act as the bias voltage for Q4. The current in Q1 was chosen to make the voltage across the load zero when both inputs were grounded.

The overall amplifier has a gain of $15 \times 33 = 490$ and a bandwidth of about 30kHz when driving a high impedance load. The gain falls quite a lot as the load gets smaller and smaller and the gain of the output stage falls. With a 10Ω load the gain was only 184. It is rather difficult to work with because small variations in temperature upset the delicate DC bias and cause the output voltage to drift around by many volts. However, I was able to record some very nice data (Figure 19-31) by letting the temperature stabilize while I kept the signal level constant.

Note The gain of the amplifier is now so high that I had to expand the input by a factor of 100 to show the input and output on the same graph.

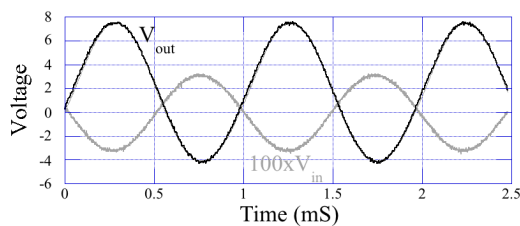


Figure 19-14 Output of 3-stage amplifier

The amplifier was driving a 10Ω load so the peak output currents are about $+700\text{mA}$ and -400mA . The amplifier is delivering about 1.5W RMS to the load.

The output of the amplifier is the solid black line and the thin gray line (look closely, its there) behind it is the best fit sine wave. The fit is extremely close and so the linearity of the circuit is very good. The actual distortion is only visible on the ascending portion of the output, just after it crosses zero. This is a small amount of crossover distortion.

We can get a better look at the non-linearity if we plot the difference between the actual output and the best fit. this difference is too small to plot on the same graph as the output so I multiplied the error by 50 and got Figure 19-32.

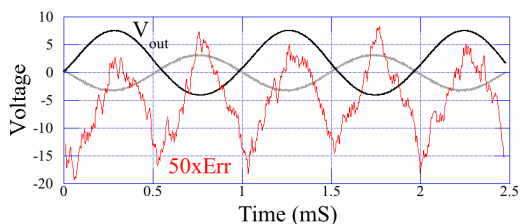


Figure 19-15 Open Loop Output Error

The error is largely at twice the frequency of the signal and is a little more than $1/50\text{th}$ of the amplitude of the output. All of the largest error values occur near the zero crossings, as we would expect for cross-over distortion.

19.4.1 Stabilizing a Multi-Stage Amplifier

As mentioned above, it is extremely difficult to keep the DC bias of a multi-stage amplifier constant. The result is that the output offset drifts around, sometimes by many volts. This is clearly unsatisfactory.

One cure for the bias problem is to AC couple the amplifier. Then each stage can be biased independently of the others and tiny variations in the first stage bias will not be passed along through the amplifier, getting bigger as they go. This solution used to be very popular but it has fallen out of fashion. One obvious disadvantage is that the gain of the system must fall off below some low frequency, although that frequency can be made quite low by choosing large enough coupling capacitors. The other disadvantage, which is much less obvious, is that the modern solution has extra benefits in addition to stabilizing the bias.

The modern solution is to subtract a portion of the output signal from the input. Essentially, this allows the amplifier to monitor its own output offset and adjust the bias to keep the offset at zero. The cost is a decrease in the total gain of the system but the benefits include not only bias stabilization but also greatly improved linearity, bandwidth, and output impedance. This technique is called **negative feedback** and it forms the major topic of the next two chapters.

We can apply the modern solution to our amplifier by adding two resistors as shown in Figure 19-33.

Info We call this form of feedback negative because it involves subtracting the feedback from the original signal. Negative feedback makes decent amplifiers into much better amplifiers. If we add the feedback instead of subtracting it then we get positive feedback. This has the opposite effect, it makes amplifiers oscillate. That is, it makes them generate signals all by themselves, even with no input at all.

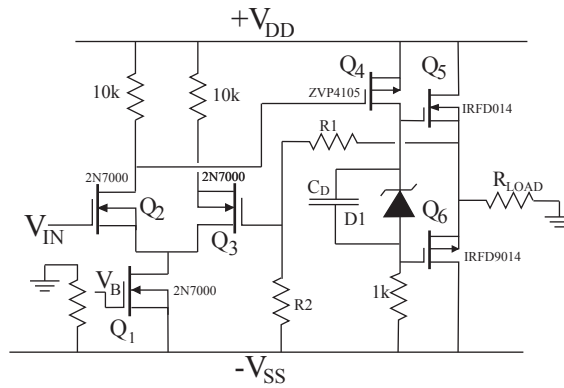


Figure 19-16 3-Stage Amplifier with Feedback

As we shall see in the next chapter, the gain of the circuit is now set by R1 and R2 according to the formula

$$Gain = \frac{R1 + R2}{R2}$$

This equation is valid so long as the basic gain of the bare amplifier, what we call the **open-loop gain**, is greater than this number.

I added resistors R1=200k and R2 = 20k to the amplifier to give it a theoretical gain of 11 and re-measured the behavior. I have compared the behavior with and without feedback in the following table.

	Open Loop	Closed Loop
Gain (No Load)	490	10.7
Gain (10Ω Load)	184	10.4
Bandwidth (No Load)	30kHz	600kHz
Bandwidth (10Ω Load)	10kHz	400kHz
Output Offset	Wildly Unstable	Stable 0.2V

Table 19-2: 2

So, at the cost of about a factor of 50 in gain we have made the amplifier perfectly stable, made it insensitive to variations in load, and increased the bandwidth by at least a factor of 20. In addition, the amplifier is now much more linear. Here is a plot of the output driving a 10V p-p sine wave into a 10Ω load at 1kHz (Figure 19-34).

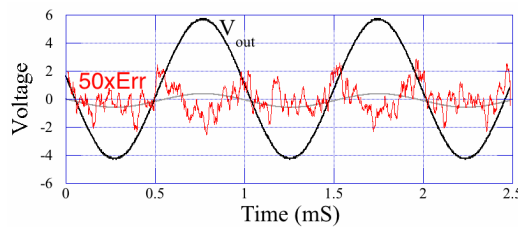


Figure 19-17 Closed Loop Output with Error

The input is the small grey line and the output the black line. There is actually a grey line under it to show the ideal output but they agree so perfectly that the grey line is not visible anywhere. In order to see the error I had to magnify it by a factor of 50 and that is the jagged line. It is almost pure noise but there is a slight indication of remaining crossover distortion in the peaks at the zero crossings of the output. It is hard to measure under the noise but the distortion is at least a factor of 10 smaller than it was in the open loop amplifier.

So the addition of feedback to our 3-stage amplifier has

- improved linearity,
- made the output stable,
- increased the bandwidth, and
- decreased the effect of output load.

Not surprisingly, feedback is used in almost all modern amplifiers and we shall study its effects and some of the neat circuits that it allows in the next two chapters.

Chapter 20: The Ideal Operational Amplifier

20.1 Introduction

Before digital computers became small enough and cheap enough to be incorporated into almost any kind of electronic hardware, there was a niche for electronic systems that could model and control real world systems. Probably the best known example is the automatic pilot that can hold an aircraft in level flight on a predetermined course but systems have also been in such diverse places as chemical engineering plants and the design of new ship hulls. These systems are all examples of Analog Computers, computers that represent real world variables by continuously varying voltages instead of by binary numbers. At the heart of such analog computers are circuits to perform the basic arithmetic operations, addition, subtraction, multiplication, integration, etc. These circuits in turn are built around special amplifiers called **Operational Amplifiers** after the mathematical operations that they perform. Originally these amplifiers were very expensive circuits built first from vacuum tubes and then from transistors but in the early 1970s a revolution in electronics arrived with the first useable integrated circuit operational amplifiers. These amplifiers were so good, so cheap, and so easy to use that they quite quickly found uses in all kinds of low frequency electronic apparatus. By end of the decade integrated circuit operational amplifiers had come to dominate low frequency analog electronics and had replaced discrete transistor amplifiers in all but high power and high frequency applications.

20.2 The operational amplifier

An operational amplifier is a very high-gain DC-coupled amplifier with differential inputs. That means that

- It can amplify signals all the way down to zero frequency, static unchanging voltages.
- It has two inputs; one of which is inverted in the sense that when the input goes more positive, the output goes more negative.

Figure 20-1 shows the symbol for an operational amplifier. The triangle is the standard symbol for an amplifier. This one has two inputs, V_{in+} and V_{in-} , and one output, V_{out} . It also has two wires for the power supply, $V+$ and $V-$. These are usually connected to a split power supply that provides both positive and negative voltages of the same magnitude, e.g. +15V and -15V or +12V and -12V. There are some op-amps made, chiefly for battery operation, that only need a single positive power supply but they are the exception. Note that the amplifier is not explicitly connected to ground. However, the power supplies are connected to ground so the amplifier still has a good ground reference and all the voltages in the circuit are measured relative to ground.

When we analyze op-amp circuits we usually ignore the power supply inputs. They are there only to provide the power to make the circuit work and usually do not affect the output voltage. In fact, we usually don't even show the power supply connections in op-amp circuits, although they must always be present or the amplifier *will not work*.

Ignoring the power supply connections means that we can treat the circuit as a 3-terminal device with two inputs and one output. A further result of ignoring the power supply connections is that we usually just label the inputs V_+ and V_- rather than the longer V_{in+} and V_{in-} . I will mostly use the shorter forms from now on.

The operational amplifier has become such a popular component because of its extreme simplicity. The inputs have a very high input impedance and so draw very little current, little

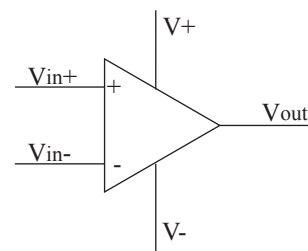


Figure 20-1 Op-Amp Symbol

Info Noise on the power supply lines, especially noise that is related to changes in the output of the op-amp, is a serious problem for real op-amps. To remove this noise you must always put a decoupling capacitor of about 0.01 μ F value from each power supply lead to ground. These capacitors will rarely be shown in circuit diagrams but they must *always* be present!

enough to be ignored in most cases. The output is a very low impedance and can source or sink currents up to about 20mA. Finally, the transfer function is extremely simple

$$V_{out} = G \times (V_+ - V_-)$$

where the gain G is a very large number, at low frequencies 200,000 is typical. Because of the minus sign in this equation we call the V_- input the **inverting input** and the V_+ input the **non-inverting input**. The output is in phase with signals applied to the non-inverting input but 180° out of phase with ones applied to the inverting input.

While the op-amp is a very simple device it is not a very good amplifier! The gain is very strongly dependent on frequency and is not necessarily very linear. A nice sinewave input may be quite distorted by passing through an op-amp used alone. Fortunately, there is a simple trick that allows us to throw away gain, of which we have too much, in exchange for dramatic improvements in the quality of the amplifier. That trick is called **Feedback**.

20.3 Feedback

Feedback is the trick of taking a portion of the output of a circuit and mixing it into the input. There are two ways to do this. You can add the feedback signal to the input or you can subtract it from the input.

Info Negative feedback was invented in 1928 by Harold S. Black at a time when vacuum tubes were just being perfected and getting a gain of 5-10 out a circuit was seen as a major triumph. The idea that you might throw all that hard earned gain away with negative feedback was ridiculed and his patent application was "treated in the same manner as one for a perpetual-motion machine" (IEEE Spectrum, December 1977). Time passed and gain became cheap. Negative feedback was universally adopted as the cure for practically all the ills to which amplifiers are subject. Harold Black's ridiculed idea now lies at the heart of almost all amplifier and signal processing systems.

- If you add the feedback, then an increase in the output results in an increase in the input. That increase in the input results in an increase in the output, which results in a further increase in the input and so on until the amplifier goes wild. This is called **positive feedback**.
- If you subtract the feedback, then an increase in the output results in a decrease in the input so the output goes down; the system is made more stable. This is called **negative feedback**.

Negative feedback is used in nearly all op-amp circuits. It dramatically improves the performance of a circuit at the expense of reducing its gain. Positive feedback is used in a very small number of op-amp circuits, chiefly oscillators (circuits that make their own AC signals) and comparators. It occasionally appears in circuits where it is not wanted and causes them to behave in a whole host of undesirable ways. We are concerned almost exclusively with negative feedback.

Figure 20-2 is a schematic view of how negative feedback is applied to an amplifier. It will help us to understand how negative feedback works.

Info The funny circled-cross thing is a summing point. It adds together V_{in} and $-f \cdot V_{out}$. The input circuitry of the op-amp performs this function.

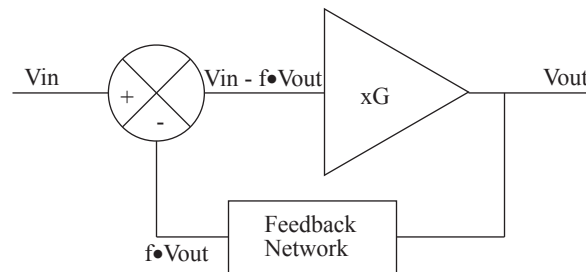


Figure 20-2 General Feedback Circuit

The feedback network takes a fraction f of the output voltage, $f \cdot V_{out}$, and feeds it back to the input. There it is subtracted from the input voltage, V_{in} , to make a modified input signal, $V_{in} - f \cdot V_{out}$, that is amplified by the amplifier to make the output voltage, V_{out} .

$$V_{out} = G \times (V_{in} - f \times V_{out})$$

Now the output voltage depends on both the input and itself. We can solve this equation to find how the output depends on the input.

$$V_{out} + G \times f \times V_{out} = G \times V_{in}$$

$$V_{out} = G \times \frac{V_{in}}{1 + G \times f} = \frac{V_{in}}{f + \frac{1}{G}}$$

Now the one thing that we know about the gain, G , of the bare amplifier is that it is very high. That means that $1/G$ is a very small quantity, often a very, very, very small quantity. So long as $1/G \ll f$ we can simplify the equation to find

$$V_{out} = \frac{1}{f} \times V_{in}$$

So the output is proportional to the input and the new gain depends *only* upon the feedback network. That means that, even though the original amplifier may have been a poor thing—with an inconstant gain and poor linearity—the final amplifier can have a very constant, very linear gain. If the feedback network is a resistive divider, then the fraction f is a constant real number regardless of frequency. That means that amplifier will have a constant, or linear, gain.

The feedback circuit forms a loop around the amplifier and leads to the following terms. The gain of the underlying amplifier is called the **open-loop** gain. It is the gain that the circuit would have if we disconnected the feedback loop. Because of this, we call the gain of an operational amplifier itself open-loop gain. By contrast, the gain of the complete circuit is called **closed-loop** gain.

20.4 The non-inverting amplifier

An operational amplifier already has the subtracting circuit built into its input stage so it is easy to take our generic feedback system and apply it to our op-amp to give us a new amplifier (Figure 20-3).

We can analyze this circuit in exactly the same way as the generic one, so long as we remember the voltage divider equation. We start from the op-amp equation.

$$V_{out} = G \times (V_{+} - V_{-})$$

Then we apply the voltage divider equation to $R1$ and $R2$ to find

$$V_{-} = \frac{R1}{R1 + R2} \times V_{out}$$

Noting that $V_{+} = V_{in}$, we substitute into the op-amp equation and find

$$V_{out} = G \times \left[V_{in} - \frac{R1}{R1 + R2} \times V_{out} \right]$$

We can solve that for V_{out} to find

$$V_{out} = \frac{R1 + R2}{R1} \times \left[\frac{V_{in}}{1 - \frac{R1 - R2}{G \times R1}} \right]$$

So long as $G \times R1 \gg R1 + R2$, the equation reduces to

$$V_{out} = \frac{R1 + R2}{R1} \times V_{in}$$

Thus the closed-loop gain of the circuit is just

$$Gain = \frac{R1 + R2}{R1}$$

Now the gain depends only on the value of two resistors and we know that resistors are very well behaved components. Their values do not depend on the frequency or size of the voltages across them. They do depend slightly on temperature but even that variation cancels so long as both resistors are made of the same material and are at the same temperature. The behavior of the amplifier is extremely well controlled, so long as the open-loop gain \gg the closed loop gain.

What does the constraint really mean? Well, first it means that you can't get too much gain out of the amplifier. Since G is typically $> 10^5$ this is not usually a problem. Second it means that above some frequency the amplifier ceases to work properly. As we shall see in Chapter 21, the open-loop gain of an op-amp falls rapidly as the signal frequency increases. Above some frequency the gain will have fallen so far that our gain condition is no longer valid. After this the amplifier is no longer well behaved. We treat this as the maximum useful frequency of the amplifier. The lower the closed-loop gain, the higher the maximum useful frequency and so

Info Resistors and Capacitors are very nearly ideal components. The current in a real resistor or capacitor behaves in almost the same fashion as that in an ideal resistor or capacitor, at least until the frequency gets up into the region above 50MHz. Thus circuits whose behavior is controlled solely by resistors and capacitors operate as nearly as possible in the way that theory predicts.

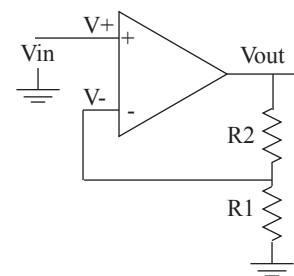


Figure 20-3 Non-Inverting Amplifier

Warning I have shown neither the power supply connections to the op-amp nor the noise suppressing capacitors on those lines. This is a common practice that makes diagrams easier to read. However, the connections and capacitors **must** still be there. The op-amp will not work without power and it will often not work well without the capacitors.

Note Because the input signal is fed directly into the non-inverting input of the op-amp, the input impedance of the whole circuit is extremely high. It is easy to have an input impedance of tens of $M\Omega$. This is very useful for some circuits, such as the amplifiers in electrocardiographs, which have to work with high impedance signal sources. However, it also means that these amplifiers are easily affected by noise pickup. A short piece of wire attached to the input will act as an antenna and will pick up mains frequency interference from all the electrical apparatus in the area so great care has to be taken to shield these amplifiers carefully.

the greater the bandwidth of the amplifier. Thus, while a gain of 1000 amplifier might only work well up to about 2500Hz with a particular op-amp, a gain of 10 amplifier would perform well up to 250kHz.

20.5 The Golden Rules

When an op-amp is connected in a circuit with negative feedback, its behavior is extremely well summarized in two rules, often called the Golden Rules. They are

Note Strictly speaking the Golden Rules are both false. However, they are such very good approximations in almost all cases that we treat them as if they were true!

1. The inputs draw no current
2. The inputs are at the same voltage

The first rule is quite simple. The inputs of real op-amps do draw current but the amount is so small compared with most signals that we can normally make this simplifying assumption.

The second rule is much less clear. First of all it is never really true. Since the output voltage $V_o = G \times (V_+ - V_-)$, if $V_+ = V_-$ then the output is zero. Obviously that is rarely true of a working amplifier. However, the maximum output voltage from an op-amp is equal to its power supply voltage, which is usually 15V. So, the maximum difference between V_+ and V_- is only $15V/10^5 = 150\mu V$. That is so small, compared to the normal signal levels that we encounter, that it is usually safe to ignore it.

Although the first rule is always true, the second rule only applies if the circuit is functioning properly. For example, if we try to make the output of an op-amp go to 20V when the power supplies are only $\pm 15V$ then the circuit fails to function properly. The output will try to go to 20V but will get stuck a little below 15V. It will **saturate**. When the circuit is saturated all bets are off; the input voltages need not be equal and probably won't be. As soon as the input alters enough that the output ceases to saturate, the circuit will resume working and the rule will again be true.

These two rules can be used to analyze the behavior of almost all op-amp circuits. We will use them to analyze the rest of our circuits in this chapter.

20.5.1 The Non-Inverting Amplifier Revisited

First I shall re-analyze the non-inverting amplifier of Figure 20-3 using the golden rules. We start from the facts

$$V_+ = V_{in}$$

so that

$$V_- = \frac{R1}{R1 + R2} \times V_{out}$$

and substitute these into the second Golden Rule to find

$$V_{in} = \frac{R1 + R2}{R1} \times V_{out}$$

so that we get the same gain equation as before, only with a lot less work.

$$V_{out} = \frac{R1 + R2}{R1} \times V_{in}$$

Note At first glance it seems that we did not need the first Golden Rule. However, we used it implicitly to find V_- . If the input current were not zero then we could not use the voltage divider equation but would have to take into account the effect of the input current.

One further thing to note about the non-inverting amplifier is its input impedance. Back in Chapter 4 we learned that when we connect two circuits together it is the input resistance of the second circuit that determines how easy it is to force a signal into the circuit. The higher the resistance the easier it is to drive. Since the input of this circuit is connected directly to the op-amp input, which draws no current, the non-inverting amplifier has a very high input impedance. In theory, Golden Rule 1 says that it should be infinite. In practice it is merely very high. A typical cheap op-amp will have an input impedance $\sim 1 \text{ G}\Omega$ and more expensive low current amplifiers can reach input impedances as high as $10,000 \text{ G}\Omega$.

20.5.2 The inverting amplifier

One constraint of the non-inverting amplifier is that the gain cannot be less than one. Even if we make R_1 infinite, the gain only reaches one. A much more flexible circuit is the inverting amplifier shown in Figure 20-4. Here we apply both the input and the feedback signals to the inverting input of the op-amp.

We can analyze this circuit with our Golden Rules. The first rule says that no current flows into the inputs. That means that all of the current I_1 flowing in resistor R_1 goes on to resistor R_2 , none enters the input. Therefore, we have

$$I_1 = I_2$$

The second rule tells us that, because the non-inverting input is connected to ground, the inverting input will also be at zero volts. The op-amp will adjust its output as needed to keep the non-inverting input at zero volts.

Now we can use Ohm's law to relate the currents and voltages.

$$I_1 = \frac{V_{in} - 0}{R_1} = \frac{V_{in}}{R_1}$$

$$I_2 = \frac{0 - V_{out}}{R_2} = \frac{-V_{out}}{R_2}$$

If we substitute these into the current equation, we relate V_{out} to V_{in} .

$$\frac{V_{in}}{R_1} = \frac{-V_{out}}{R_2}$$

So that our final gain equation is

$$V_{out} = -\frac{R_2}{R_1} \times V_{in}$$

Once again, the gain of the circuit is determined purely by two resistors. As in the inverting amplifier, we have the constant gain and high linearity of an amplifier whose performance depends only on the properties of resistors rather than on the messy details of the transistors that are doing the real work.

This circuit differs in three important respects from the non-inverting amplifier.

1. Its output is inverted, as shown by the minus sign in the gain. Thus the output is 180° out of phase with the input, a positive input giving a negative output and vice versa.
2. The gain can take any real value; this amplifier can reduce as easily as it can increase.
3. The input impedance is much lower.

We can calculate the input impedance by applying a voltage to the circuit and measuring the input current. In fact, we already found

$$I_{in} = I_1 = \frac{V_{in}}{R_1}$$

so the input impedance R_{in} is given by

$$R_{in} = \frac{V_{in}}{I_{in}} = R_1$$

So, while the non-inverting amplifier had a nearly infinite input impedance, the inverting amplifier has an impedance of only R_1 .

20.6 Virtual Ground

As we just saw, the negative feedback allows the output of an op-amp to hold its inverting input at 0V, so long as the amplifier is not saturated. That means that the inverting input pin has some of the properties of ground. However, unlike ground, no current flows into the input pin; it all flows out through the feedback resistor. This behavior makes this point rather special and we call it a **virtual ground**. A normal ground point simply absorbs any amount of current and loses all information about it. A virtual ground point absorbs the current just as well, but it passes on the information about the current to the output of the circuit. There are a lot of circuits that are based on this idea of the virtual ground. The simplest is the **current-to-voltage converter** or **trans-resistance amplifier**.

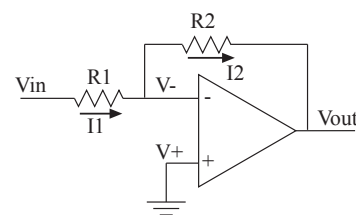


Figure 20-4 Inverting Amplifier

Note I must emphasize again that the second rule works because the op-amp adjusts its own output to make it work. So long as the output voltage can find a value that makes the inputs have the same value, the second rule holds. When the output can't control the input the second rule disappears.

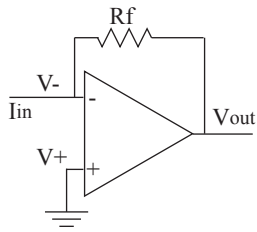


Figure 20-5 Trans-Resistance Amplifier

Note This is of course not true. The input resistance is only zero so long as the output does not saturate. If large currents must flow into the input then the feedback resistor must be chosen small enough that the output will not saturate.

20.6.1 Current-to-Voltage Converter

As Figure 20-5 shows, this is really just the core of an inverting amplifier used on its own. All of the input current flows in the feedback resistor, R_f , so we have

$$I_{in} = \frac{0 - V_{out}}{R_f}$$

or

$$V_{out} = -R_f \times I_{in}$$

So a current input has been converted to a voltage output, hence the common name of the circuit. It is also called a trans-resistance amplifier because the equivalent of the gain is R_f , a resistance.

This circuit has the very curious property that its input resistance is 0. It will absorb any amount of current but the input voltage is always 0.

20.6.2 The Summing Amplifier

The virtual ground point at the input to a current-to-voltage converter allows us to build a circuit that can add together several different voltages. This circuit is called a summing amplifier. In its simplest form it adds just two signals (Figure 20-6).

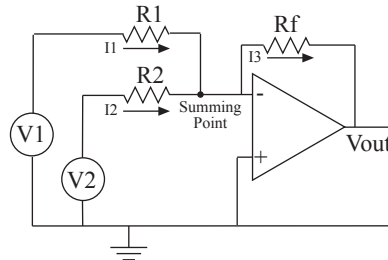


Figure 20-6 Summing Amplifier or Mixer

We can analyze this circuit using the Golden Rules. According to the first Golden Rule and Kirchhoff's current law

$$I_1 + I_2 = I_3$$

Because the summing point is a virtual ground, the voltage there is 0V and we have

$$I_1 = \frac{V_1}{R_1}, I_2 = \frac{V_2}{R_2}, \text{ and } I_3 = -\frac{V_{out}}{R_f}$$

Substituting these into the current equation and solving for V_{out} we find

$$V_{out} = -\left[\frac{R_f}{R_1} \times V_1 + \frac{R_f}{R_2} \times V_2 \right]$$

Thus the output is the scaled sum of the two input voltages.

We can extend this principle to add many voltages. Each input gets its own resistor and adds another term to the equation. Each term is completely independent of the other terms. The only limit to this is the limitation on the output voltage. As we add more and more terms the output voltage will, in general, grow. The whole circuit will cease to function if the output voltage gets close to the power supply voltages, causing the op-amp to saturate.

20.6.3 The Mixer

One of the most important applications of this circuit lies at the heart of the modern music industry. A modern recording is made not with one microphone but with dozens. When recording popular music, each instrument usually has its own input. All of those different signals are then combined by the recording engineer to make the final signal that is what you buy. The signals are combined using a variant of the summing amplifier in which the individual input resistors are made variable. This is called a Mixer.

The whole of modern music recording is completely dependent on the availability of mixers that can combine many signals without one signal affecting the other. When the engineer

turns a knob or moves a slider to increase the volume of one of the channels he does not want the sound from the other channels to change as would be the case with a mixer made only from resistors.

Example

We can see the problem if we consider trying to add two sinewave signals with resistors alone. Figure 20-7 shows the sort of circuit we might try. Note that we make R1 and R2 variable so that the recording engineer can adjust how much of each signal appears in the output. For this to work well, each resistor must affect the amount of only one signal in the output.

We can easily compute the output voltage, though it is a little messy. Assuming that no current is drawn from the output we have

$$I1 + I2 = I3 \quad I1 = \frac{V1 - V_{out}}{R1} \quad I2 = \frac{V2 - V_{out}}{R2} \quad I3 = \frac{V_{out}}{R3}$$

If we substitute the individual currents into the current equation and collect terms we find

$$V_{out} \times \left[\frac{1}{R1} + \frac{1}{R2} + \frac{1}{R3} \right] = \frac{V1}{R1} + \frac{V2}{R2}$$

Now we introduce Rp, where

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

and we have

$$V_{out} = \frac{Rp}{R1} \times V1 + \frac{Rp}{R2} \times V2$$

That looks quite good. Increase R1 and you will decrease the amount of V1 in the output and vice versa. The problem is that Rp also depends on R1. So if you increase R1 then you increase Rp and affect not only the amount of V1 in the final signal but also the amount of V2!

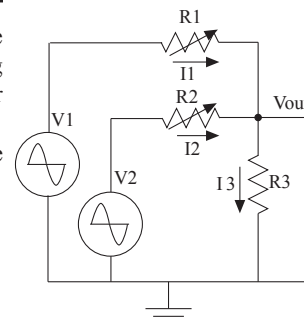


Figure 20-7 Resistive Mixer (BAD)

Real mixers use the summing amplifier circuit of Summing Amplifier or Mixer to make the gains of the channels independent. The feedback resistor, Rf, is fixed and potentiometers are used for the channel resistors R1 and R2. This way the output can be set to any combination of the two input signals. This principle can be extended to as many signals as you like. It is not unusual to see mixers in recording studios that can combine 48 or more separate signals into 1. These are very complex looking because of the number of knobs that they have.



Figure 20-8 16-channel Studio Mixer

Figure 20-8 shows a 16-channel Mackie mixer. The row of sliders at the bottom perform the main mixing function. The sets of knobs above them are used as mini-mixers. Each of the 16 channels can take input from a microphone and several auxiliary sources. The knobs mix those sources together to make a channel and then the sliders mix the channels.

20.7 A few good circuits

The rest of this chapter is taken up with a small gallery of useful op-amp circuits. This is only a small sample of the range of uses for op-amps. There are many, many more circuits available. Op-amp manufacturers include lots of them in their data books. It is very instructive to look through these and figure out how each circuit works. There are some very clever circuits among them!

20.7.1 Unity Gain Buffer

The unity-gain buffer is a special case of the non-inverting amplifier in which R1 is made infinite leaving the circuit of Figure 20-9. This has the extremely simple transfer function

$$V_{out} = V_{in}$$

It is so simple that it seems as though the circuit is useless, since a piece of wire has the same transfer function. The big difference is that the buffer has an infinite input impedance. Thus the circuit draws no current from the source of Vin but can deliver current to Vout. It possesses current gain and power gain even though it has no voltage gain.

The unity gain buffer is used in all kinds of situations where you want to connect two circuits together without the second circuit affecting the first. It is used to isolate one circuit from another (see Chapter 24).

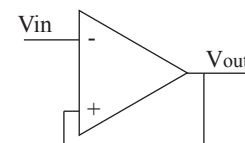


Figure 20-9 Unity Gain Buffer

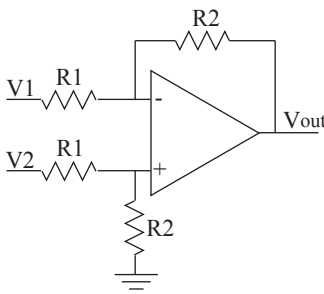


Figure 20-10 Difference Amplifier

Note Although multiplication or division of a signal by a constant is easy, multiplication or division of two signals is quite difficult. It requires specialized computation circuits that are beyond the scope of this book.

20.7.2 Arithmetic operators

We have already met two arithmetic operations. A simple inverting or non-inverting amplifier multiplies a signal voltage by a constant and the mixer adds two signals together. The next circuit (Figure 20-10) subtracts one signal from another. It is sometimes called a subtractor and sometimes a difference or differential amplifier, since it amplifies the difference between two signals.

We can analyze this circuit using the Golden Rules in much the same way that we analyzed the summing amplifier. I have left the actual calculation as an exercise. In exercise 9 you are asked to show that

$$V_{out} = \frac{R2}{R1} \times (V2 - V1)$$

Beware, this circuit depends on the cancellation of two terms, one coming from the upper pair of resistors and one from the lower pair. That means that this circuit depends for its accurate operation on the exact matching of the resistors. Both resistors R1 must be accurately equal as must both resistors R2. This is so critical that normal practice is to use 1% or better resistors and to make one of the resistors, usually the lower R2, adjustable. The circuit is tuned by applying the same sinewave signal to both inputs and then adjusting the variable resistor to minimize the amplitude of the output.

20.7.3 Integration

Surprisingly, integration and differentiation are very easy. Figure 20-11 is the circuit for an op-amp integrator. Historically, this circuit was used in analog computers to perform the mathematical operation of integration. Its transfer function is

$$V_{out} = \frac{-1}{R \times C} \times \int_0^t V_{in} dt$$

Today, this circuit is chiefly used in ramp generators, circuits to generate a linearly rising voltage. This relies on the fact that

$$\int C dt = Ct$$

where C is the constant input voltage. If we switch C between positive and negative values then the output from the circuit will be a triangle wave as shown in Figure 20-12.

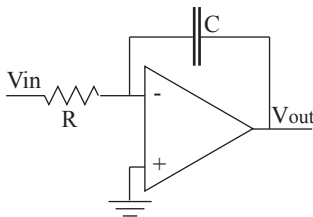


Figure 20-11 Integrator

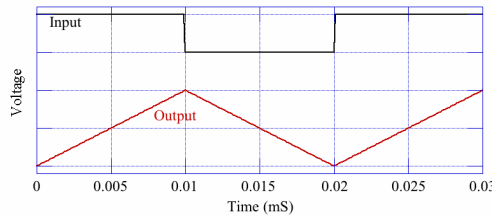


Figure 20-12 Integrating a Square Wave

Info The analysis of this one requires calculus. First we remember that the charge on the capacitor is related to the voltage across it by

$$Q = C \times V$$

and that the charge is the integral of the current that has flowed onto the plates.

$$Q = \int_0^t i(t) dt$$

Now the current flowing into the capacitor is equal to the current flowing in the input resistor, R.

$$i(t) = \frac{V_{in}}{R}$$

and the voltage across the capacitor is just $-V_{out}$ because one end is at virtual ground. That means we have

$$-C \times V_{out} = \int_0^t i(t) dt = \int_0^t \frac{V_{in}}{R} dt$$

or

$$V_{out} = \frac{-1}{R \times C} \times \int_0^t V_{in} dt$$

So the output voltage is proportional to the integral of the input voltage and the circuit functions as intended.

Beware, this circuit has a nasty tendency to get out of control. It is very important that the input signal have an average level of exactly zero volts. If the average is not zero, then the non-zero constant will be integrated and the integral of a constant is a steadily rising voltage. Thus the output voltage will rise and rise until it comes up against one of the power supply limits and the circuit stops working.

20.7.4 The differentiator

If we interchange the capacitor and the resistor in the circuit of Figure 20-11, then we get the circuit of Figure 20-13. This circuit differentiates the input voltage according to the equation

$$V_{out} = -R \times C \times \frac{dV_{in}}{dt}$$

It is much safer than the integrator; the derivative of a constant is zero! This circuit does place great demands on the op-amp however. Accurate differentiation requires very good high frequency response so this circuit works best with a very fast amplifier.

Like the integrator, this circuit was designed for use in analog computers. Its principle use today is as a signal modifier. For example it can take a triangle wave and convert it into a square wave.

Info The analysis is rather similar to that of the integrator except that this time we need to differentiate the capacitor equation.

$$Q = C \times V \text{ therefore } I = C \times \frac{dV}{dt}$$

Now the current from the capacitor forms the input for a trans-resistance amplifier so we have

$$V_{out} = -R \times C \times \frac{dV_{in}}{dt}$$

20.7.5 Photodiode amplifier

Probably the best example of a trans-resistance amplifier is seen in the photodiode amplifier. A photodiode is a current source whose strength depends on the intensity of the light falling on the diode. The relationship between incident light and photo-current is most linear if there is no voltage across the diode. This makes the trans-resistance amplifier an ideal match for a photodiode (Figure 20-14).

Obviously, $V_{out} = -R \times I_D$. You can make the output voltage either positive or negative depending on which way round you put the diode. This is the most linear and the fastest kind of photodiode circuit. Because there is no voltage across the diode, its capacitance does not slow the response. Of course, the speed of the overall circuit depends on the speed of the op-amp so if you want speed you have to use a fast op-amp.

Summary

An operational amplifier is a high quality differential amplifier with the transfer function

$$V_{out} = G \times (V_{in+} - V_{in-})$$

We call the input V_{in-} the **inverting** input and the V_{in+} the **non-inverting** input.

It is usually operated with negative feedback, which trades a loss of gain for a set of amplifier properties that depend only on passive components such as resistors and capacitors. When operating with negative feedback the amplifier obeys the **Golden Rules**

1. The inputs draw no current.
2. The output adjusts itself to keep the voltage at the inverting input the same as the voltage at the non-inverting input.

The two most common op-amp circuits are the non-inverting amplifier and the inverting amplifier.

The **non-inverting amplifier** has the transfer function

$$V_{out} = \frac{R1 + R2}{R1} \times V_{in}$$

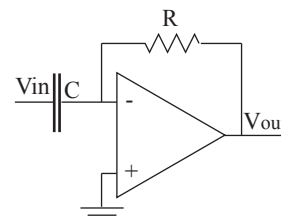


Figure 20-13 Differentiator

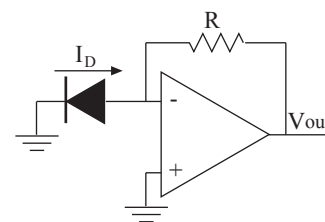


Figure 20-14 Photodiode Amplifier

and a nearly infinite input resistance.

The **inverting amplifier** has the transfer function

$$V_{out} = -\frac{R_2}{R_1} \times V_{in}$$

and the input resistance R1.

The **transresistance amplifier** or **current-voltage converter** has the transfer function

$$V_{out} = -R_f \times I_{in}$$

It takes an input current and outputs a voltage. The input point is always at 0V, the same potential as ground. Unlike ground, however, the current flowing into the input is not lost back to the power supply. We call this point **Virtual Ground** or **Virtual Earth**.

A n extension of the transresistance amplifier is the **summing amplifier** or **mixer**. It has the transfer function

$$V_{out} = -\left[\frac{R_f}{R_1} \times V_1 + \frac{R_f}{R_2} \times V_2 \right]$$

You can extend this to as many inputs as you like. It has the property that it adds the inputs together but does not allow one input to affect any other.

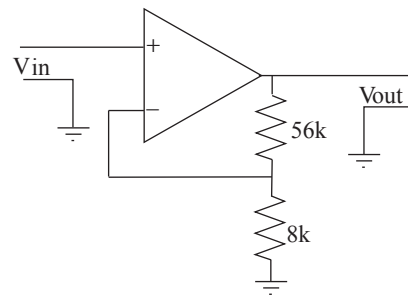
The opposite of a summing amplifier is a **difference amplifier** or **differential amplifier**. It has the transfer function

$$V_{out} = \frac{R_2}{R_1} \times (V_2 - V_1)$$

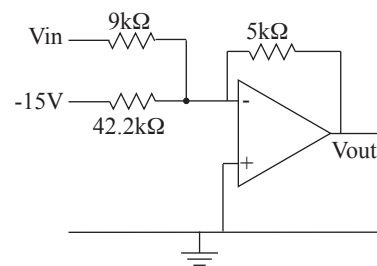
For this to work properly the resistors with the same name must be well matched, that is of very nearly identical value. This circuit is usually built with 1% or better precision resistors.

Exercises

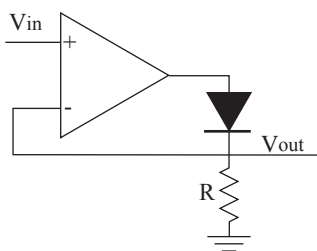
1. The input to the following circuit is a 0.2V P-P sine wave at 2.4kHz. Plot the input and output on a single graph



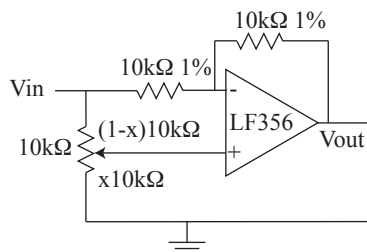
2. Use the golden rules to find the equation that gives the output voltage from the circuit on the right as a function of its input voltage. This is actually quite a useful circuit as it can convert temperatures from Fahrenheit to Celsius. In order to keep the values within the ±15V range of the power supplies we have to use a representation where the voltage = 0.1×temperature. There is also the slight oddity the output is inverted, ie. negative.



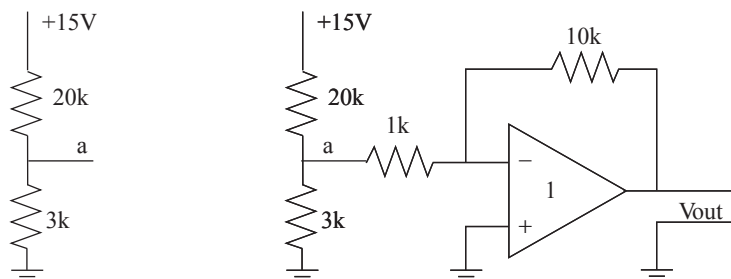
3. The circuit on the right is called a precision rectifier. By considering separately the behavior of the circuit for positive input voltages and for negative input voltages, deduce the output of the circuit when the input is a 0.2 V amplitude sinewave at 100Hz. How does the behavior of this circuit differ from that of a simple diode rectifier?



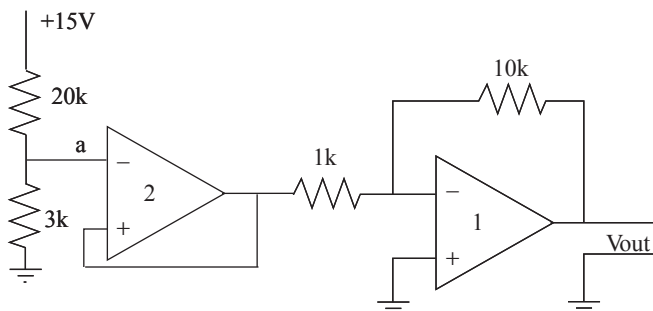
4. The circuit on the right is supposed to produce any gain between -1.0 and +1.0 as the variable resistor is moved from the bottom to the top. You can treat the variable resistor as a pair of series resistors with values $x \times 10k$ and $(1-x) \times 10k$. Then moving the slider on the resistor corresponds to varying x from 0 to 1. Use the op-amp golden rules to find the gain of the circuit as a function of x



5. a) Calculate the output voltage (V_{out}) from the Voltage divider below left.
 b) The output of the voltage divider is then connected to the input of an op amp giving the circuit below. Calculate the output from the amplifier. Note that the answer is NOT the obvious one.



6. Explain how adding a second op-amp (2 below) makes the output voltage have its obvious value.



7. Use the Golden Rules and Kirchoff's Laws to show that the Difference Amplifier of Figure 20-10 obeys the equation.

$$V_{out} = \frac{R_2}{R_1} \times (V_2 - V_1)$$

Chapter 21:Real Op-Amps

21.1 Introduction

In the last chapter, we designed circuits around idealized operational amplifiers. Their inputs drew no current, their outputs were limited only by the power supply voltages, and their bandwidth limited only by the fall of the open loop gain. Real op-amps are not quite so perfect, although you can buy near perfection in any one area for a relatively small price. In this chapter we shall look at the limitations of real op-amps, see how they limit the circuits in which they are used, and look at a few of the hundreds of op-amps available.

There are three main areas in which real op-amps differ from the theoretical ideal.

- 1) Real op-amps operate satisfactorily only up to some (quite modest) frequency. After that, they run out of gain and cease to behave like op-amps. The limits are often different for small signals and for large signals.
- 2) The inputs of real op-amps do draw some current and they have other problems.
- 3) The output of a real op-amp can handle only a limited amount of voltage and current.

We shall deal with these areas one by one in the following sections and will then look at some real op-amps. Most of the discussion of limitations will use the example of the LF356, a cheap general purpose op-amp that is suitable for most non-specialized applications.

21.2 Frequency Response

No amplifier can operate equally well for all signals at all frequencies. Operational amplifiers are mostly used in circuits operating at fairly low frequencies, up to about 1MHz. Above that, more specialized devices and very different kinds of circuits have traditionally dominated. However, op-amps are increasingly found in faster circuits as higher speed devices become more and more economical. Specialized op-amps are now found in some circuits operating at tens of MHz as video technology moves into the digital era. This section deals with the factors that limit the frequency response of op-amp circuits.

21.2.1 Gain-bandwidth Limit

As was already mentioned, the open loop gain of an op-amp is extremely high at low frequencies but falls rapidly. Figure 21-1 is a graph of the open loop gain of an LF356 taken from the manufacturer's data sheet

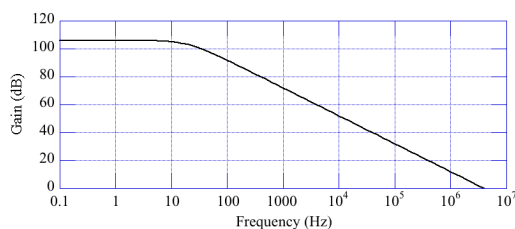


Figure 21-1 Open Loop Frequency Response of the LF356

The gain starts to fall from its maximum of 200,000 at about 10Hz. It soon falls at a steady rate, dropping by a factor of 10 every time the frequency increases tenfold. We can describe the gain mathematically with the equation

$$G(f) = \frac{G_0}{\sqrt{1 + \left[\frac{f}{f_0}\right]^2}}$$

Note You should recognize this gain curve. It is exactly that of a low-pass filter as described in Chapter 8.

where G_0 is the gain at very low frequency, and f_0 is a characteristic frequency at which the gain starts to fall (strictly, it is the frequency at which the gain has fallen by 30%). For the LF356, $G_0 = 250,000$ and $f_0 = 20\text{Hz}$.

***Why does the gain fall like this?**

One of the chief design goals for an operational amplifier is **stability**. We say that an amplifier is stable if it is not prone to oscillate—to generate a signal on its own. It turns out that uncontrolled high-frequency gain is a very good source of oscillations. Op-amp designers respond by putting a capacitor into the circuit, whose charging and discharging slow the circuit and reduce the high frequency gain. This capacitor forms a low-pass filter with various internal resistances. It is the effect of this low pass filter that gives us the fall in open-loop gain. If the capacitor were omitted then the amplifier gain would still tend to decrease at high frequencies but the details would be determined by the poorly controlled high-frequency behavior of the transistors in the amplifier. It would be much harder to make the amplifier stable. Adding the internal capacitor overrides the effects of the transistors and results in a controlled fall of gain with frequency.

Closed-Loop Bandwidth

How does the fall in the open-loop gain affect the closed-loop gain of a real amplifier? Well, the condition for a closed loop amplifier to work well was that the open-loop gain \gg closed-loop gain. When that relation fails the closed loop gain will fall below its ideal value. We will have reached the upper limit of the closed-loop bandwidth. Because we have a detailed model of the behavior of an op-amp based non-inverting amplifier (see section 20.4), we can compute the exact gain of a closed loop amplifier at any frequency (Figure 21-2).

Info *Computing the Closed-Loop Frequency Response

In Chapter 20 we found that the gain of a non-inverting amplifier was given by

$$V_{out} = \frac{R1}{R1-R2} \times \frac{V_{in}}{\left[1 + \frac{R1-R2}{G \times R1}\right]}$$

If we substitute our expression for $G(\omega)$ into the closed-loop gain expression and simplify it, writing

$$A = \frac{R1+R2}{R1}$$

we find

$$\frac{V_{out}}{V_{in}} = \frac{1 + \frac{A}{G_0} \times \sqrt{1 + \left[\frac{f}{f_0}\right]^2}}$$

This is the formula that is plotted in Figure 21-2

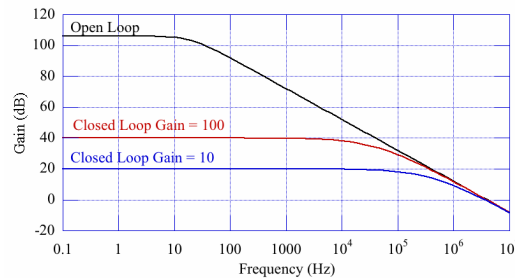


Figure 21-2 Closed-Loop Frequency Response of LF356

As we see, the larger the closed-loop gain, the smaller the bandwidth of the amplifier. In fact, because the open loop gain falls at exactly the same rate as the frequency rises, the gain-bandwidth product for the closed-loop amplifier is constant. For the LF356 the gain-bandwidth is given on the data sheet as 5Mhz and the figure appears to agree quite well with this.

Once we know the gain-bandwidth, we can compute the bandwidth of any closed loop amplifier

$$\text{Closed-Loop Bandwidth} = \frac{\text{Gain-Bandwidth}}{\text{Closed-Loop Gain}}$$

What happens to the signal as the frequency approaches the bandwidth limit?

The answer depends to some extent on the amplitude of the signal. So long as the signal is small, the gain simply falls fairly smoothly as the frequency increases. The output amplitude is lower than it should be but the signal shape is correct. However if the amplitude is large there is a second limitation of real amplifiers that distorts the signal.

21.2.2 The Slew-rate Limit

In addition to the basic frequency limitation of the gain-bandwidth limit, there is also a limit to how fast the output voltage of the op-amp can change. This limit is called the **slew-rate** limit and is specified in volts/ μSec . It ranges from $<1\text{V}/\mu\text{S}$ for older, slower chips such as the 741 and some single-supply op-amps, through $12\text{V}/\mu\text{S}$ for the LF356 to $>1000\text{V}/\mu\text{S}$ for some ultra-fast chips.

Info *Origin of the Slew-Rate Limit

The slew-rate limit arises from the same source as the bandwidth limit. At the heart of all integrated op-amps there is a capacitor whose job is to swamp the effects of all the stray capacitances inside the transistors of the op-amp. That capacitor is there to force the gain of the op-amp to roll off as we have already seen but it has the secondary effect of limiting the rate of change of the output voltage. There is only a certain amount of current available to charge or discharge this capacitor and that sets an upper limit to the rate at which the voltage across the capacitor can change because

$$I = C \times \frac{dV}{dt}$$

This limit on the rate of change of a key internal voltage affects the rest of the op-amp and thus limits the rate of change of the output voltage.

Consider a gain of 10 non-inverting amplifier made with an LM741 op-amp. This older amplifier has nearly the same gain bandwidth as an LF356, 1.5Mhz, but a slew-rate limit of only $0.5\text{V}/\mu\text{S}$. The closed-loop bandwidth of the circuit at a gain of 10 is 150kHz (using the formula above). Let us look at the output that we would expect if we drove the amplifier with various sinewaves at 100kHz. Figure 21-3 shows what we would expect to see at the output for input amplitudes of 0.01V, 0.03V, and 0.1V.

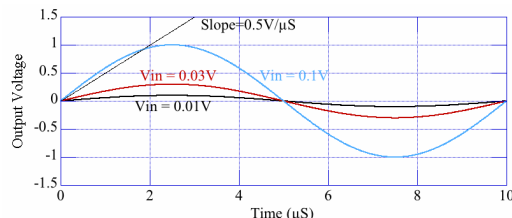


Figure 21-3 Ideal Outputs of Gain-of-10 Amplifier

The fastest rate of change, the steepest slope, occurs where the voltage crosses zero. If we look at those regions, we can see that the slope increases as the amplitude increases. Now, the output of a 741 cannot change faster than the slew-rate limit of $0.5\text{V}/\mu\text{S}$. I have marked this slope in Figure 21-3 with a thin black line. The two lower curves lie well below that line; their steepest slopes are less than $0.5\text{V}/\mu\text{S}$ and they can be produced without distortion. The top-most curve, however, lies above that line for some considerable distance. Its maximum slope is $0.63\text{V}/\mu\text{S}$, which is too fast for the op-amp to produce. That means that the op-amp does its best and produces an output that follows the black line until the slope falls low enough for the op-amp to catch up (where the black line crosses the dark gray curve). After that, the output is correct. The same thing happens on the downward slope so the real output of this circuit looks like Figure 21-4.

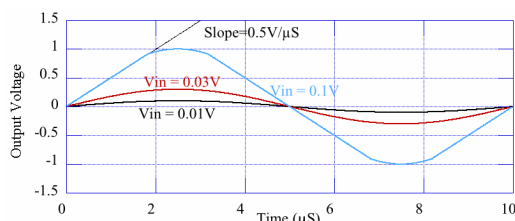


Figure 21-4 Actual Outputs of Gain-of-10 Amplifier

The slew-rate limiting of the op-amp **distorts** signals that try to change faster than $0.5\text{V}/\mu\text{S}$ and so we get an amplifier that behaves sensibly for small amplitude signals but that distorts larger signals badly. This means that the bandwidth calculated from the gain-bandwidth formula is only valid for small amplitude signals. The large-signal, high-frequency behavior of the amplifier is dominated by the slew-rate limit. The only cure for this problem is to use a faster op-amp. If we replaced the 741 in this example with an LF356 then we would not see any distortion up to the full output of the op-amp. Even a 13V, 100kHz sinewave has a maximum slope of only $8.2\text{V}/\mu\text{S}$, well below the $12\text{V}/\mu\text{S}$ slew-rate limit of the 356.

Remember A signal is said to be **distorted** when its shape is altered. Two signals have the same shape if one can be laid exactly on-top-of the other by altering only the scales and offsets between the two signals. If the signal is an audio signal then distortion affects the sound that the signal makes. Two signals with the same shape will sound alike but a distorted shape will sound different.

21.2.3 Transient response

The most challenging signals for any amplifier are those with very fast, discrete, changes in voltage. Such a rapid change is called a **transient** because it does not last very long. The transient response of an amplifier is affected by both its bandwidth and its slew-rate limit. As usual, the finite bandwidth is most important at small amplitudes and the slew-rate most important at large ones. The usual way to measure the transient response is to use a very sharp-edged square-wave signal and to show the output for both small and large signals. Figure 21-5 and Figure 21-6 show the transient response of the LF356. In each case, the input signal was a square wave with a rise and fall time of a few nanoseconds, instantaneous on the scale of these graphs. These curves were obtained with the op-amp connected as a unity gain, non-inverting amplifier.

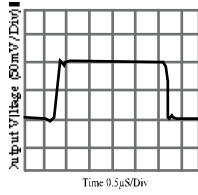


Figure 21-5 Small-Signal Transient Response

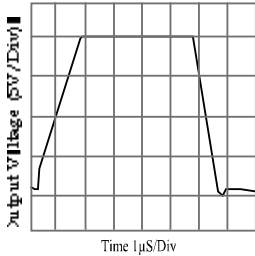


Figure 21-6 Large-Signal Transient Response

Note We have seen exactly the same sort of asymmetry in rise and fall times before. The NMOS and PMOS inverters suffered from the same problem. Rise and fall times were very different because the charging and discharging currents for a capacitor flowed through two different paths.

As Figure 21-5 shows, the small-amplitude transient response is very good. The output is slightly rounded at the end of the transition and there is a small wiggle that lasts about 100nS before the level settles down. It does the same thing for a rising and for a falling signal. At unity gain the LF356 has a bandwidth of 5Mhz so that the signal rise can contain components this fast without loss of amplitude. In fact the rise and wiggle only last about 100nS so that the amplifier is still giving us some gain up to about $1/100\text{nS} = 10\text{Mhz}$. The small signal response is completely consistent with the gain-bandwidth limit.

By contrast, Figure 21-6 shows the large signal response. The large amplitude response is much poorer; the signal takes nearly $2\mu\text{s}$ to rise to its full voltage. Since this corresponds to a frequency of only 0.5Mhz the bandwidth is obviously not the limiting factor. The rising edge is a nearly straight line, rising at the slew-rate limit, which totally dominates this behavior. Interestingly, the falling edge is quite a lot faster. This asymmetry happens because the capacitor that limits the slew rate has a different path for its charging and discharging currents so the time constants for the charge and discharge are different.

21.2.4 How to get to high frequencies

There is really only one way to get round the frequency limitations of op-amps; spend more money. Faster op-amps than the LF356 are available but often, the more speed you want, the more money you have to spend. Table 21-1 shows a few of the higher speed op-amps available today with their gain-bandwidths, slew rates, and approximate single-unit prices in 2000.

Table 21-1: 1 Some High Speed Op-Amps

Part No.	GBW(MHz)	Slew Rate (V/µS)	Supply Voltage	Price (2000)	Comment
LM6132	7	12	0,5V	\$5	Low power, single supply 2 to a package
LM6152	45	30	0,5V	\$10	Faster version of the 6132
LM6361	50	300	±15V	\$5	Fast, standard power, min gain +2 or -1
LM6364	175	300	±15V	\$5	Faster, min. gain +5
LM6365	725	300	±15V	\$5	Fastest, min. gain +25
LM7121	200	1000	±5V	\$3	High slew-rate, tiny package.
LM7171	200	4100	±5V	\$3	Even higher slew rate, tiny.

21.3 Output Current Limit

Op-amps are basically low power devices; their output stages only deliver a small current to the load. For example, our LF356 is rated by the manufacturer to source or sink only 25mA of current. If you try to exceed that current then the output voltage will fall and the op-amp will cease to obey the golden rules.

The low current limit of ordinary op-amps means that you cannot use them in a whole variety of useful tasks for which they are otherwise ideal. Examples of such tasks include driving loudspeakers, controlling DC motor speeds, driving signals on long cables, and building very stable power supplies for all sorts of equipment.

Example

A typical loudspeaker in a small device such as a transistor radio has a resistance of 8Ω and requires 1W of power to produce a useable sound. In order to dissipate 1W in an 8Ω resistor we need a current of

$$I = \sqrt{\frac{P}{R}} = \sqrt{\frac{1}{8}} = 0.35A$$

350mA is 14 times more current than we can get from an LF356. We cannot drive a loudspeaker with an ordinary op-amp.

There are several reasons why op-amps usually have such small output current limits.

- High-current transistors must be physically larger than low current ones. That means that high-current op-amps take up more silicon than their lower power brethren and that makes them cost more.
- The large size of high-current transistors means that they have larger stray capacitances and are thus slower than lower current devices.
- High-power devices dissipate a lot of heat.

The high heat dissipation causes two problems.

- It means that you have to use a larger, more expensive package in order to get rid of the heat so it doesn't destroy the op-amp.
- It means that the silicon of a high-current op-amp runs hotter than a corresponding low-current device. The higher temperature degrades most of the other properties of the device. In particular the input current and voltage errors are much worse at high temperatures and so the small signal behavior of the devices is not very good.

21.3.1 High current solutions

There are two ways to increase the current drive capacity of op-amps. The simple one is to buy a special high-current device and the fancy one is to build a power stage out of discrete transistors driven by a low-current device. As usual, there is a trade-off between cost and convenience. The high power op-amps are very convenient but can be more costly than adding a discrete power stage to a low-power op-amp.

Table 21-2: 2 Some Power Op-Amps

Part No.	V_{OUT}	I_{OUT}	Power Bandwidth	I_{BIAS}	Comments
LH0041C	$\pm 15V$	0.2	20k	1μ	Getting a bit elderly, still useful
LH0021C	$\pm 12V$	1	15	1μ	
LH0101C	± 12	5	300k	1μ	A hefty device and quite fast
ULN3751Z	± 5	3.5	20k	1μ	Comes in a TO-220 package
L272	± 12	1A	10k	2.5μ	Comes in a Dip package
L165	± 12	3A	100k	1μ	Quite fast for its current, TO-220 package

Power op-amps are available with output currents up to a few amps. They are typically not very fast and have rather poor input characteristics (see below). They are often best used as unity-gain buffers in partnership with an ordinary low-current op-amp that provides gain and gives the overall circuit its nice input characteristics. Table 21-2 shows some of the devices available in 2000.

The other way to get more current is to add one or more external discrete devices to a low-current op-amp. This has several advantages.

- It keeps the heat of the power stage well away from the temperature sensitive input circuitry of the op-amp.
- It is cheaper than using both a low-power and a high-power op-amp.
- It allows you to handle very large currents because power transistors are available that can handle currents of tens of amps, higher than even the best power op-amps.

Figure 21-7 shows the simplest kind of output buffer. Each FET is connected as a source follower. It draws only enough current from the op-amp to charge its input capacitance, typically a few μA , but can deliver the full source-drain current of the transistor to the load, 200mA in this case. By itself this is a terrible amplifier. The upper FET has a turn-on voltage of about +2.5V and the lower FET a turn on voltage of -2.5V so that there is a 5V range of input where neither transistor is turned on. This huge gap would make the circuit useless if were not for the magic of feedback. If we put the circuit inside the feedback loop of an op-amp (Figure 21-8) then all is well.

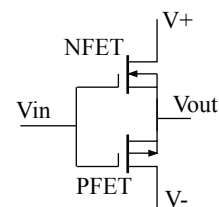


Figure 21-7 Simple Power Buffer

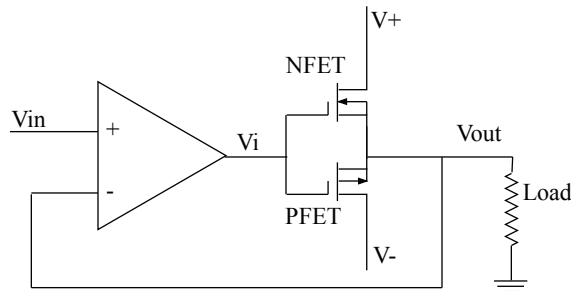


Figure 21-8 Amplifier with Simple Power Buffer

The op-amp has to try to obey its golden rule and keep the output voltage, V_{out} , equal to the input voltage. For example, if the input voltage is 0.3V then the op-amp will try to make $V_{out} = 0.3V$. It can do that by making V_i some voltage higher than 3V. Then current will flow in the NFET and the output voltage, V_{out} , will rise. The op-amp will settle on producing a V_i that is just sufficient to make $V_{out} = 0.3V$. Now, if we put in a small amplitude sinewave into V_{in} , V_{out} will try to follow V_{in} . That means that V_i will have to do something quite dramatic as the input voltage goes from just less than 0V to just above 0V. V_i will have to swing from just below -3V to just above +3V as fast as possible. The signals will look something like Figure 21-9.

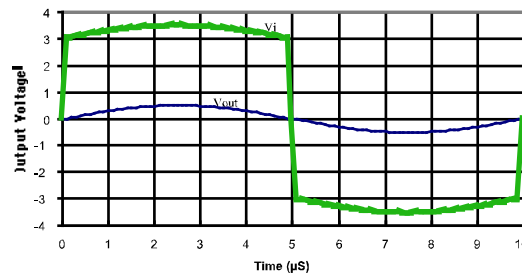


Figure 21-9 FET Drive Voltage

The intermediate voltage, V_i , jumps as fast as it can from +3V to -3V as the output crosses zero. At low frequencies, the 0.4 μ s that an LF356 takes to skip across 6V is so short that the output shows no sign of it. At high frequencies, like this, the time taken to make the jump becomes significant and the output wave is slightly distorted for about 1 μ s. This distortion, occurring where the output signal crosses zero, is called **crossover distortion**. It is the chief problem with this sort of power booster. With a fast op-amp like the LF356, the problem is barely noticeable at audio frequencies and this circuit is the basis for most audio power amplifiers.

Crossover distortion can be reduced by **biasing** the output stage so that there is a small current flowing down the output chain even at 0V output (Figure 21-10). The improvement comes at the expense of increased current drain from the power supplies, and increased heating of the output transistors.

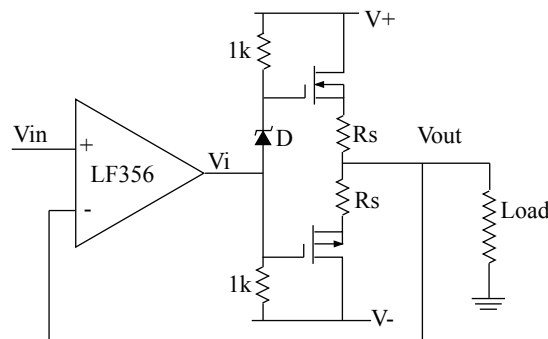


Figure 21-10 Improved High-Current Amplifier

The Zener diode, D , is chosen to hold the gates of the output FETs about 6V apart so that when V_i is 0 both FETs are slightly turned on. The source resistors R_s limit the current flowing in the FET chain to some chosen level. Now, if V_i rises by 0.1V then the output voltage will also rise as the NFET turns on more and the PFET turns off. There is no longer a dead zone and the crossover distortion is greatly reduced.

By adding more components, it is possible to improve the biasing and further reduce the distortion. It is also possible to add short circuit protection and other protections so that the output transistors will not be damaged no matter what someone does to the output. Such precautions are normal in stereo amplifier circuits, which often use output stages like this for the final power amplifier stage. The lower end of the stereo market is dominated by special audio-power ICs that pack all of this up into one package.

21.4 Input Characteristics

All of the limitations that we have met so far affect the output of an op-amp. Now we come to the imperfections in the input circuitry. Remember that the ideal op-amp allows no current to enter or leave its input pins and the output voltage is strictly given by

$$V_{out} = G \cdot (V_{in+} - V_{in-}).$$

Real op-amps fall short of both these ideals.

21.4.1 Common-Mode Rejection

The least serious of these shortfalls is a dependence of the output on the values of the inputs instead of only on their difference. Mathematically, this means that the formula for the output should really be written

$$V_{out} = G_D \cdot (V_{in+} - V_{in-}) + G_C \cdot (V_{in+} + V_{in-})$$

where G_D is called the **differential gain** and G_C is the **common-mode gain**. The effect of this is that the average level of the inputs has an effect on the output. This is not supposed to happen and in real op-amps the effect is very small. For example, the LF356 has $G_D = 250,000$ and $G_C = 2$, so that the average value would have to change by thousands of times the difference signal in order to have any perceptible effect on the output. The common mode gain is usually specified indirectly using the **Common-Mode Rejection Ratio** or **CMRR**. This is just the ratio of the two gains

$$CMRR = \frac{G_D}{G_C}$$

and is usually specified in dB. That is

$$CMRR(dB) = 20 \times \log \frac{G_D}{G_C}$$

Most op-amps offer CMRRs of 80dB or more. That is, the differential gain is at least 10,000 times the common-mode gain.

The only time that we normally have to worry about the CMRR is when we are looking at a very small differential signal, mV or less, riding on top of a lot of common-mode noise. For example, a small signal is carried on very long wires. The wires act as an antenna, picking up electrical noise from the surroundings, especially noise from AC powered equipment such as motors and fluorescent lights. Because both signal wires take the same path, they both pick up the same noise. So the input to the op-amp can consist of a 2mV differential signal hidden in 2-5V of common-mode noise. In circumstances like these, where the CMRR is really important, a special multi-op-amp circuit called an **instrumentation amplifier** is often used (Figure 21-11). This arrangement of three op-amps can easily achieve CMRRs of 100dB and with careful choice of op-amp and careful matching of the resistors can reach 120dB.

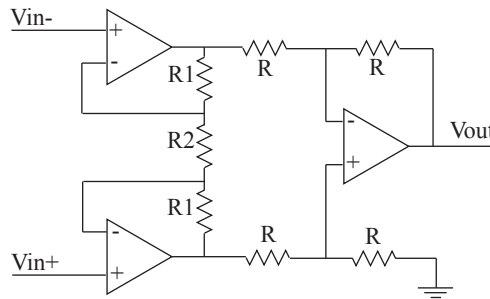


Figure 21-11 Three Op-Amp Instrumentation Amplifier

The first two op-amps form a curious amplifier that has both a differential input and a differential output. Its differential gain is

$$G1 = 1 + 2 \times \frac{R1}{R2}$$

and its common mode gain is 1. The third op-amp is a conventional difference amplifier that has a differential gain of 1 and a common-mode gain that depends on the matching of the resistors. Even with only 1% resistors it is easy to get a CMRR of >100dB with this circuit, so long as the input op-amps have fairly good CMRR. LF356's are quite adequate and OP-27's are superb.

21.4.2 Input Offset Voltage

In theory, if the CMRR is good enough, the transfer function of an op-amp is

$$V_{out} = G \cdot (V_+ - V_-)$$

which implies that the output voltage will be zero when the two inputs are exactly equal. In practice this is never the case. Instead the output is zero when there is some small voltage, called the **input offset voltage**, between the inputs. We can account for this in our formula by introducing V_{OS} thus

t

There is a wide range of variation in input offset voltage among commercial op-amps. Our favorite LF356 has a V_{OS} of 1mV at 25°C and, like most op-amps, it gets worse as the temperature increases. Precision op-amps such as the popular OP-27 have input offset voltages of only 10μV, while a power op-amp such as the 3A LM675 can only manage 10mV.

The input offset voltage is normally only important for high precision DC amplifiers where the relationship between input and output must be accurate to a fraction of one percent. The constant offset is not a problem in AC amplifiers even when precision is required since it does not vary in time. An example of a circuit that requires excellent offset behavior is a photodiode amplifier used in low light conditions. It must collect the tiny current from the photodiode and convert it into a larger voltage that is proportional to the current. When no light falls on the photodiode you want the output voltage to be zero. Any input offset voltage is added directly to the output of the circuit. Since the full-scale output of such a circuit might only be 10mV the 1mV offset voltage of an LF356 would represent a 10% error in the output. This would obviously be a place to use a much higher precision op-amp.

21.4.3 Input Currents

One of our golden rules of op-amp behavior is that the inputs draw no current. This is not true for any op-amp; some small current is needed to make the device work. This small current is called the **input bias current**. Its value ranges from >10μA in some high-speed devices, through 100pA for the LF356, to only 60fA for the best low-current devices. The input current causes problems in two different ways. First, it leads to errors in much the same way as the input offset voltage and second it limits the performance of some low current circuits.

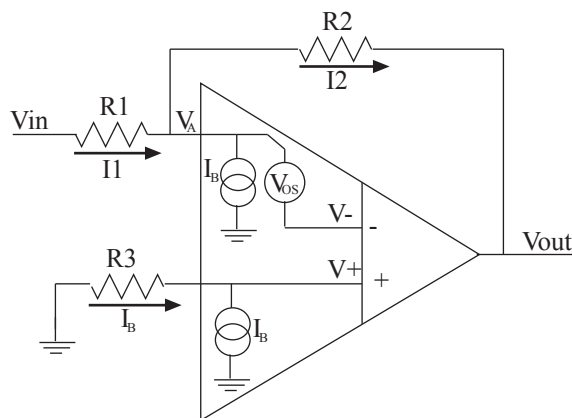


Figure 21-12 Op-Amp Input Error Terms

21.4.4 *Total Input Error

The first kind of error can be seen if we reanalyze the inverting amplifier including the effects of various errors as shown in Figure 21-12. I have shown the input error sources as explicit components within the op-amp, which they certainly are not. The errors are simply the results of the way the real circuitry works but it is very convenient to show them this way for analysis. The little empty triangle at the point is then supposed to behave as an ideal op-amp. I have also added an extra resistor, R3, for reasons that will become clear.

We first calculate V+. It is not 0V anymore because the input bias current flows in it. If the bias current flows into the chip then we have

$$V_+ = -R3 \times I_B$$

Next we calculate VA in terms of Vin and Vout. Looking first at R1 we have

$$V_A = V_{in} - I1 \times R1$$

and then looking at R2

$$V_A = V_{out} + I2 \times R2$$

Because of Kirchhoff's current law, we know that

$$I1 = I2 + I_B$$

and can thus write

$$V_A = V_{in} - I2 \times R1 - I_B \times R1$$

We now have two simultaneous equations for I2 and VA that we can solve by eliminating I2.

From the R2 equation

$$I2 = \frac{V_A - V_{out}}{R2}$$

so that

$$V_A = V_{in} - (V_A - V_{out}) \times \frac{R1}{R2} - I_B \times R1$$

and

$$V_A + \frac{R1}{R2} \times V_A = V_{in} + \frac{R1}{R2} \times V_{out} - I_B \times R1$$

Now, the voltage V- seen by the ideal op-amp is VA - VOS so that the second golden rule tells us that

$$V_- = V_S - V_{OS} = V_+ = -I_B \times R3$$

and so

$$V_A = V_{OS} - I_B \times R3$$

We substitute that back into the expression for VA and find

$$V_{out} = -\left(\frac{R2}{R1} \cdot V_{in}\right) + \left(1 + \frac{R2}{R1}\right) \cdot V_{OS} + \left(R3 - \frac{R2 \cdot R1}{R2 + R1}\right) \cdot I_B$$

So two new error terms are added to our usual formula for an inverting amplifier. There is nothing that can be done about the VOS error except choose a better op-amp. However, the IB

Note Different devices may have their bias currents flowing either into or out of the chip depending on the details of the input circuitry. The bias current always flows in the same direction for both inputs of a single device. Thus if IB+ flows in, then so does IB-.

error has a coefficient that is a mixture of positive and negative terms and so can be made zero by a suitable choice of R3. We can eliminate the bias current error by choosing

$$R3 = \frac{R2 \times R1}{R2 + R1}$$

That is, we set R3 to the parallel combination of R1 and R2.

Note Because the input bias currents of modern FET input op-amps are so small compared to V_{OS} , the error is usually dominated by V_{OS} and we omit R3. However, you will regularly find R3 in precision circuits that use older op-amps with their higher bias currents.

So the input bias current causes an error that is small in most cases and which can be completely removed by adding R3. Actually, things are not quite that good. We have assumed that the input bias current at the negative input is the same as that at the positive input. In real chips this is not the case. We call the difference between the two input currents the **input offset current** and it is typically comparable in magnitude to the input bias current. That makes the compensation resistor, R3, rarely worth using!

21.4.5 Correcting offset errors

We have seen that both the amplifier input voltage offset and the input bias and offset currents lead to offset errors in amplifiers. There are some circumstances where you cannot tolerate such errors. One example is a precision, low-level, DC amplifier used to process the signal from a transducer such as a light sensor or a force sensor. In this case, offset errors in the amplifier produce wrong answers from the device. Another is the operational integrator where any input offset gets integrated and produces a steadily rising drift at the output. In these cases you have to provide a means to cancel out the offset. One way is to use the offset-compensation terminals built into many IC op-amps. The other is to add external circuitry to introduce a deliberate, canceling, offset.

Many IC op-amps have terminals for an offset correction. Exactly how they are used differs somewhat from device to device, but the LF356 is fairly typical. A 25kΩ potentiometer is connected between pins 1 and 5 with the wiper connected to the positive supply as shown in Figure 21-13.

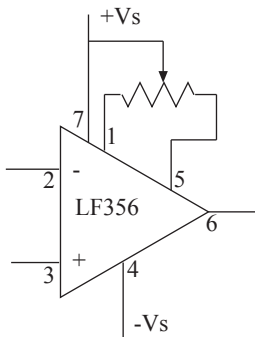


Figure 21-13 Op-Amp Offset Compensation

The range of the built-in offset correction is usually quite small and the unbalance that it introduces into the internal working of the IC tends to degrade the CMRR. In addition, the offset voltage varies somewhat with temperature and so a circuit that is trimmed to zero at one temperature will not be trimmed at another. This temperature drift is actually made worse if you use the built-in correction circuit. The best route to real precision is to choose a special, high-quality op-amp such as the OP-27E, which has a typical V_{OS} of 10μV and drift of 0.2μV/°C.

21.4.6 Input current limitations

There are a number of places where the very existence of the input bias current is the problem rather than the errors that it introduces into the output voltage. There are some applications where the bias current places the ultimate limit on the performance of the device. For example, the lower limit to the sensitivity of current-to-voltage converters, trans-resistance amplifiers, is set by the input bias current. In order for the amplifier to work, the current entering the amplifier has to be small compared to the current flowing in the feedback resistor. In that sense, the bias current sets a lower limit on the current that you can measure with such a device.

Consider the ionization gauge, a device to measure ionizing radiation such as X-rays or γ-rays (Figure 21-14). In this, a loop of wire at high voltage is put round the input lead of a trans-resistance amplifier.

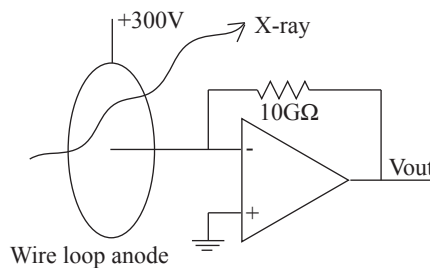


Figure 21-14 Ionization Gauge

X-rays that pass through the loop of wire knock electrons off a few of the atoms of the air near the loop. The strong electric field caused by the high voltage on the loop makes the electrons drift towards the loop and the air ions drift towards the input wire. When they hit the input wire they collect electrons from it to return to neutrality, causing a tiny current to flow in the wire. That current passes through the feedback resistor to produce the output voltage

$$out = -10G\Omega \times I_{in}$$

Each X-ray that passes through the loop leads to a few electrons being stolen from the input wire and so to a very, very tiny current. If we tried to use an LF356 for this task then the lowest current that it could measure with even 10% accuracy would be 1nA. 1nA of current corresponds to about 10^9 X-rays per second passing through the loop. That is so large that a synchrotron would be needed to generate the X-rays. For a task like this, you need the lowest input bias current op-amp that you can find. The best available today have typical bias currents of only 60fA, 6×10^{-15} A. With such an amplifier you could measure X-ray fluxes as low as 10^5 X-rays per second, well within reach of even moderate laboratory X-ray generators.

Summary

Real op-amps are imperfect creatures, striving to reach the ideals of the golden rules but falling short. Principle problems are

- Gain-bandwidth limit: sets the maximum frequency up to which the amplifier will work for a given gain.
- Slew-rate limit: the fastest rate at which the output can change. This reduces the effective bandwidth for large signals.
- Output current limit: maximum current that the op-amp can deliver to a load; usually about 20-25mA for a non-power device.
- Input Bias Current: the actual current that flows into the input of the op-amp in contravention of Golden Rule 1. Its effects can be minimized in the non-inverting configuration by adding a resistor between ground and the non-inverting input.
- Input Offset Current: the difference between the bias currents of the two inputs. It produces an error in the output voltage.
- Input Offset Voltage: the input voltage difference needed to make the output voltage 0. It leads to small errors in output voltage.

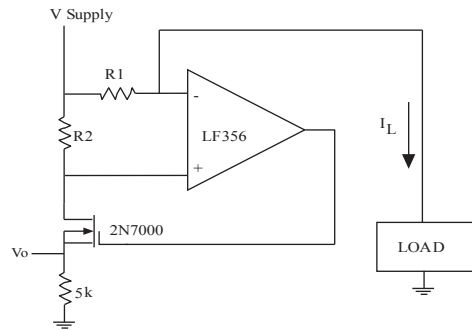
Each of the imperfections can be minimized by choosing the right op-amp but no one device can optimize more than one or two of the problems.

- High precision op-amps are usually not very fast and have poor current drive capabilities.
- Fast op-amps have relatively poor input offsets and biases.
- High power op-amps are slow and have high input offsets and biases.

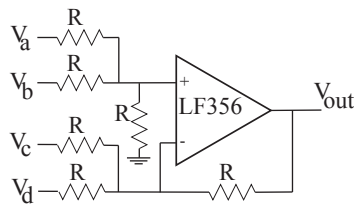
When you choosing an op-amp for a particular circuit you have to think carefully about which characteristics are important and which are not. In a precision DC amplifier you will need to use low offset op-amps and to correct for bias currents. In a video output amplifier you will need maximum speed and plenty of drive capability but will not care about the input errors. If you need the best possible characteristics in all respects then you will have to pay a lot of money and/or combine several devices. For example, you could use a high-power amplifier after a precision amplifier to get a high power device with precision inputs.

Exercises

1. The circuit below is designed to measure the current being drawn by the box labeled LOAD. The voltage at V_o is proportional to the current I_L drawn by the load. Use the op-amp golden rules to show how the circuit works and to derive the formula for V_o as a function of I_L .



2. The circuit of exercise 2 in Chapter 20 takes as input a voltage equal to the temperature in $^{\circ}\text{F}/10$ and outputs a voltage equal the temperature in $^{\circ}\text{C}/10$. (That is, a temperature of 56°F would go in as 5.6V and come out as 1.33V corresponding to 13.3°C). What is the largest error in the output temperature that could be caused by the 5% tolerance of the resistors?
3. Several small temperature sensors are available that produce outputs of $10\text{mV}/\text{K}$. Since $1^{\circ}\text{C} = 1\text{K}$ and 0°C corresponds to 273K , we can write the output voltage as $V = 2.73 + 0.01T$ where T is the temperature in $^{\circ}\text{C}$. This means that the temperature range from 0°C to 100°C corresponds to a voltage range of $2.73\text{--}3.73\text{V}$. A typical analog-digital converter has an input range of $0\text{--}5\text{V}$. Design an op-amp circuit that takes an input voltage in the range $2.73\text{--}3.73\text{V}$ and transforms it into the range $0\text{--}5\text{V}$. Hint: You will probably need a mixer and an inverting amplifier.
4. The circuit below is called an Adder-Subtractor. Use the Golden Rules to find a formula relating V_{out} to the 4 input voltages V_a , V_b , V_c , V_d .



Chapter 22:Comparators

22.1 The Open-Loop Comparator

There are a few practical circuits that use an op-amp without negative feedback. The most common is the comparator. A comparator is a circuit that compares two voltages and produces a binary output to tell which is the greater. This is exactly the behavior of an open-loop op-amp (Figure 22-1).

The basic theory of the op-amp says that $V_O = G \cdot (V_A - V_B)$ but the gain is too high to make this very useful as it stands. If $V_A > V_B$, the output voltage tries to get extremely high and of course it can't because the output is limited by the power supplies. Instead the output goes to V_{S+} and stays there. On the other hand, if $V_A < V_B$, the output heads for a massive negative voltage and is limited to V_{S-} . So the output is binary, though with rather different voltages from those we are used to. The op-amp is a comparator. It obeys the rule

$$V_O = V_{S+} \text{ if } V_A > V_B$$

$$V_O = V_{S-} \text{ if } V_A < V_B$$

which results in the transfer function plotted as Figure 22-2

The most common use for a comparator is with the inverting input held at a fixed **reference** voltage (V_R) and the non-inverting input connected to a signal (V_i). The output is then high when the signal is greater than the reference and low when the signal is less than the reference

Note that it is possible to turn this upside down and build an inverting comparator! If you use the non-inverting input as a reference and apply the signal to the inverting input then the output will be high when the signal is *less than* the reference, and vice versa..In that case the transfer function will be the other way up!

22.1.1 Problems with comparators

Running the op-amp in the open-loop mode means that there is enormous gain available and this leads to problems, particularly with slowly varying input signals. An ideal, slowly varying, input signal should lead to an output like that of Figure 22-3.

In practice, few input signals are this clean. Even if the signal source itself is ideal, there is noise in the input stages of the comparator that can lead to rather different behavior. Since the behavior is the same whether the noise arises inside the device or outside, Figure 22-4 is a typical view of what happens in a real situation.

This output is reminiscent of the output of a simple switch, where mechanical contact bounce leads to the same sort of messy output transition. If the output of the comparator goes to a slow circuit then this behavior may not be a problem. By the time the slow circuit has noticed what is happening the input change is complete and the comparator output is stable. However, if the comparator output goes to a fast circuit such as a counter then this situation leads to trouble. Consider a bicycle speedometer that measures the speed of the bicycle by counting how many spokes pass a fixed point in 1 second. If each spoke passing produced a noisy transition like that of Figure 22-4, then instead of seeing 1 spoke pass the counter would think that several spokes had passed. A 6 m.p.h. bicyclist could think that he was doing 25 m.p.h.!

22.2 Hysteresis to the Rescue

The cure for this is to add **hysteresis** to the comparator. Hysteresis means that the voltage at which the comparator switches from 0-1 is different from the voltage at which it switches from 1 to 0. Figure 22-2 was a plot of the output vs. V_i for a standard op-amp comparator and Figure 22-5 shows the same plot for a comparator with hysteresis.

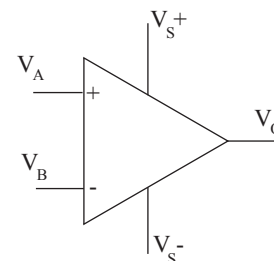


Figure 22-1 Open-Loop Op-Amp Comparator

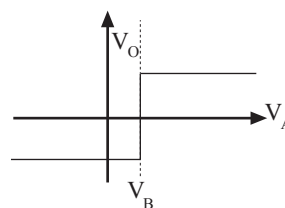


Figure 22-2 Comparator Transfer Function

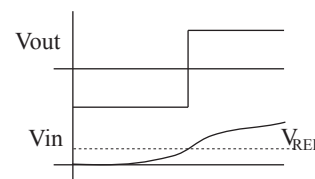


Figure 22-3 Output of Comparator for Ideal Input

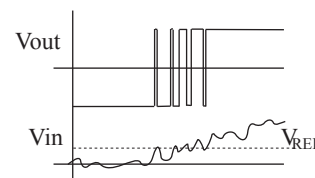


Figure 22-4 Output of Comparator for Noisy Input

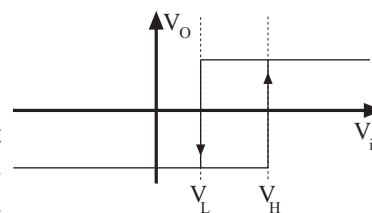


Figure 22-5 Transfer Function with Hysteresis

The single switching level, V_R , has now been replaced by two levels, V_L and V_H . Let us follow the behavior of the comparator as the input voltage changes from a value well below V_L to a value well above V_H .

Initially, the output of the comparator is at V_{S-} since the input voltage is very low. As we increase V_i it first reaches V_L . At this point nothing happens. The output remains at V_{S-} . We continue to increase V_i until it reaches V_H . At this point, the comparator switches from V_{S-} to V_{S+} where it remains through any further increase in V_i . Now if V_i falls again, something different happens. This time, when V_i reaches V_H the output remains high. V_i has to fall all the way to V_L before the comparator switches back to V_{S-} . A noisy input will not produce oscillations on the output unless the noise is big enough to cross the hysteresis gap V_A^- to V_A^+ .

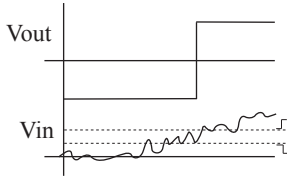


Figure 22-6 Output of Comparator with Hysteresis for Noisy Input

If we replace the comparator in the example shown in Figure 22-4 by a comparator with hysteresis then we get the remarkable improvement in the output shown in (Figure 22-6). The output does not switch from low to high until the noisy input passes the upper trigger level. Once the switch has taken place, the input can drop back below the trigger level without altering the output. The output will not switch back until the input falls below the lower trigger level. Thus, the noise would have to be larger than the hysteresis gap before it caused problems in the output.

22.2.1 Positive Feedback and Hysteresis

We produce hysteresis with positive feedback instead of negative feedback.

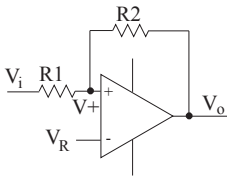


Figure 22-7 Op-Amp Comparator with Hysteresis

Op-Amp Comparator with Hysteresis shows the circuit for a comparator with hysteresis. It looks a lot like an inverting amplifier except that the inputs are reversed so that the feedback taken to the *positive input* instead of the negative input. This is a little more difficult to analyze than usual because the second golden rule *does not apply!* That rule only applies when the circuit is operating with negative feedback and this one is not. A complete analysis (see below) shows that the point at which the comparator switches from a negative output to a positive output (V_H in Figure 22-7) is given by

$$V_H = \frac{R1+R2}{R2} \times V_R - \frac{R1}{R2} \times V_{S-}$$

and the other switch point, V_L , by

$$V_L = \frac{R1+R2}{R2} \times V_R - \frac{R1}{R2} \times V_{S+}$$

Since we usually operate op-amps with $\pm 15V$ power supplies these equations normally become

$$V_H = \frac{R1+R2}{R2} \times V_R + \frac{R1}{R2} \times 15$$

and

$$V_L = \frac{R1+R2}{R2} \times V_R - \frac{R1}{R2} \times 15$$

It is useful to notice that the two set points lie equally spaced on either side of the voltage

$$\frac{R1+R2}{R2} \times V_R$$

The curious thing is that this central voltage is no longer V_R but is somewhat larger, amplified by the gain of a non-inverting amplifier.

In the most common situation, you only want to apply a very small hysteresis—a few mV is quite common—so that $R1 \ll R2$. In that case $(R1+R2)/R2 \gg 1$ and so the central voltage is essentially equal to V_R . We are back to the simple situation but with better noise performance.

Deriving the Comparator Voltages I

Although we cannot use the second Golden Rule, the comparator rule that we just found in section The Open-Loop Comparator still applies. In this case V_i and V_R play the roles of V_A and V_B , so that

$$V_O = V_S^- \text{ if } V_R > V_i$$

$$V_O = V_S^+ \text{ if } V_R < V_i$$

Of course, the actual input to the circuit is not V_i but V_+ , so we have to compute V_+ in terms of V_i . Because the first golden rule is still true, no current flows into the input of the op-amp. Thus

$$\frac{V_i - V_+}{R1} = \frac{V_+}{R2} \text{ OR } V_+ = \frac{R2 \times V_i + R1 \times V_O}{R1 + R2}$$

When $V_R < V_+$ the output is V_S^- . That will remain true until V_i reaches the switching point V_H , which will occur when $V_+ = V_R$. All during this time we have

$$V_+ = \frac{R2 \times V_i + R1 \times V_S^-}{R1 + R2}$$

so the output stays low until $V_+ = V_R$ and $V_i = V_H$ which occurs when

$$V_R = \frac{R2 \times V_H + R1 \times V_S^-}{R1 + R2} \text{ OR } V_L = \frac{(R1 + R2) \times V_R - R1 \times V_S^-}{R2}$$

If V_i rises above this value then the comparator will switch and the output will go to V_S^+ . This changes the relationship between V_i and V_+ . Now we have

$$V_+ = \frac{R2 \times V_i + R1 \times V_S^+}{R1 + R2}$$

So the output stays high until the next time that $V_+ = V_R$, which occurs when $V_i = V_L$, giving

$$V_R = \frac{R2 \times V_L + R1 \times V_S^+}{R1 + R2} \text{ OR } V_L = \frac{(R1 + R2) \times V_R - R1 \times V_S^+}{R2}$$

This gives us the voltage at which the output switches from V_S^+ to V_S^- . So the switch points are

$$V_{H'} = \frac{R1 + R2}{R2} \times V_R - \frac{R1}{R2} \times V_S^-$$

and

$$V_{L'} = \frac{R1 + R2}{R2} \times V_R - \frac{R1}{R2} \times V_S^+$$

22.2.2 tIC Comparators with Digital Output

There are some occasions when this sort of $\pm 15V$ output is exactly what we need, but more often we want a standard 0-5V output instead. It is quite easy to convert the high voltage output to a logic level output using an FET as shown in Figure 22-8.

The only trouble with this is that the output has now been inverted. When $V_A > V_B$, $V_G = +15V$, which means that the FET is turned fully on. Since the FET is on, current flows in the 1k resistor and V_O is brought down to 0V. On the other hand, when $V_A < V_B$, $V_G = -15V$ and the FET is turned off. With the FET off, the 1k resistor pulls the output up to +5V. Thus the output is now a standard logic level output but the meaning of the inputs has been reversed. The operating rule is now

$$V_O = +5V \text{ if } V_A < V_B$$

$$V_O = 0V \text{ if } V_A > V_B$$

So the positive, non-inverting, input of the whole comparator is V_B and the negative, inverting, input is V_A .

We can apply hysteresis to the digital-output comparator in two different ways. The simpler way to analyze is to add the hysteresis to the op-amp alone as in Figure 22-9.

In this case the set levels are exactly the same as those that we found for the simple comparator. The only change is the inversion of the output so that

$$\text{Output} = 1 \text{ if } V_i < V_L$$

$$\text{Output} = 0 \text{ if } V_i > V_H$$

V_H and V_L have the same values as before.

The other way is to take the positive feedback from the digital output. This alters the equations and is usually not worth the trouble. The only time we take the feedback straight from the digital output is when using an integrated, single-supply comparator, as in the next section.

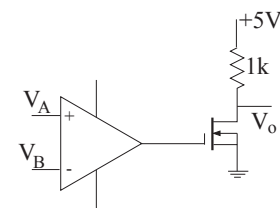


Figure 22-8 Comparator with Digital Output

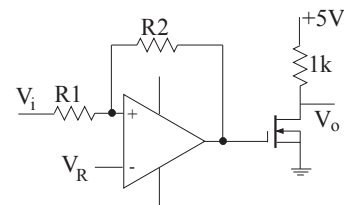


Figure 22-9 Comparator with Hysteresis and Digital Output

22.3 Commercial comparators

As you might expect, IC manufacturers make single chip comparators that generate a logic output. They are designed to switch much more rapidly than a comparator made from a standard op-amp and to be easy to connect to logic circuitry. A good example is the LM311, which can operate over a very wide range of supply voltages, from a single 0-5V logic supply to the ±15V supplies of a standard op-amp (Figure 22-10).

The output is a single transistor switch with one end connected to ground and the other left unattached so that you can generate a variety of different output voltages regardless of the power supply voltage. By itself the output of the LM311 does nothing. It requires a **pull-up** resistor to set the output voltage. This is a small resistor (usually about 1-10K) that pulls the output up to the desired voltage when the switch FET is off. When the FET is on, the FET pulls the output down to 0V. Thus we have the transfer function

$$V_o = V_D \text{ if } V_{Non} > V_{Inv}$$

$$V_o = 0V \text{ if } V_{Non} < V_{Inv}$$

where V_D is the digital power supply voltage to which the resistor is attached. This is usually 5V.

22.3.1 Adding Hysteresis to the LM311

We can add hysteresis to the LM311 in exactly the same way that we added it to the plain op-amp. We add two resistors to get the circuit of Figure 22-11

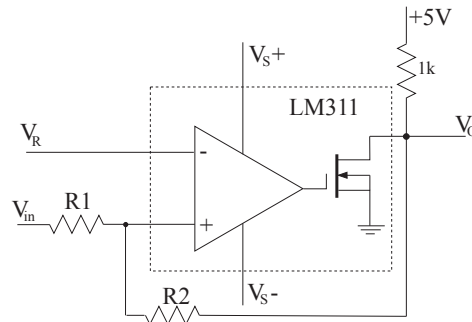


Figure 22-11 LM311 with Hysteresis

A simple adaptation of the general equations for a comparator with hysteresis gives us the rules for this circuit

$$V_o = +5V \text{ if } V_{in} > V_H = \frac{R1+R2}{R2} \times V_R$$

$$V_o = 0V \text{ if } V_{in} < V_L = \frac{R1+R2}{R2} \times V_R - \frac{R1}{R2} \times 5$$

22.3.2 Applying the Comparator

One common application for such a comparator is to change a small analog signal into a digital signal. For example, chapter 1 of the Introduction to Computers book shows a basic optical encoder that consists of a set of photodetectors and lights with a moving plate that interrupts the light falling on the photodetectors. A typical light source/photodetector arrangement (Figure 22-12) can produce a large output signal change under ideal conditions but a much smaller change under real conditions.

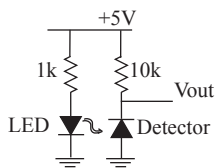


Figure 22-12 Photodetector

Using one such detector I obtained an output voltage of 0.72V when the light fell on the detector and an output of 5V when all light was blocked from the detector. However, when I interrupted the light beam with a piece of card while room light still fell on the detector the output voltage only changed from 0.72V to about 1.4V. That change is too small to pass to a logic input but a comparator can easily detect that difference and convert it to a logic level using the circuit of Figure 22-13.

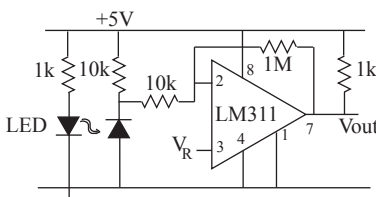


Figure 22-13 Opto-Detector with Logic Output

The 1MΩ feedback resistor combined with the 10k input resistor give a hysteresis of $10k \cdot 5V / 1M\Omega = 0.5V$, which is a little large but within the range of the input signal and it pro-

Warning The pin-out is very similar to the standard op-amp pin-out but there are some crucial differences. The positive supply has moved to pin 8 and the output to pin 7. Most confusingly, the inputs are switched; the non-inverting input is on pin 2 on the comparator and on pin 3 on an op-amp.

Warning This looks suspiciously like an inverting amplifier. Of course it is not. The key difference is that the feedback is taken to the positive input and not to the negative input.

vides excellent noise immunity. The final circuit produces a clean 0→5 transition every time a piece of paper is slid between the LED and the detector.

Summary

A comparator is an analog circuit that produces a digital output. It compares two voltages and produces a digital output depending upon which of the two voltages is the greater.

This simplest form of comparator is an open-loop op-amp. It compares the voltages on its two inputs, V_+ and V_- and produces an output

$$V_{out} = +15 \text{ if } V_+ > V_-$$

$$V_{out} = -15 \text{ if } V_+ < V_-$$

The output can be converted to a digital output (0-5V) by adding an FET switch to the comparator at the cost of switching round the inputs.

The simple comparator is very sensitive to noise on the input signals. The cure for this is to add **hysteresis** to the circuit. This makes the voltage at which the output goes from lo→high greater than that at which it goes from high→lo.

We add hysteresis to a comparator by adding **positive feedback** using two resistors, R_1 and R_2 . If the output levels of the comparator are V_{DD} and V_{SS} then the comparator equations become

$$V_{out} = V_{DD} \text{ if } V_{in} > V_H = \frac{R_1+R_2}{R_2} \times V_- - \frac{R_1}{R_2} \times V_{SS}$$

$$V_{out} = V_{SS} \text{ if } V_{in} < V_L = \frac{R_1+R_2}{R_2} \times V_- - \frac{R_1}{R_2} \times V_{DD}$$

Where V_{in} is the input to the resistor chain and V_- the voltage on the inverting input of the comparator.

Exercises

1. Design a circuit whose output will switch from 0-5V when its input increases past +3V and will then stay high until the input falls below -3V.
2. A photo-sensor acts as a very high ($>1M\Omega$) resistance when it is in the dark and decreases its resistance as light falls on it until it reaches about $10k\Omega$ when fully illuminated. Design a circuit to control a light based on the ambient light level. Use a photo-sensor as described, some resistors, and a comparator with a logic output to produce an explain a circuit that turns a light on when the sensor is in the dark and turns the light off when the sensor is brightly illuminated.
3. Add hysteresis to your circuit from problem 2 so that small changes in the light level near the set point don't make the light turn on and off. Once the sensor is dark enough to turn the light on, it should have to get quite a lot brighter to turn the light off again. Explain your choice of hysteresis value.

Chapter 23:Oscillators

23.1 Introduction

So far we have looked at circuits which modify existing electrical signals, usually signals coming from a signal generator. Now it is time to look at where those signals come from. A circuit that generates a time varying signal is called an **oscillator**. Oscillators are found in many kinds of electronic circuit. They are at the heart of both radio transmitters and receivers, electronic musical instruments are full of oscillators, and you hear the output of oscillators every time you use a touch-tone phone. In the digital world all sequential digital circuits rely on oscillators to generate their clock signals, oscillators drive all the electronic clocks and watches we see around us, and oscillators make all of our computers keep going.

There are two basic kinds of oscillators. The classic kind consists of an amplifier and a frequency selective circuit connected in a positive feedback loop. Such an oscillator runs as linearly as possible and generates a more or less pure sinewave. The other kind uses some essentially non-linear switching mechanism and drives itself back and forward between two states. The basic principle, which underlies both kinds of oscillator, is positive feedback—amplifying a signal that, after some modification, then becomes the input to the amplifier.

23.2 Linear Oscillators

This class of oscillator is based directly on the idea of positive feedback. If we take a linear amplifier and connect it in a feedback loop with a frequency selective circuit as shown in Figure 25-1 then we get an oscillator.

The best way to understand the working of the circuit is to start with a tiny sinewave, amplitude V_0 and feed that into the amplifier. Out comes a sinewave with amplitude $G \cdot V_0$ (Figure 25-2).

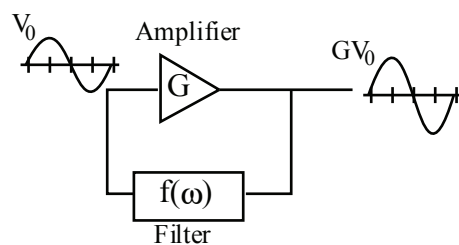


Figure 23-2

The amplified sinewave goes into the filter which attenuates it and may add a phase shift. The signal that comes out has amplitude $G \cdot f(\omega) \cdot V_0$ and this becomes the new input to the amplifier (Figure 25-3).

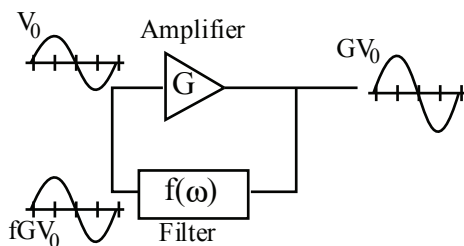


Figure 23-3

If, as shown in the figure, the new signal is larger than the starting signal then the amplitude will grow larger and larger forever.

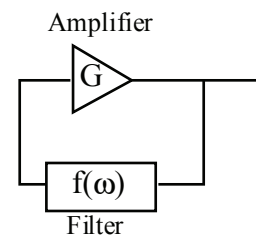


Figure 23-1 Generalized Oscillator

If the new signal is smaller than the starting signal then the signal will fade away to nothing. If the new signal is exactly equal to the starting signal then the system will sit happily circulating the same signal round all the time at a constant amplitude. That is a **stable** oscillator.

So , in order for an oscillator to operate at all we must have

$$G \cdot f(\omega) \geq 1$$

and the oscillator is stable only if the two sides are exactly equal.

An oscillator with a loop gain (G·f) of exactly 1 is very difficult to build. The slightest alteration in either f or G, which might be caused by, for example, temperature changes, will destroy the equality.

Such an oscillator is also impossible to start. We can make the oscillator self-starting by making the loop gain slightly greater than one. That way a stray bit of noise will be amplified, filtered, reamplified, refiltered, and so on. It will go round and round building up in amplitude and heading to a frequency for which $G \cdot f(\omega) > 1$.

Now the opposite problem arises, the amplitude keeps growing without bound and soon the amplifier will not be able to keep up. We have to add some small non-linearity to the system to limit the final amplitude. We must alter the circuit so that the gain decreases as the amplitude of the signal increases. That way, the gain at low amplitudes can be high enough to ensure good starting but the amplitude will not rise past the point at which the overall gain has fallen to one. After that, any increase in amplitude will lead to a reduction in the gain so the signal will fall back to its equilibrium value and vice versa.

The actual frequency of oscillation is controlled by the filter. The circuit is capable of oscillating at any frequency for which $G \cdot f(\omega) > 1$ but for oscillation we also require that the final wave, $G \cdot f(\omega) \cdot V_0$ be **in phase** with the starting wave. So the oscillator will oscillate at that frequency for which the total phase shift is 0. Note that every circuit contains a small amount of noise. That noise will nudge the operating frequency around in a random fashion. The size of this effect is affected by the range of frequencies for which the loop gain is > 1 . If the filter has a wide bandwidth then the oscillator will wander around through a wider range of frequencies than if the filter is very sharp. Thus the most precise and quiet oscillators, the best oscillators, have very sharp filters. The filter loss and phase shift change very rapidly with frequency.

Remember A high quality oscillator must have a very narrow filter in order to produce a pure output frequency.

23.2.1 The Wien-Bridge Oscillator

This is the classic sinewave oscillator. It uses a simple filter invented by Wien and shown in Figure 25-4.

Note This filter has been drawn backwards. The input is on the right and the output is on the left. This is to make it easier to see how the filter is used in the final circuit.

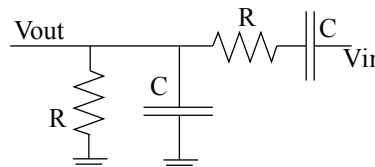


Figure 23-4 Wien-Bridge Filter

Using advanced analysis methods we can find the transfer function and the phase shift for this filter. The amplitude portion is

$$V_{//} = \frac{V_{out}}{V_{in}} = \frac{\omega RC}{\sqrt{1+7\omega^2 R^4 C^2 + C^4 R^4 \omega^4}}$$

and the phase shift is

$$f = \tan^{-1} \left[\frac{1 - C^2 R^2 \omega^2}{3CR} \right]$$

Those are very nasty equations. The best way to understand them is to look at the Bode plot of Figure 25-5.

The phase shift changes from $+90^\circ$ to -90° quite rapidly, making most of the change over two decades of frequency centered on $\omega RC = 1$. At the same time the amplitude has a broad maxi-

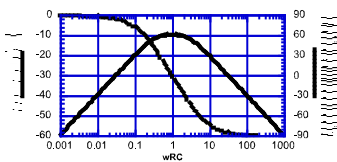


Figure 24-5 Wien Filter Bode Plot

Figure 23-5 Wien Filter Bode Plot

imum centered at the same place. The loss at the top of the maximum (actually the minimum loss) is -9.54dB corresponding to a gain of 1/3 . So this will make a good oscillator if we team it with an amplifier whose gain is 3.

We can use a standard non-inverting op-amp circuit but we need to modify the circuit slightly so that the gain depends on the signal level. We must provide a non-linear element to control the gain. The usual choice is a light bulb. A light bulb has a low resistance when it is cold (small signal flowing through it) and a higher resistance when it is hot (large signal flowing through it). Consider the circuit of Figure 25-6. This is a standard non-inverting amplifier and so it has a gain of

$$G = \frac{R_f + R_L}{R_L}$$

Now if we choose R so that $R = 2R_L$ when the lamp is just lit then the gain will be exactly 3 at that point. For signals that are too small to light the lamp the resistance R_L will be lower and the amplifier gain will be > 3 . For signals larger than the minimum needed to light the lamp the resistance R_L will be higher and the amplifier gain will be reduced to < 3 . Thus the gain of the circuit will adjust itself according to the average size of the signal and will stabilize the output at a desirable level. The lamp is an ideal element for this sort of circuit because it does not change its temperature quickly.

If we put these two circuits together we get the standard Wein oscillator circuit (Figure 25-7), the standard for high purity sinewaves at audio frequencies.

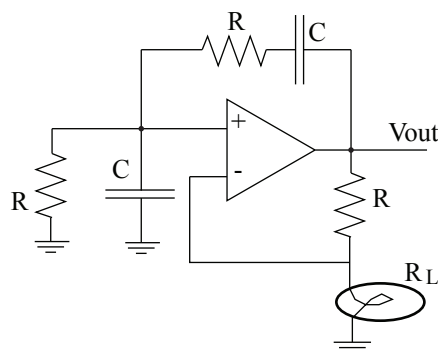


Figure 23-7 Wein-Bridge Oscillator

This circuit will oscillate at the frequency

$$f = \frac{1}{2\pi RC}$$

The circuit is reasonably tolerant of mismatches in the resistors R and C and can be tuned to different frequencies if you use a pair of variable resistors that are mechanically connected so that all changes are made to both resistors at the same time.

23.2.2 The L-C Oscillator

The Wein oscillator is mostly used at low frequencies, up to about 1Mhz. Above that we usually use LC oscillators. These are more easily tuned and do not need such high quality amplifiers. These are based on the classic parallel LCR tuned circuit of Figure 25-8.

So long as R is small (usually just the resistance of the wire from which the coil is wound), this combination has an impedance which is sharply peaked at a single frequency as we see in Figure 25-9.

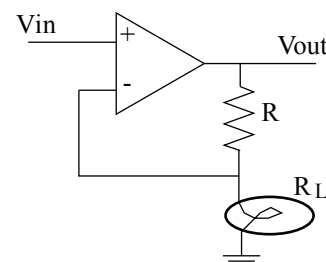


Figure 23-6 Non-linear Amplifier

Note The gain control circuit must operate very slowly compared to the frequency of the oscillator otherwise the output signal will be distorted. If the gain varies during the cycle, or even from one cycle to the next, then the amplitude of the output will vary and the output will not be a pure sinewave. In order to keep the harmonic distortion of the sinewave low, to keep the sinewave as pure as possible, the gain should be essentially constant over hundreds of cycles of the sinewave.

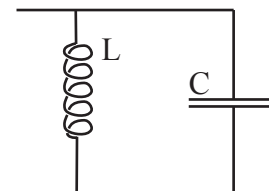


Figure 23-8

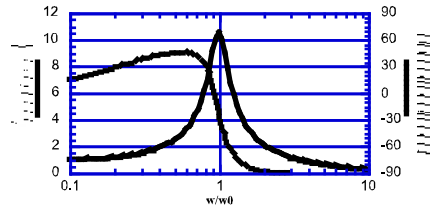


Figure 23-9

Both the amplitude and the phase change very rapidly near the frequency $\omega_0=1/LC$. The variation is very much more rapid than that for the Wein bridge filter we just discussed. Just how rapid depends on the $Q = L/CR^2$ of the circuit as Figure 25-10 shows

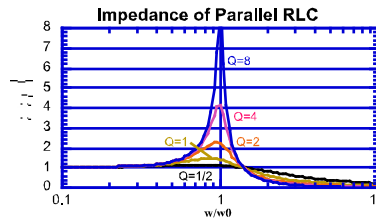


Figure 23-10

This rapid variation means that the oscillation frequency is very much more precisely defined, and the oscillator much more well behaved, than the Wein oscillator. There are several common circuits based on the tuned LC circuit.

23.2.3 The Colpitts oscillator

<USE INFO FROM ARRL HANDBOOK>

23.2.4 The Crystal Oscillator

A thin slab of quartz crystal placed between metal plates bends and stretches when a voltage is applied to the plates. This effect is called **piezoelectricity**. If the frequency of the applied voltage is just right then the quartz crystal will resonate mechanically, ringing like a tiny bell. The mechanical oscillation interacts with the electrical signal to make the quartz crystal behave like a tuned RLC circuit with a remarkable Q; values round 10,000 are typical! The crystal can replace the tuned circuit in an LC oscillator to give a sinewave oscillator with remarkable stability. A good crystal oscillator will only drift by a few ppm over times measured in hours.

<A couple of working crystal circuits>

Quartz crystal oscillators are widely used as clocks for digital systems such as computers and watches. These oscillators don't have to produce good sinewaves, indeed the final output should be as clean a square wave as possible. They do need to be as compatible with digital circuitry as possible and so most of them use the following kind of circuit.

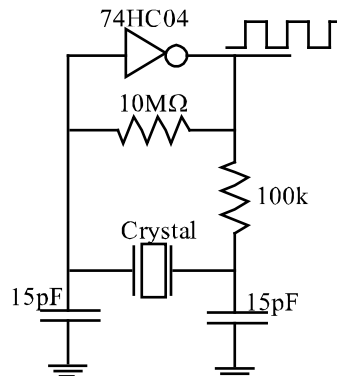


Figure 23-11

One of the great things about this sort of circuit is that it is simple enough to be designed into a computer chip so that you usually only have to add the crystal (and sometimes the capacitors) to have a complete clock circuit, all the rest is already there.

Rather than use a separate crystal and associated components you can buy complete digital crystal oscillators packed in a metal can. These usually have the same footprint as a standard digital IC and so are particularly easy to add to a circuit. They are available in a wide range of frequencies and only a little more expensive than the crystals alone.

23.3 Non-linear Oscillators

This class of oscillators, sometimes called **relaxation oscillators**, operate in a highly non-linear mode and produce square or triangle wave outputs. They normally operate by charging and discharging a capacitor between two limits set by comparators. The capacitor and charge circuitry are linear and the comparators provide the non-linearity. The simplest such circuit, though little used, is shown in Figure 25-12. Let us follow its operation because the more complicated oscillators work in a similar way.

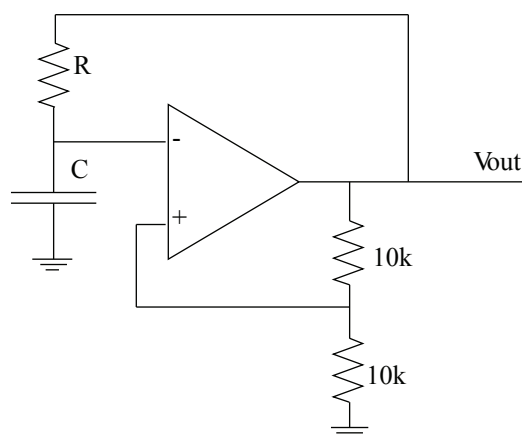


Figure 23-12 Relaxation Oscillator

First, we shall break the circuit up into its component sections. The first section is the R-C series circuit that we studied in some detail back in section 7.2. The other is a comparator like those we discussed in Chapter 22. We shall start by examining the comparator.

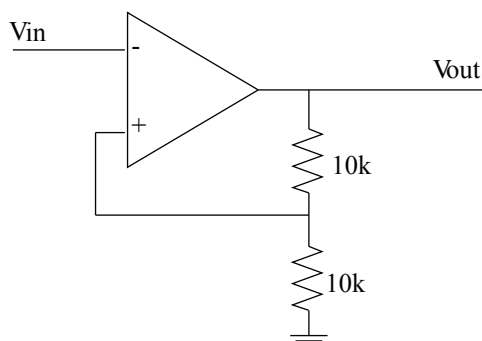


Figure 23-13

We know that, although this looks suspiciously like a non-inverting amplifier, it is really a comparator because the feedback is positive. Where are the switching points? Well, when $V_{out} = V_{DD}$ (the positive supply voltage) then $V_{+} = V_{DD}/2$ and when $V_{out} = -V_{SS}$ (the negative supply voltage) then $V_{+} = -V_{SS}/2$. So, assuming that $V_{SS} = -V_{DD}$, the two switching points are at $\pm V_{DD}/2$. The transfer function is shown in Figure 25-14.

When V_{in} is down near $-V_{SS}$ then V_{out} is high, at $+V_{DD}$. This high V_{out} makes the voltage at V_{+} positive and so the comparator stays in the high state until V_{in} reaches $+V_{DD}/2$.

As soon as V_{in} passes $+V_{DD}/2$, the comparator switches and the output goes to $-V_{SS}$. Now the voltage at V_{+} is $-V_{SS}/2$ and so the comparator will stay low so long as $V_{in} > -V_{SS}/2$. If V_{in} falls

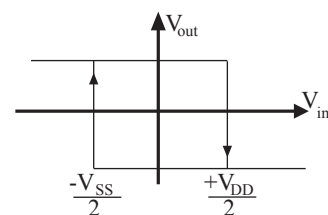


Figure 23-14 Hysteresis Curve

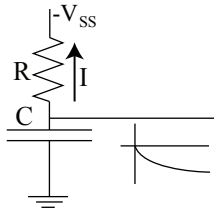


Figure 23-15 Charging the Capacitor to $-V_{SS}$

below $-V_{SS}/2$, the comparator switches back and we are back where we started from. As you see from the figure, there is a region between $V_{in} = -V_{SS}/2$ and $V_{in} = +V_{DD}/2$ where the comparator has two stable states. The oscillator operates in this region.

Now we are ready to examine the behaviour of the complete circuit. When we first turn the power on the capacitor C is discharged so the $V_- = 0V$. The comparator is in its bistable range and could turn on with either polarity, it is completely random. For the sake of our example we will assume that it starts with $V_{out} = -V_{SS}$ but you can easily check that everything works just as well if it starts out with V_{out} negative. So at the start we have the situation shown in Figure 23-15.

The capacitor end of the resistor is at a higher potential than the other end so current flows out of the capacitor, as shown, and the voltage on the capacitor starts to go negative. We are now in familiar territory. The RC circuit charges in its happy exponential fashion towards $-V_{SS}$ and drives V_- negative as shown.

The capacitor is trying to charge to $-V_{SS}$ so fairly soon V_- reaches the trigger point at $-V_{SS}/2$. When this happens the comparator switches state and the voltage at the top of the resistor rises to $+V_{DD}$.

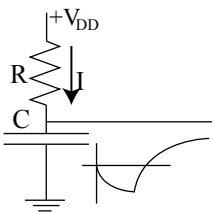


Figure 23-16 Charging the Capacitor to $+V_{DD}$

Now the top of the resistor is positive with respect to the bottom. The current reverses and the capacitor starts to charge up towards $+V_{DD}$ (Figure 23-16). Now the voltage at V_- starts to rise. It can rise past the original trigger point because of the hysteresis. V_- rises, crosses 0V and keeps rising until it eventually reaches the upper trigger point at $+V_{DD}/2$. As soon as V_- crosses the upper trigger, the comparator switches and V_{out} falls to $-V_{SS}$ again. Now V_- heads back towards $-V_{SS}$ and the cycle starts over.

Figure 25-17 shows both the output voltage and the capacitor voltage as a function of time so that you can see how they are related.

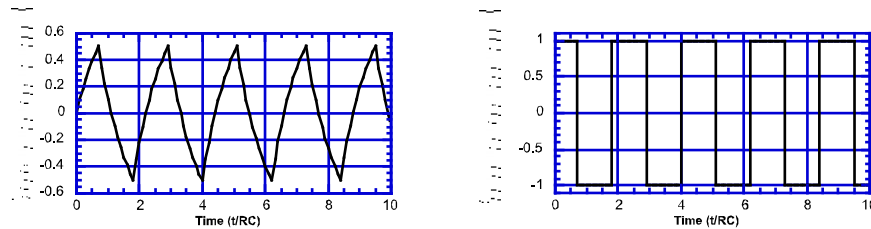


Figure 23-17 Capacitor and Output Waveforms

Note Because the output has to swing all the way from $-V_{SS}$ to $+V_{DD}$, the slew-rate limit of the operational amplifier is the major limitation on how high a frequency you can get from this circuit!

The very first fraction of a cycle is different from all the rest but after that one the oscillator settles down and produces regular oscillations with a period of $2 \cdot \ln 3 \cdot RC = 2.2RC$. The output swings over the full power supply voltage range while the voltage on the capacitor oscillates through the hysteresis interval of the comparator.

23.3.1 The 555

As you might expect, IC manufacturers have provided us with a wide range of specialized oscillator chips to choose from. These range from simple square wave relaxation oscillators to elaborate voltage controlled systems with square, triangle, and even sine outputs. By far the most popular are the 555 and its many relatives.

In the beginning (well, 197?) was the LM555.

This chip can perform a number of functions including generating single pulses as well as operating as an oscillator. Rather than using a single comparator with hysteresis, the 555 uses a pair of comparators and some logic to set the interval within which it oscillates. Here is the internal circuit

23.3.2 Voltage-Controlled Multivibrators

23.3.3 Frequency Synthesizers

Summary

An **oscillator** is a circuit that generates a periodic wave all by itself. All oscillators make use of **positive feedback** to create a self-sustaining signal. We distinguish two kinds of oscillators.

Linear Oscillators operate in a fully linear fashion and combine a linear amplifier with a frequency selective network (filter). They produce sine waves.

Non-Linear Oscillators use at least one non-linear element. They usually combine a non-linear amplifier with some form of time varying circuit. They produce square waves, triangle waves, sawtooth waves, or other non-sinusoidal waves.

A Linear Oscillator: The Wein Bridge Oscillator

This uses a simple RC bandpass filter to set the operating frequency. The filter has a slowly varying gain curve but a rapidly varying phase that reaches 0° at the same time that the filter gain reaches its greatest value, $1/3$.

Since the filter has a gain of $1/3$ we can create a self-sustaining oscillation if we couple the filter with a gain of $+3$ amplifier. This gives us a Wein Bridge oscillator. In order to make the output stable we have to make the gain of the amplifier depend upon the amplitude of the signal. If the signal level falls we want the gain to rise in order to restore the signal. If the signal gets too big we want the gain to fall so that the signal will shrink back to its proper size. The standard solution is to use a light bulb for the lower resistor in the non-inverting amplifier. As the signal level rises the light bulb filament will warm up and its resistance will increase, decreasing the gain of the amplifier and producing the stabilizing effect.

The circuit will oscillate at the frequency for which the phase shift round the filter is exactly 0° . That is

$$f = \frac{1}{2\pi RC}$$

The resistor and lamp must be chosen so that the gain is 3 at the operating amplitude.

A Non-Linear Oscillator: The Relaxation Oscillator

The relaxation oscillator is a classic form of non-linear oscillator. It uses an RC series circuit to generate a steadily rising/falling voltage. That voltage is passed to a comparator with Hysteresis and the output of the comparator is used to drive the RC circuit, completing the feedback loop. The oscillator operates between the upper and lower hysteresis set points of the comparator and generates a square wave out from the comparator output and a new triangle wave at the output of the RC circuit.

Chapter 24: Linear Regulated Power Supplies

24.1 Introduction

Back in Chapter 10 we studied the design of unregulated power supplies, power supplies that have no way to control their output voltage as the current drawn from the supply varies. Such supplies produce fairly poor quality power. The output voltage varies with the current drawn from the supply—falling as the current increases—and with any variations in the input voltage such as the ripple from the AC supply. In this chapter we shall see how to use operational amplifiers to control the output voltage or current accurately despite wide variations in the source voltage or load current. It will be our aim to design supplies that can produce any desired output voltage and hold it steady to within <1% while the current drawn goes from 0 to several amperes.

Regulated power supplies use feedback to keep the output voltage steady. A controller monitors the actual output voltage and compares it with an internal standard then it adjusts the resistance of a control element to make the output voltage stay at the correct value. So, if we want to produce precisely controlled power supplies then we need a standard of voltage against which to measure the output.

24.2 Voltage references.

There are two main types of voltage reference in common use. The first kind relies on the diode forward current curve with its steady 0.6V drop across the diode. The second relies on the controlled reverse breakdown voltage of a Zener diode. Power supplies are more likely to use a Zener reference so we shall first look at those and then come back to the forward-drop references when we talk about IC references in section 24.3.

24.2.1 Zener references

Back in chapter 9 we met the Zener diode; a diode whose reverse breakdown voltage is carefully controlled in the manufacturing process. Zener diodes are available with breakdown voltages from 2V to 200V in a set of standard values, just like resistors. They can have power ratings from <1W to about 50W but the higher voltages and powers are very expensive and have fairly poor characteristics. The best Zeners, and the most common, have voltages around 6V and power ratings of a fraction of a Watt. If we look at the I-V curve for a typical Zener diode, the 1N715A, then we can see both its features and some of its flaws (Figure 27-1). The reverse breakdown voltage is clearly visible as a “knee” in the curve at -5.1V. Below this voltage the diode does not conduct at all and above this voltage it conducts so well that the voltage across the diode is nearly constant regardless of the current. It is that constant voltage that we use as a voltage reference.

When we look more carefully we see that the knee is not perfectly sharp; the current below the knee is not zero and the I-V curve beyond it is not perfectly vertical. This means that the voltage across the diode is not strictly constant. Thus, if we want a good, stable, repeatable reference voltage then we have to control the current through the diode quite carefully. We need to choose an operating current that is large enough that the curve is as vertical as possible, while keeping the current low enough that the diode will not overheat. Fortunately, the manufacturer usually tells us the best operating current for the diode.

Two flaws that we cannot see from this graph are the variation of diode voltage with temperature and the variation from one unit to another. The whole I-V curve moves from side-to-side as the operating temperature of the diode changes resulting in a temperature coefficient for the

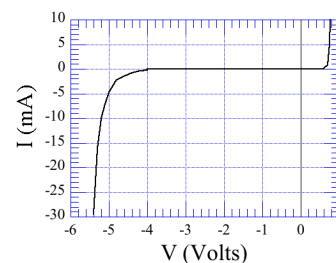


Figure 24-1 I-V curve for a 1N715A 5.1V Zener

Note We characterize the deviation from vertical of the Zener curve using the slope resistance. This is simply the resistance for a straight line fit through the moderate current portion of the Zener's I-V curve. It is given the symbol R_z and should ideally be zero. Our 1N751A has a slope resistance of about 15 Ω .

Note A coefficient of x ppm means $\alpha = x \cdot 10^{-6}$ so that, e.g., a coefficient of 100ppm means $\alpha = 100 \cdot 10^{-6} = 10^{-4}$. For example a 5.1V Zener with a 100ppm/ $^{\circ}\text{C}$ temperature coefficient will change its output voltage by 0.51mV for every 1°C change in its operating temperature.

diode voltage that ranges from $\pm 5\text{ppm}/^\circ\text{C}$ for the best diodes (e.g. the 1N4895) to more than $0.1\%/^\circ\text{C}$ for the poorest.

The change in output voltage is given by the formula

$$\Delta V = V \cdot \alpha \cdot \Delta T$$

where V is the diode voltage, α is the temperature coefficient, and ΔT is the temperature change in $^\circ\text{C}$.

The variation from one unit to the next varies with the price of the diode. Cheaper devices may vary as much as 20% from one unit to the next while the more expensive models vary by only 1%. This variation is mostly a problem when you are trying to make a large number of identical power supplies. When you are only trying to make one or two you can just buy several diodes and choose the ones that have the best voltages.

Finally, many Zener diodes are quite noisy; the output voltage fluctuates randomly and rapidly over some small range. An average Zener diode produces several microvolts of noise at frequencies from below 1Hz up to the MHz region. Indeed, at least one kind of Zener diode is sold specifically as a noise source! This noise is intrinsic to the physics of the device and will be transferred to the output of the power supply. Thus so precision supplies need to use low noise devices. Because the noise arises principally on the surface of the semiconductor, Zeners that are buried inside the silicon of an IC rather than formed on the surface are much quieter than normal Zeners. These **buried Zeners** have the best noise performance and are the devices of choice for low noise operation.

24.2.2 Using a Zener diode

We can make a simple voltage reference using a Zener diode in series with a current limiting resistor (Figure 27-2). We choose the resistor R to supply the desired current to the diode. Using Ohm's law we have

$$R = \frac{V_+ - V_Z}{I_Z}$$

Example

The 1N821 is a 6.2V diode designed to operate at 7.5mA current that has a slope resistance near 7.5mA of 15Ω . So, if we want to operate that from a 15V supply, then the resistor has to drop $15 - 6.2 = 8.8\text{V}$ at a current of 7.5mA. So we need a resistor of

$$R = 8.8/0.0075 = 1173\Omega$$

The nearest standard value is 1200Ω which will give us a current of $8.8/1200 = 7.33\text{mA}$. What effect does that have on the output? Well, the incremental form of Ohm's law tells us that

$$\Delta V = R_Z \cdot \Delta I$$

so that the voltage error is $15\Omega \cdot 0.17\text{mA} = 2.5\text{mV}$. If we want to do better then we should use a variable resistor and trim it to give exactly the desired voltage.

This circuit has the problem that variations in the supply voltage will pass through, diminished in size, to the output voltage. In general the variation ΔV_o in the output voltage due to a variation ΔV_i in the input voltage is

$$\Delta V_o = \frac{R_Z}{R_Z + R} \times \Delta V_i$$

where R_Z is the diode slope resistance and R the value of the current limiting resistor.

Example

If there is 0.5V of ripple on V_+ then the current through the Zener will vary by about $0.5\text{V}/1200\Omega = 0.42\text{mA}$. That current ripple will make the output voltage vary by $15\Omega \cdot 0.42\text{mA} = 6.25\text{mV}$. So the output voltage still has the ripple, though the depth has been decreased from 3% to 0.1%. Still, a voltage reference that ripples is less than ideal.

24.3 IC references

The ordinary Zener, with its sensitivity to variations in supply voltage and temperature, is clearly far from a perfect reference. Integrated circuit manufactures have seized this opportu-

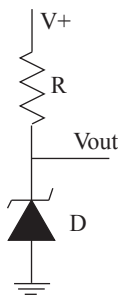


Figure 24-2 Zener reference

nity and provide a wide range of special purpose voltage reference IC's. Internally these use either Zener diodes, usually buried Zeners, or forward biased diodes, with their constant 0.6V forward voltage drop. These are the so-called **bandgap references**. The IC's include circuitry to control the diode current very carefully, isolating it from changes in the supply voltage, and play clever tricks to cancel the temperature coefficients of their reference elements by mixing them with other temperature dependent components.

Example

A Zener diode with a positive temperature coefficient can be put in series with a forward biased diode, which has a $-2.1\text{mV}/^\circ\text{C}$ temperature coefficient. If you choose the Zener voltage carefully, then you can get very good cancellation and produce a circuit where the output voltage has an overall temperature coefficient of only $1\text{-}2\text{ppm}/^\circ\text{C}$.

24.3.1 Bandgap references

We saw in Chapter 9 that a diode has a forward current-voltage curve with a rather sharp turn-on so that the voltage drop across a diode varies only slightly with the current through the diode. If we control the current through the diode then we can build a 0.6V reference source. Because the diode I-V curve is far from vertical near 0.6V the output voltage will vary somewhat with diode current and so we must make sure the current is held quite steady. Once that condition is met, a diode makes a very stable voltage reference. The main limitation is that the turn-on voltage varies with the temperature of the diode, falling 2.1mV for every 1°C rise in temperature.

There are some clever tricks that integrated circuit manufacturers play to offset the temperature coefficient by adding in a voltage with a positive temperature coefficient that just cancels the diode's own negative one. They also add circuitry to produce larger output voltages with the same stability as the internal reference.

Note The $-2.1\text{mV}/^\circ\text{C}$ temperature coefficient of a diode forward-drop reference can be used to measure temperature as shown in Chapter 23.

Table 24-1: 1 Some Voltage References

Model	$V_{\text{REF}}(\text{V})$	I (mA)	Tempco (ppm/ $^\circ\text{C}$)	$\Delta V/\Delta I$ (mV/mA)	Approx Price (\$/2000)	Comments
1N5231	5.1	20	300	17	0.2	Simple Zener
LM385-1.2	1.2	$10\mu\text{A}$ -20	150	0.4	2	Bandgap reference
LT1004-1.2	1.2	$10\mu\text{A}$ -20	20	0.2	2	Improved bandgap ref.
LM399H	6.95	.5-10	0.3	0.5	8	Heated buried Zener
LT1019	2.5	.5-10	<2	0.02	11	Heated bandgap ref

These **bandgap references** come in small 2-pin or 3-pin packages that look just like transistors. You connect them in series with a current setting resistor and they produce a steady output voltage. Bandgap references are particularly well suited for battery-powered equipment as they retain their stability and precision down to very low currents. For example, the LT1004 operates happily from 20mA all the way down to $10\mu\text{A}$. It comes in two different models. One produces an output of 1.235 ± 0.004 Volts and the other an output of 2.50 ± 0.02 Volts. These have temperature coefficients of only 20 ppm/ $^\circ\text{C}$ so that the 1.235V output voltage changes by only 0.25mV for a 10°C change in temperature! The LT1019, which comes in 2.5V, 4.5V, 5V, and 10V models, offers even better performance at higher cost. Its temperature coefficient is $<2\text{ppm}/^\circ\text{C}$! Exceptional performance like this is not usually needed in power supplies but we shall see references with this sort of performance when we study digital-analog converters in Chapter 25.

24.4 A simple voltage regulator

Now that we have a range of voltage references we can go ahead and produce a complete voltage regulator. Figure 27-3 shows the simplest circuit to start with.

Note The op-amp used must be able to work with its inputs all the way down to its negative supply voltage. There are special single-supply op-amps such as the LM324 that are designed specifically for this purpose.

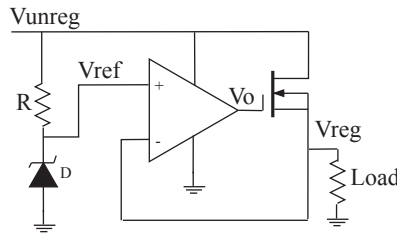


Figure 24-3 Simple Voltage Regulator

This does not work very well! The op-amp tries to make its two inputs equal in voltage. The non-inverting input is held at the reference voltage, V_{ref} , by the Zener, D. The inverting input is at the output voltage, V_{reg} . If the output voltage is below V_{ref} then V_o rises, turning the FET on more and causing V_{reg} to rise. Similarly, if $V_{reg} > V_{ref}$ then V_o falls, the FET moves towards cut-off and the output voltage falls.

The major problem with this circuit is that the reference is driven from the unregulated supply and so will pass all the ripple and other supply voltage variations on to the output of the regulator. We can improve this circuit either by replacing the Zener reference with an IC reference or by replacing the resistor with a constant current source. A good constant current source will hold the current flowing in the diode at a constant value despite variations in the supply voltage. One of the easiest ways to do this is to supply the diode current from the reference voltage rather than from the unregulated voltage, as in Figure 27-4. That sounds circular but it works.

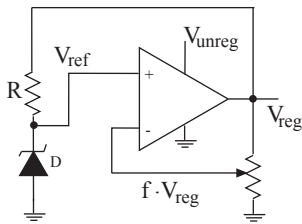


Figure 24-4 Improved Voltage Reference

Once this circuit is working it is easy to see how it functions. The potentiometer samples off a fraction, f , of the op-amp's output voltage for the inverting input. The op-amp will adjust its output voltage to make the voltage at the inverting input equal to V_{REF} so that the output voltage is regulated at V_{REF}/f . Since the regulated voltage is $>V_{REF}$, it can supply the power to the current control resistor, R_D . Now variations in the supply voltage do not affect the current through the Zener and so do not affect the output voltage. Note that again we must use a single-supply op-amp because the inverting input has to operate all the way down to 0V.

The circuit is self starting because, when power is first supplied, the op-amp's output will tend to head towards the middle of the supply range causing some current to try flow to the diode. If this initial voltage is high enough, the diode will start to conduct and the positive input will go straight to V_{REF} and then all is well. If the initial voltage is too low to turn the diode on, the non-inverting input will be driven to V_{out} while the inverting input is driven only to $f \times V_{out}$. Since V_{in+} is now $>V_{in-}$, the output voltage will rise and will keep rising until the diode turns on and normal operation starts.

You can take the output from this circuit either from the diode itself, giving you a reference voltage of V_{REF} or from the output of the op-amp, V_o , giving you a reference voltage V_{REF}/f . This is useful for obtaining any desired reference voltage from a single Zener. Obviously, the noise performance and temperature coefficient are still those of the underlying Zener, so that you should choose a temperature compensated, buried Zener reference for the best performance.

24.4.1 An improved voltage regulator

If we put our improved voltage reference into our regulator and add a couple of extra components we get a useable voltage regulator (Figure 24-5)

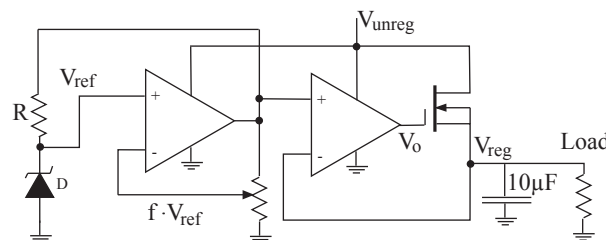


Figure 24-5 A Practical Voltage Regulator

The principle is the same as before but we have three improvements. First, the reference is now driven by a stable supply and so the output voltage is extremely well isolated from the unregulated power supply. Second, we can choose the output voltage independently of the reference voltage by selecting R1 and R2 or by adjusting f . The actual output voltage is

$$V_{REG} = \frac{R1 + R2}{f \times R2} \times V_{REF}$$

Third, we have added a capacitor to the output to help mop up any residual noise and to improve the transient response. This should be a high quality capacitor such as a tantalum electrolytic.

24.5 Current regulation

Our regulator is nearly complete. All it lacks is some protection from the inevitable accidents that can destroy a real world device. In particular, the output transistor is easily destroyed by excessive current draw. It is quite easy to short the output of a regulator and so apply the full unregulated supply voltage to the transistor with nothing to limit the current. This will result in excessive power dissipation in the transistor and will probably destroy it. The cure is to limit the maximum current that the circuit can deliver. There are two common ways to do this. Some fairly expensive laboratory power supplies offer an adjustable current limit that allows the supply to be used a constant current source as well as a constant voltage source. If we plot the output voltage as a function of the load current for such a device then we get a curve like Figure 27-6.

The output transistor for such a supply has to be very rugged and very well cooled. The maximum power it has to dissipate is $V_{unreg} \times I_{MAX}$ when the output is short circuited. This means that a heavy duty power supply such as the Kepco 100-10M, which is rated to deliver 100V @ 10A, has to dissipate more than 1kW under worst case conditions. Not surprisingly, it weighs 60lbs and has fans and massive heat sinks for its multiple output transistors

If we do not require the current limited mode of operation then we can save a lot on the robustness of our output transistor by using **foldback current limiting**. In this case, once you try to draw more than the rated current from the supply, the current limiting circuit drastically reduces the current down towards some safe value for the short-circuit current, I_{SC} . Figure 27-7 is the V-I curve for such a device

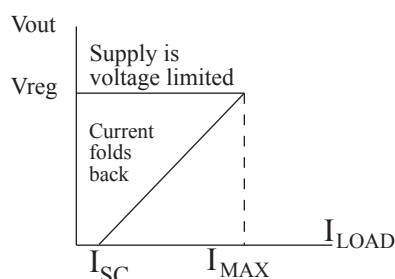


Figure 24-7 Fold-Back Current Limiting

When the current rises to I_{MAX} the protection circuit starts to work, limiting the current to a lower value. The more severe the short—the lower the output voltage across the short—the lower the current limit. This foldback current limiting is the usual way to protect voltage output power supplies. It is achieved by sensing the output current and when it gets too large drastically lowering the drive to the output transistor, thus shutting off the output current. The circuits for this are somewhat beyond the scope of this book but they are built into the commercial voltage regulator ICs that we normally use.

24.6 IC voltage regulators

The tasks of voltage regulation are so common and so standard that there are special purpose IC's available to meet most needs. These usually incorporate all the best reference and pro-

Info **Transient response** is the ability of the regulator to respond to sudden changes in the output current. If the output current draw suddenly rises—digital circuits can cause changes of 10s of mA on a 10ns time scale—then the feedback loop takes some time to respond. Thus the output voltage drops and then recovers as the amplifier notices and restores the regulation. Since the op-amp in the regulator has a finite frequency response it may take many μS for the regulator to respond to a really rapid change in current. A capacitor across the output can supply the excess current for those few microseconds and so prevent the voltage from dipping. Looking at it another way, the capacitor forms an RC filter with the output impedance of the regulator and this will not let the output voltage change over times much less than RC.

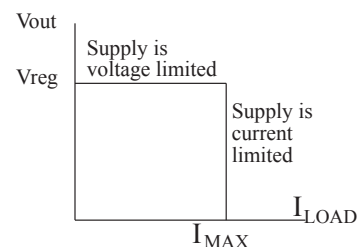


Figure 24-6 Current Limited Supply

tection technologies and are the recommended method of building all but the fanciest power supplies. They are available in a wide range of operating currents and voltages, including both fixed voltage and variable voltage models. We will look at several of the most popular, all of which have only three terminals and need only two or three external components to make a high quality regulator.

24.6.1 The LM340

This is a fixed, +5V, regulator available in several different packages and so in several current ratings. It is standard practice to package the same chip in several different ways to make a family of IC's, ranging from small low current devices to large, high current ones. The current rating of one of these regulators is controlled by how well the package allows heat to be removed from the chip. A regulator in a tiny plastic transistor package has a current limit of 100mA, while the same device in a hefty metal TO-3 package, connected to a large heat sink, can carry up to 1A.

The LM340-5 can operate with no external components but it is best to add an output capacitor to improve the transient response. If the regulator is put on a circuit board some distance from the filter capacitors of the unregulated supply then it is also a good idea to add a small input capacitor. Figure 26-8 below shows the circuit for an LM340-5 regulator.

Note The chip does not have two separate ground connections. The connection shown at the top of the package is actually a large metal tab that lies under the whole package and is internally connected to the center pin. This tab is used to carry heat away from the actual IC chip which is bonded to the tab. In the LM340 regulators the tab ends up being connected to ground which is quite convenient. The negative regulators of the related LM320 family have the unregulated input voltage on the tab which is a lot less convenient!

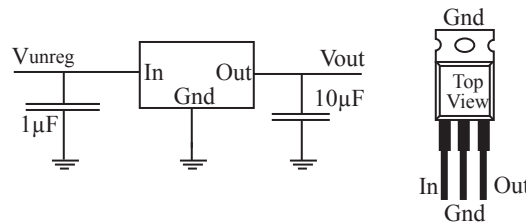


Figure 24-8 LM340-5 Three-terminal Voltage Regulator

A three-terminal regulator like this provides all the features that we can want in an everyday regulator. The internal reference is compensated to provide an overall temperature coefficient of only 0.6mV/°C or 120ppm/°C. It provides excellent isolation from ripple on the input and is fully protected against abuse. Internal circuitry limits the output current and shuts the device down if it overheats or is in danger of transistor burnout.

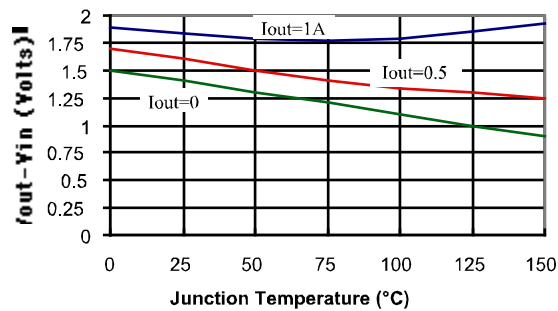


Figure 24-9 Variation of LM340-5 Drop-Out Voltage with Temperature

These devices are not perfect and we do have to be aware of some elementary limitations. The unregulated input voltage has to be large enough not only to provide the output voltage but also to operate the regulator's own circuitry. The **dropout voltage** is the voltage below which the device will not operate. It varies slightly with output current and operating temperature, as shown in Figure 26-9, but in general is about 2V more than the output voltage.

At the other end, the input voltage cannot be too high or the regulator will have to dissipate too much power. Remember that the output transistor must dissipate $I_{out} \cdot (V_{in} - V_{out})$ and the chip will suffer permanent damage if the internal temperature, the **junction temperature**, rises above 150°C. Thus, although an LM340-5 can tolerate an unregulated voltage up to 35V, a value of about 8V is ideal.

As well as needing a certain minimum voltage to operate, the chip consumes some current of its own in addition to that which it supplies to the load. This current is called the **quiescent current** because it is the current that flows when the chip is doing nothing—when the chip is **quiescent**. For the LM340-5, the quiescent current varies from 5mA to 5.5mA as the input voltage rises to the maximum of 35V. This is not important in equipment to be operated from the household current but is too large to make the chip a good choice for battery powered equipment. There are special micro-power regulators with ultra-low dropout voltages and quiescent currents for battery powered devices but a good switching regulator is often a better choice (see Chapter 30).

The LM340-5 is only one of a whole family of fixed output positive and negative voltage regulators. A survey of some of the more common types is shown in Table 24-2. They are all very similar to the LM340-5 in operation and characteristics.

Table 24-2:
3-Terminal Voltage Regulators

Model	V_{REG}	I_{MAX}	Case
LM340L-xx	5V,12V,15V	0.1A	TO-92
LM340T-xx	5V,12V,15V	1A	TO-220
LM323	5V	3A	TO-3
LM320L-xx	-5V,-12V,-15V	0.1A	TO-92
LM340T-xx	-5V,-12V,-15V	1A	TO-220

24.6.2 LM317 3-terminal adjustable regulator

Sometimes, a fixed voltage regulator just won't do; either because you need to keep changing the voltage or because you just need a voltage that the fixed regulators don't provide. Then it is time to turn to the LM317.

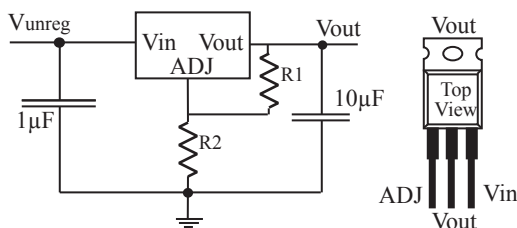


Figure 24-10 LM317 Adjustable Voltage Regulator

This is a 3-terminal device that needs only two resistors in addition to the usual transient suppression capacitors to regulate any voltage between 1.25V and 30V. Again, there are several different packages available from the TO-39 transistor package that can dissipate only 2W at currents up to 0.5A to the metal TO-3 case that can handle 20W and currents up to 1.5A. In this device the ground terminal is replaced by an adjust terminal that is held 1.25V below the output terminal. By choosing resistors R1 and R2 in Figure 26-10, the output voltage can be programmed to any value between 1.25V and $V_{in}-2V$.

The 1.25V placed across R1 makes a current $1.25/R1$ flow through R1. That current flows through R2 along with the adjustment current, I_{ADJ} , typically $50\mu A$. Because of these currents, the voltage at the adjustment pin is

$$V_{ADJ} = \left[\frac{1.25}{R1} + I_{ADJ} \right] \times R2$$

and the output voltage is

$$V_{out} = V_{ADJ} + 1.25V = 1.25 \times \left[1 + \frac{R2}{R1} \right] + I_{ADJ} \times R2.$$

In theory, you can choose R1 and R2 to give any voltage above 1.25V. In practice, the output voltage is limited by the unregulated input voltage and the dropout voltage. If you make one of the resistors, usually R2, variable then you have a variable output regulator. Even if you only need a fixed voltage, then you usually need to make R2 adjustable because most output

voltages cannot be reached using only standard value resistors. R2 is usually made a multi-turn trimpot.

24.7 A complete power supply design

We can combine an unregulated supply from chapter 9 with a 3-terminal voltage regulator to design a complete regulated 5V, 1.5 power supply. The only design decisions we have to make are the choice of power transformer and filter capacitor. We know that about 8V is ideal for the unregulated input to LM7805. From chapter 9 we know that the peak unregulated output voltage from a bridge rectifier/capacitor filter supply is given by

$$V_{pk} = 1.4 \cdot V_{RMS} - 1.2V$$

so we need a transformer with an RMS voltage of at least

$$V_{RMS} = \frac{8 + 1.2}{1.4} = 6.6V$$

Standard transformers are available with outputs of 6.3V or of 7.5V. The 6.3V is just a little too low. It would probably work most of the time but under worse case conditions the regulator would fail. Therefore we will use a 7.5V, 1.5A transformer. That will give us a peak voltage of

$$V_{pk} = 7.5 \cdot 1.4 - 1.2 = 9.3V$$

We can tolerate over 1V of ripple without risk of the regulator shutting down. The allowed ripple sets a minimum value for the filter capacitor. In chapter 9 we found that the ripple depth was

$$V_{ripple} = \frac{I}{120 \times C}$$

so we need a capacitor of at least

$$C = \frac{I}{120 \times V_{ripple}} = \frac{1.5}{120 \times 1} = 12500\mu F$$

For safety, and to improve the ripple performance, we will choose a 20,000μF, 16V filter capacitor. This can be an ordinary aluminum electrolytic capacitor since it does not have to react to rapid transient loads. The 10μF output capacitor, however, must be a high quality device, usually a tantalum electrolytic, because it does have to respond to the transients. The only other thing we have to add is a bridge rectifier and I chose an easily available bridge rectifier rated at 4A for safety. Here is our final design.

Info The RS401LMS is a single package containing four 4A diodes already connected as a bridge rectifier. It has 4 terminals; two AC input terminals marked '~' and the DC terminals marked '+' and '-'.

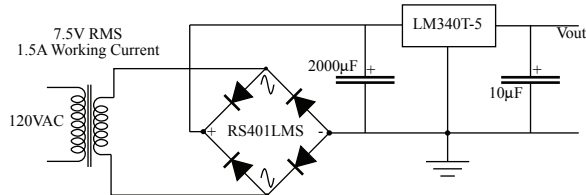


Figure 26-11 A Complete 5V Power Supply

Summary

A **regulated power supply** is one where the output voltage (and/or current) is controlled by circuitry to keep the output constant despite variations in the input supply or in the load. It consists of an unregulated supply plus a voltage reference and an amplifier circuit that controls the flow of current to the load in order to keep the output at the desired level.

The unregulated supply provides the power but suffers from the usual problems of output ripple and an output voltage that falls as the load current increases.

The **voltage reference** provides a stable, fixed voltage against which the output of the supply can be compared.

The amplifier circuit does the real work. It compares the output voltage (or current) to the reference voltage and adjusts the amount of current that is delivered to the load. If the output voltage (current) rises too high, the amplifier decreases the output current and the output falls back to its correct value. If the output voltage (current) drops too low then the amplifier increases the output current. The amplifier controls the output current by varying the gate voltage on the output pass transistor.

Voltage Reference

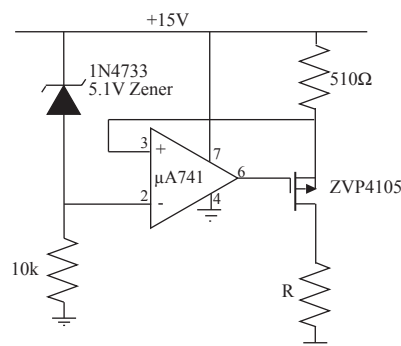
The most common voltage references are diodes.

Simple circuits use a Zener diode which has a well defined reverse breakdown voltage. This can provide a reference voltage anywhere from about 2V to about 150V. Zener diodes are cheap, easy to use, and available in a wide range of voltages but they are quite noisy.

More sophisticated circuits use the diode forward voltage drop, 0.6 for a standard Silicon diode, as a reference. This is quieter and can be controlled extremely well but is too low a voltage for most purposes so it has to be used with more external circuitry to cover a range of useful voltages.

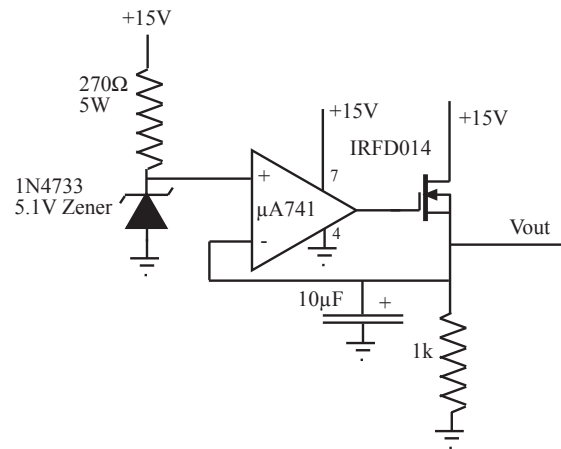
Exercises

- In this chapter we have seen power supply circuits that are designed to maintain a constant output voltage. We can also design circuits that will maintain a constant output current. There are several kinds of such circuit. Some can only source current (push current into the load), some can only sink it, and some can do both. The circuit below is a current source. Use the op-amp golden rules to find the current that flows into the load resistor R and show that it does not depend on the value of the resistor. That is what makes this a constant current source.



- What is the largest value that the load resistor R above can have if the circuit is to continue to function?
- Design, and give a complete circuit for, a voltage regulator to deliver 10V at up to 200mA. You may use any value of resistor that you like but may use only a 5.1V Zener as the voltage reference.

4. The figure below shows a voltage regulator that is designed to be able to supply up to 200mA of output current. Calculate the maximum power dissipated in the output FET and the maximum power delivered to the load.



5. A power supply is built using the circuit of Exercise 26.4. The output is connected to a high-power variable resistor. Sketch a graph of the voltage at V_{out} and the voltage at the gate of the FET as the variable resistor is turned down from its initial value of 1000Ω to a final value of 25Ω , assuming that the FET does not heat up significantly. Explain why your graphs have the shape that they do.

Chapter 25:Digital-to-Analog Conversion

25.1 Introduction

So far our computer can read the state of switches in the world and can turn switches on and off with its parallel ports. Now it is time to use those switches to generate some computer controlled voltages. **Digital-to-analog converters (DACs)** are quite straightforward devices made by crossbreeding a set of switches with an amplifier.

25.2 The DAC

To make a DAC we start with a summing amplifier and a set of switches as shown in Figure 25-1.

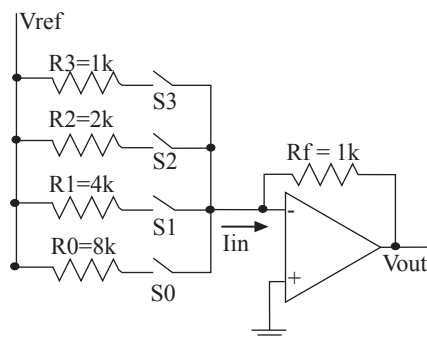


Figure 25-1 Simple Manual DAC

The magic lies in the values of the resistors, which match the binary weighting of the bits in a binary number. There are two parts to the circuit. On the right is a current-to-voltage (trans-resistance) amplifier made up of the op-amp and the 1k feedback resistor. We know that the output of this circuit will be $-R_f I_{in}$ where I_{in} is the current flowing in the input lead as shown. On the left we have a set of resistors and switches which form a current source. The left-hand end of each resistor is held at the reference voltage, V_{ref} , and can be connected by a switch to the virtual ground of the amplifier input.

When a switch is open, no current flows in its resistor. When the switch is closed, a current V_{ref}/R flows in its resistor. By Kirchoff's current law, the total I_{in} is the sum of the resistor currents.

We can calculate the output voltage for any configuration of the input switches. For example, the figure shows all of the switches open so that the current $I_{in} = 0$. If we introduce variables $S_0, S_1, S_2,$ and S_3 which are 0 if the corresponding switch is open and 1 if the switch is closed then the total current is

$$I_{in} = V_{ref} \times V_{ADJ} = \left[\frac{S_0}{R_0} + \frac{S_1}{R_1} + \frac{S_2}{R_2} + \frac{S_3}{R_3} \right]$$

$$I_{in} = \frac{V_{ref}}{8k} \times (s_0 + 2 \times s_1 + 4 \times s_2 + 8 \times s_3)$$

Thus the output voltage is

$$V_{out} = \frac{V_{ref}}{8} \times (s_0 + 2 \times s_1 + 4 \times s_2 + 8 \times s_3)$$

So the output voltage depends on the settings of the switches in exactly the same way that the value of a binary number depends on the setting of its bits. By setting the switches we can make every value from $V_{out} = 0$ to $V_{out} = 1.875 \cdot V_{ref}$ as shown in Table 25-1.

Table 25-1: DAC Output Voltages

S_3	S_2	S_1	S_0	V_{out}	V_{out} ($V_{ref}=5V$)
0	0	0	0	0·Vref	0
0	0	0	1	-1/8·Vref	-0.625V
0	0	1	0	-2/8·Vref	-1.25V
0	0	1	1	-3/8·Vref	-1.875V
0	1	0	0	-4/8·Vref	-2.5V
0	1	0	1	-5/8·Vref	-3.125V
0	1	1	0	-6/8·Vref	-3.75V
0	1	1	1	-7/8·Vref	-4.375V
1	0	0	0	-1·Vref	-5V
1	0	0	1	-9/8·Vref	-5.625V
1	0	1	0	-10/8·Vref	-6.25V
1	0	1	1	-11/8·Vref	-6.875V
1	1	0	0	-12/8·Vref	-7.5V
1	1	0	1	-13/8·Vref	-8.125V
1	1	1	0	-14/8·Vref	-8.75V
1	1	1	1	-15/8·Vref	-9.375V

In general, if we call the binary number N , then the output voltage corresponding to that number is given by

$$V(N) = \frac{-N \times V_{ref} \times 1k}{8k} = -N \times \frac{V_{ref}}{8}$$

The circuit shown in Figure 27-1 is our first digital-to-analog converter, a circuit that converts a binary number on a set of switches into a current which is proportional to the binary number and then to a voltage, also proportional to the number. It is a rather primitive DAC since it relies on a human to set the switches.

25.2.1 FETs replace Fingers

To get a more useful circuit we need to replace the manual switches by electronic switches. As we know, an FET makes a very good switch so we can use an FET in place of each switch and turn the switch on and off with a digital output bit. That gives us the circuit of Figure 25-2, which converts a digital input (D0-D3) into an output current.

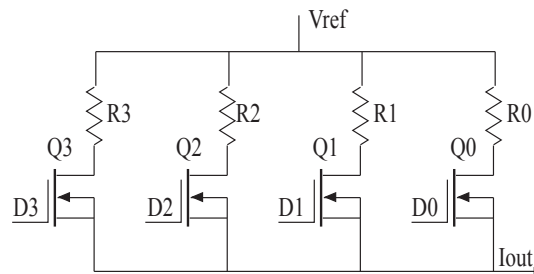


Figure 25-2 Simple 4-Bit Current-Output DAC

This circuit is called a Current-Output DAC because the output is exactly, a current whose value is set by the number on the switches. Current outputs DACs are useful in some high speed applications but we usually want a Voltage-Output DAC, a DAC whose output is a voltage proportional to the number on the switches.

To get a voltage output, we connect the current output to an op-amp connected as a current-to-voltage converter. Since the input is a virtual ground point, we can be sure that the source of each FET is at 0V. Thus, when a digital input is a 0 (0V) the corresponding FET is turned off and no current flows in the resistor. When a digital input is 1 (5V) the corresponding FET is turned fully on and current flows in the resistor. If we connect the digital inputs, D0-D3, to some bits of a computer parallel port, then we can convert a digital number from the computer into a voltage.

25.3 The R-2R Ladder

We have built a 4-bit converter using 4 resistors. The smallest resistor is only 1/8th the value of the largest one. That means that the largest resistor must be made very accurately. Its value has to be accurate to a fraction of the value of the smallest resistor. You could build an adequate 4-bit DAC with 1% resistors but not an 8-bit DAC. In an 8-bit DAC the smallest resistor is only 1/128th of the largest resistor so that you would need resistors with a precision of better than 0.07%. It is very difficult, and very expensive, to make resistors with that kind of precision and so a different circuit is usually used. It needs only two different values of resistors, R and $2R$, and is called an R-2R ladder. Here is the circuit for a 4-bit R-2R ladder.

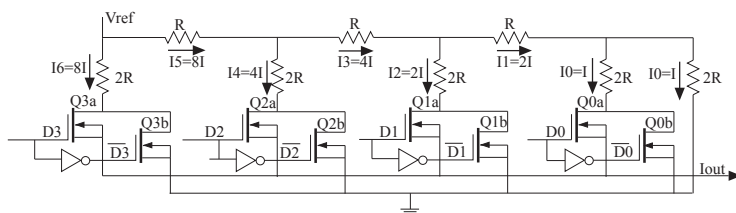


Figure 25-3 R-2R Ladder Current-Output DAC

Things have become a little more complicated. Let's look first at the 2 FETs making up each of the four switches. We'll use bit 0 for an example but each of the others is just the same.

- When D0 is a 0, D0 is a 1. That means that Q0a is turned OFF and Q0b is turned ON. The bottom of the resistor is connected to the bottom wire, which is connected to ground. Thus current flows in the resistor but it simply flows to ground and does not reach the output.
- When D0 is 1, D0 is a 0. In this case Q0a is turned ON and Q0b is turned OFF. This time the resistor is connected to the output wire. Now this is held at 0V by the virtual ground input of the op-amp so that *exactly the same current flows in the resistor* but this time the current flows to Iout instead of to ground.

No matter what the state of D0 the same current always flows in the resistor; the switch merely selects whether it flows to the output or to ground. This means that, so far as the current is concerned, the circuit behaves just like the simpler circuit of Figure 25-4.

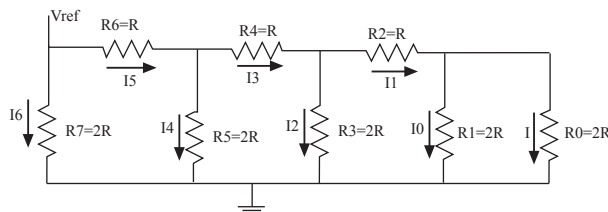


Figure 25-4 Simplified R-2R Ladder

We start analyzing this at the right hand end. The last two resistors, R0 and R1, are in parallel and so have the same voltage across them. Since they are equal in value, they also have the same current flowing in them. We shall call that current I. Two equal resistors in parallel have a total resistance of half the value of one of the resistors. Thus the two resistors together look like a single resistor R as we see in Figure 25-5.

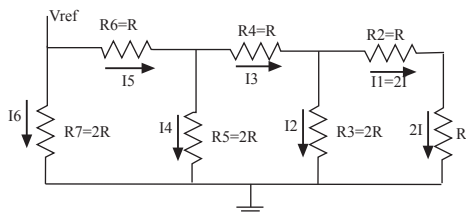


Figure 25-5 Simplified R-2R Ladder Step 1

The combined current 2I flows in this resistor which is in series with the original resistor R2. These two resistors together make up a resistor of value 2R and the whole circuit is equivalent to the circuit of Figure 25-6.

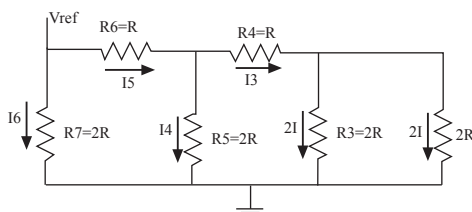


Figure 25-6 Simplified R-2R Ladder Step 2

If you compare this circuit to the original one then you will see that this is almost identical to the rightmost 3 sections of the original. The only difference is that the currents are all twice as big. You should now be able to understand the whole pattern. Equal currents flow in the two parallel $2R$ resistors so that $I_2 = 2I$. The sum of those equal currents flows in the R_4 so that $I_3 = 4I$. Together, the three right-hand resistors are equivalent to a single resistor of value $2R$ with a current of $4I$ flowing in it.

Again, the new $2R$ resistor is in parallel with the $R_5=2R$ and so $I_4 = 4I$. Both $4I$ currents flow in R_6 so that $I_5 = 8I$. The whole combination of all the resistors to the right of R_7 is again equal to a resistor $2R$ so that the current in R_7 is the same as that in R_6 and $I_6 = 8I$. Now we know that the voltage across R_7 is V_{ref} and so we have

$$8 \times I = \frac{V_{ref}}{2R}$$

which means that

$$I = \frac{V_{ref}}{16R}$$

Figure 25-7 shows the full circuit with all the currents drawn in place.

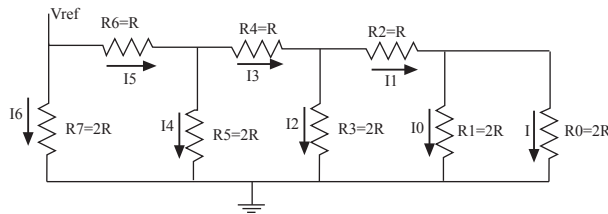


Figure 25-7 Simplified R-2R Ladder Showing Currents

Note The resistors of the R-2R ladder still need to be quite precise since the errors accumulate as you go along the ladder. In a 16-bit ladder the resistors have to be individually adjusted during production. This is done using a computer-controlled laser to adjust the resistors while the computer measures the output voltages from the circuit.

Like the first DAC circuit, we have a set of currents which form a binary sequence, $8I$, $4I$, $2I$, and I . Unlike the first DAC circuit, we have used only two values of resistor. Instead of having to be able to make a wide range of different precision resistors we only have to be able to make two kinds, R and $2R$.

25.4 Commercial DAC chips

Obviously a circuit as complicated as the R-2R DAC circuit is a candidate for making into a single chip. DACs are available in a variety of forms tuned for a range of different tasks. At the low end of the market there are simple 8-bit DACs with either voltage or current outputs. These are used in a wide variety of control applications, such as motor speed controllers, that need only moderate accuracy. Then there are higher precision 10- and 12-bit DACs that are used when more accuracy is required. These are found in such places as computer controlled power supplies and digital plotters. Then there are the 16 and more bit audio DACs that lie at the heart of all the digital audio devices around us, from CD players to digital cell phones and high quality music synthesizers. Finally, the most expensive DACs are the ultra-high speed ones found in digital video applications such as computer video displays. These can generate hundreds of millions of samples a second and can cost hundreds of dollars a piece. We shall look at a few common examples.

25.4.1 A cheap 8-bit DAC: the DAC0808

This is a long-time favorite for low-end applications. It is an 8-bit **current output** DAC that costs only \$2 a piece. It is the most primitive form of DAC containing only the R-2R ladder and current switches. You must add an external reference voltage and operational amplifier to make a complete DAC. Figure 25-8 shows a typical circuit.

The manufacturer claims an accuracy of 0.19% for this DAC meaning that the largest difference between any single output value and the correct value is only 0.19% of the full scale output. The output is guaranteed to settle to within $\pm 1/2$ of a least significant bit in only 150nS but that is the time for the current to settle. The final output voltage may take rather longer

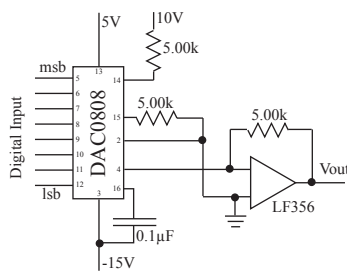


Figure 25-8 DAC0808 Typical Circuit

to settle because of the speed of the output op-amp. You can typically convert a few million samples-per-second with this system.

25.4.2 A precision 12-bit DAC: the DAC80P

For applications where more precision is called for, we turn first to the 12-bit DAC. The DAC80P is the industry standard 12-bit part, offering excellent performance at a reasonable price (about \$20). Unlike the DAC0808, the DAC80P has both the op-amp and the reference voltage built in so that no active components are required for voltage output operation. However, in order to achieve optimum performance over the full output range you do have to add some external trimming resistors.

This more precise DAC has a more extensive error specification as shown in the next section. It can generate a new sample every 2 μ S and so can operate at 500,000 samples per second.

25.4.3 A 16-bit Audio DAC: the PCM56P

At the heart of every CD player there is a digital-to-analog converter. It takes the stream of digital data coming from the CD and converts it into an analog voltage that is then amplified and sent to the speakers or headphones. The huge market for these devices keeps the prices of such converters very reasonable and a single PCM56P costs only \$12 in 1998. Like the DAC80P, the PCM56P has the reference voltage and op-amp built into the chip. Unlike the DAC80P, the input to the chip is in serial form rather than parallel. That is, the 16-bits are sent to the DAC one-at-a-time over a single input pin, rather than sending all 16 at once over 16 input pins.

An audio DAC does not have to operate at such a high frequency as some other applications since sound only contains information up to 20kHz. Most consumer audio systems run at a rate of 44,100 samples per second and the 1.5 μ S settling time of the PCM56P is short enough to allow operation at up to 4 times that speed.

25.5 Imperfections of DACs

Every real DAC is prey to imperfections in the three subsystems from which it is built; the R-2R ladder, the FET switches, and op-amp based trans-resistance amplifier. Let us look first at imperfections in the ladder.

25.5.1 Ladder problems

It is impossible to make resistors that are perfectly identical so that the resistors in the ladder will not be exactly the correct values and so the currents will not be exactly correct. Instead of having currents that follow the sequence

$$I, 2I, 4I, 8I, 16I, \text{ etc.},$$

a real DAC will have currents that follow a slightly imperfect rule such as

$$I, 1.9993I, 4.0021I, 7.9984I, \text{ etc.}$$

This means that the outputs will not be quite what they should be. The manufacturer of a DAC provides a lot of information about these imperfections in the data sheet. We'll examine the Accuracy section of the data sheet for the DAC80P as shown in Table 25-2.

Table 25-2: Accuracy Data for the DAC80P

Parameter	DAC80P			UNITS
	MIN	TYP	MAX	
Linearity Error		$\pm 1/4$	$\pm 1/2$	LSB
Differential Linearity Error		$\pm 1/2$	$\pm 3/4$	LSB
Gain Error		± 0.1	± 0.3	%
Offset Error		± 0.05	± 0.15	% of Full Scale

The **Linearity Error** is the maximum deviation between the ideal output voltage and the real output voltage measured under all conditions. It is the ultimate test of the accuracy of a DAC. It is usually specified, as here, in units of the least significant bit.

Note For a voltage-output DAC we can find the size of 1 LSB by dividing the total range of output voltage by the number of steps possible. For example, a 0-5V, 8-bit DAC has $1 \text{ LSB} = 5V/2^8 = 19.5\text{mV}$.

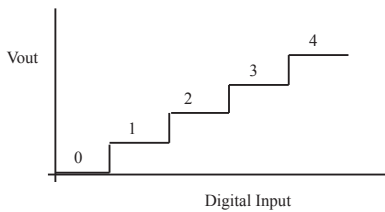


Figure 25-9
DAC Voltage vs. Digital Input

Note The specifications depend on temperature because the resistor values change slightly as the temperature changes. More details can often be found in the data sheets for individual devices.

The **Differential Linearity Error** is the largest difference between the actual size of a single step and the ideal size of a single step. Ideally every one of the steps in Figure 27-9 has the same height, 1 LSB. In reality, resistor errors lead to small variations in the step heights and these variations are measured by the Differential Linearity Error. Note that the Linearity Error is slightly better than the Differential Linearity Error because the errors are randomly distributed and so there is some cancellation.

If the Differential Linearity Error is $\pm 1/2$ LSB, then you must worry whether the output for an input value of $n+1$ is greater than the output for an input value of n . A DAC that can guarantee that the output always increases when the input value increases is said to be **Monotonic**. The manufacturer of the DAC80P guarantees that the chip is monotonic when operated over the temperature range from 0°C to $+70^{\circ}\text{C}$.

The **Gain** and **Offset** errors measure the difference between the slope and offset of the best straight line drawn through the output data on a plot like that of Figure 25-9 and the slope and offset of the ideal DAC. They tell you how closely the outputs of two different DACs of the same kind will match. Most DACs, including the DAC80P, provide pins for adjusting the Gain and Offset so that several DACs can be matched more closely than the basic specification would allow by trimming the errors in the circuit. That leaves only variations of the Gain and Offset with temperature to cause mismatch between the DACs. These drifts are typically a few ppm/ $^{\circ}\text{C}$ and so are of interest only in high precision circuits operated under extreme conditions.

25.5.2 Switch Problems

The next source of imperfection is the array of FET switches. Here the problem is that not all of the FETs switch at exactly the same time. When the input bits change state, there are slight differences in the switching times of the FETs. Thus the output does not change instantaneously from one correct level to the next but passes through unwanted intermediate states. Note that these intermediates need may well not be between the correct states. Once you jumble the bits of a binary number, you can get values from all over the place. At its worst, this leads to **glitches**; at the very least it adds to the **settling time** (see below). Although a DAC driven by the sequence 0, 1, 2, 3, 4 should produce an output that looks like Figure 25-9 the output from a real DAC might look like Figure 25-10.

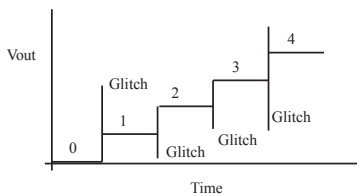


Figure 25-10 DAC Glitches

Glitches are of very short duration, often only a few nS, but can be quite large. Some (expensive) DACs have special deglitching circuitry to remove the glitches. Most DACs just specify a limit to the energy in the glitches and rely on filtering of the final signal to get rid of the effects of the glitches.

25.5.3 Op-Amp Problems

Lastly, there are the imperfections introduced by the current-to-voltage amplifier. Because this is based on an op-amp, it suffers from all of the problems of op-amp circuits. First, there is the combined effect of the input errors that produces an offset voltage. This means that when you tell the DAC to produce a 0V output it actually produces a different, small, voltage; the **offset error**. The manufacturer of the DAC usually specifies an upper limit for this offset and often provides terminals to trim the voltage to zero.

Second, there are the problems arising from the finite speed of the op-amp, which are so similar to the problems of unequal FET switching time that they can be combined in a single characteristic. The combination of the finite bandwidth of the op-amp, the slew-rate limit of the op-amp, and the unequal switching time of the FETs is usually specified in terms of the **settling time** for the DAC. This is the time it takes for the output to come within some region of its correct value following a change in the input. A common specification is time to settle within $\pm 1/2$ lsb of its final value. A current output DAC, having no op-amp to slow it down, may have a settling time as low as a few nS for an expensive video DAC, though a few

hundred nS is a more common and affordable value. A voltage output DAC will usually be slower—values in the range 1-2 μ S are common for cheaper parts such as the DAC80P.

25.6 Some DAC examples

A DAC is used anywhere a computer or other digital circuit has to control a continuously variable output. The obvious examples are sound cards for computers and other digital musical instruments. Other uses are as diverse as a video converter for a computer display, a modern digitally tuned radio or television, or a lighting controller for a whole theater. Here are a couple of simple examples.

25.6.1 Arbitrary waveform generator

An arbitrary waveform generator is a very fancy kind of signal generator. Instead of making only the simple sine, square, and triangle waves of a normal analog signal generator, an arbitrary waveform generator can generate waves of any shape and play them at any of a wide range of frequencies. The idea is to represent the desired shape by a set of numbers, each one giving the level of the waveform at an instant, and then to play the whole waveform out through a DAC at a computer-controlled rate. The complexity of the waveform that can be represented depends on the number of points in the waveform and on the number of different voltage levels available.

Commercial examples tend to be expensive and extremely flexible devices but we can design a basic one with a small computer and a DAC. We shall use a fast 8-bit DAC, giving us 256 different voltage levels, and 1k of memory, giving us 1024 different output points. It is tempting to try to use the computer to read through the memory and send the data to the DAC using a loop but this turns out not to work. The problem is that there is always some overhead to a loop. Even an infinite loop has to have a branch instruction to take it back to the start of the loop. That means that every so often you have to insert this extra instruction so the output points are not perfectly evenly spaced. Instead, we will use a set of counters to drive the RAM and connect the RAM outputs directly to the DAC inputs (Figure 25-11).

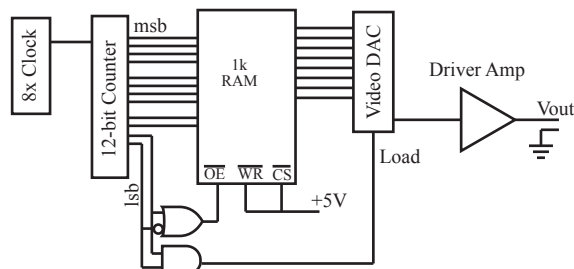


Figure 25-11 Arbitrary Waveform Generator

I have shown only the waveform generator part of the circuit. There must also be a computer part, which shares the same RAM so that there is a way to get the data into the RAM. Once the RAM is loaded, the circuit is completely self-contained.

The chip is wired so that it is always enabled and accepting address inputs. The bottom two bits of the counter are decoded to provide the output enable signal for the RAM and load signal for the input register of the DAC. The remaining 10 bits make up the address for the RAM. As the counter counts, the circuit goes through the following sequence

- 1) Change the address
- 2) Wait through the RAM access time
- 3) Send the OE signal to the RAM
- 4) Wait while the RAM outputs settle
- 5) Load the data into the DAC and generate the next voltage
- 6) Repeat from 1

Figure 25-12 shows how the negative true OE signal and the positive true Load signal are generated from the bottom two bits of the counter output.

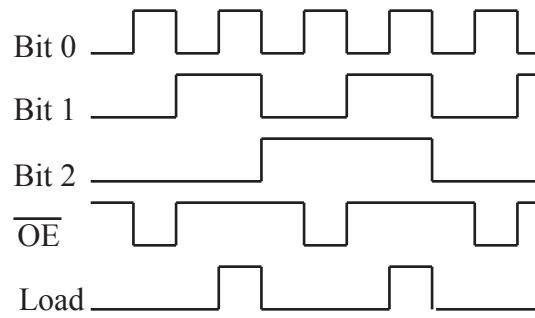


Figure 25-12 Signal Timing for Waveform Generator

The maximum frequency of this system is limited by the speed of the video DAC. With a fast DAC such as the DAC600 the system could operate at 256 million samples per second, running from a clock at 2GHz and producing a complete waveform every 4 μ S. Thus, if a single wave takes up all 1024 samples, the maximum output frequency from a 256MHz DAC would be only 256kHz!

25.6.2 Greenhouse climate control

This application is at the opposite end of the time scale from the first example. There we used an extremely fast video DAC to generate waveforms in the kHz range. Here we can use as slow a DAC as we like since conditions inside a greenhouse change very slowly, on a time-scale of minutes. The system we are to control consists of a greenhouse with a stove and a mechanical blind system that can cover the windows if the sun is too fierce. The heater can only be turned on and off but the blind system can be set to any position. The larger the voltage applied to the blind controller, the further down the windows the blinds go.

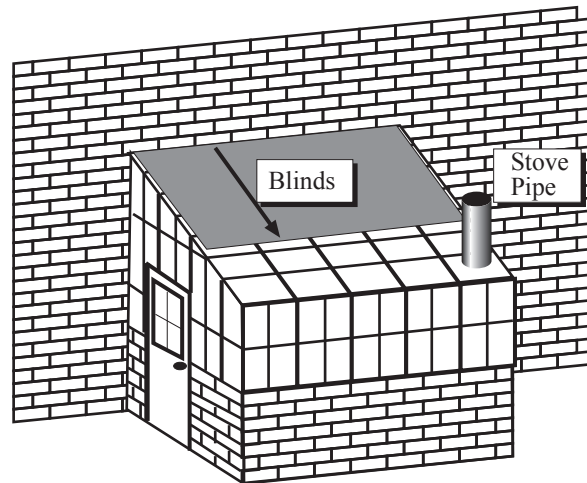


Figure 25-13 Greenhouse

The controller will need to borrow a little from the next chapter because it needs to measure temperature in order to control it. In fact, the computer has two temperature sensors; one for the inside of the greenhouse and one for the outside. Figure 25-14 is a circuit based on an MC9S08 single chip microcomputer.

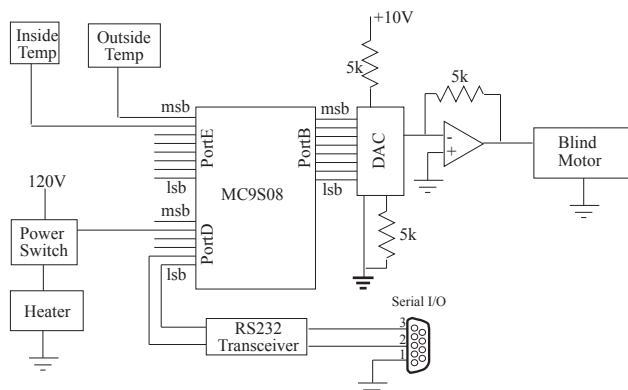


Figure 25-14 Greenhouse Controller

We drive the voltage output DAC from port B, which is always an output port. We drive the power switch from one of the upper bits of port D since we do not need the serial peripheral port. We use the bottom two bits of port D as the SCI input and output bits so that we can program the thermostat from an external computer or a video terminal. Finally, we use two bits of port E as analog inputs reading the voltages from the temperature sensors.

A commercial controller would probably have an elaborate front panel so that you could alter the temperature set points and might well allow you to set several different set points for different times of day. I have left all this out to keep the example simple. Instead of a front panel I have provided a serial link so that an external computer, called a **host computer**, can tell the controller what to do. The host sends short commands consisting of a single letter and a number. For example, to set the operating temperature you might send the command “T18” to set the temperature to 18°C.

The complete program is too long to show here (and also uses quite a lot of material from the next chapter) but the overall flowchart is shown in Figure 25-15. Note that the outside temperature is not used in the control algorithm but the host computer can ask about the outside temperature. It sends an “O” command and the controller responds by reading the outside temperature sensor and sending the temperature back over the serial link. Similarly, it can read the inside temperature by sending an “I” command.

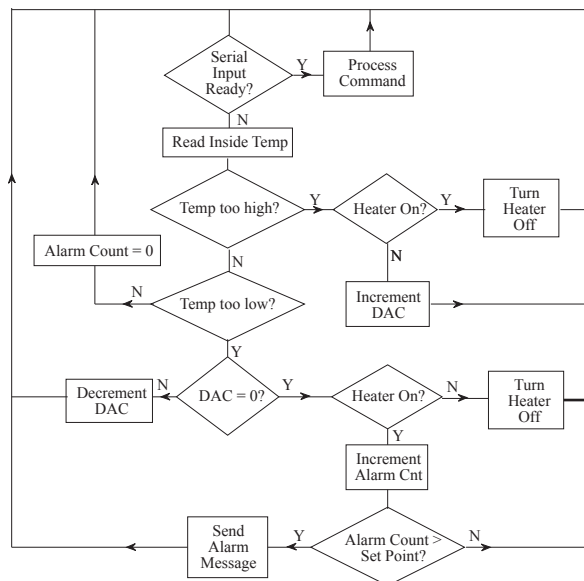


Figure 25-15 Greenhouse Controller Flow Chart

Summary

A **Digital-to-Analog Converter** or **DAC** takes a multi-bit digital input and converts it to an analog voltage. The analog voltage is proportional to the value of the number.

The most common form of DAC works by adding together a set of binary weighted currents. Each bit of the incoming binary number controls a switch that either includes or excludes the relevant current from the sum. The basic output is then a current that is proportional to the binary number. This is called a **current output DAC**.

If we take the current output of the DAC and feed it to a trans-resistance amplifier then we generate an output voltage proportional to the current and so to the original number. This is called a **voltage output DAC**.

An N-bit DAC produces 2^N different output voltages. For a unipolar DAC these usually range from 0V to some maximum V_{max} . In this case, if the incoming number is m then the output voltage is

$$V_{out} = m \times \frac{V_{max}}{2^N - 1}$$

Various different conventions allow DACs to produce bipolar (both positive and negative) output voltages.

DACs are not perfect devices. The output voltage does not exactly fit the formula. We characterise the DAC errors in several ways. If the ideal formula for the output voltage due to an input number m is

$$V = A \times m$$

then the real output is

$$V = A' \times m + O + E_m$$

Gain Error: This is the difference between the real slope of the gain curve (A') and the ideal value (A). It is expressed as a percentage of A .

Offset Error: This is the constant error term O . It is expressed as a percentage of the full scale output.

Differential Linearity Error: This is the difference between the ideal size of a step and the real size. It involves both the gain error and the step error, E_m . It is usually expressed as a fraction of the step size.

Linearity Error: This is the maximum difference between the real output voltage and the ideal value. It is the best measure of the overall quality of the DAC. It is expressed as a fraction of the step size.

Monotonicity: This is not a separate measurement but a convenient rough summary. A DAC is **monotonic** if $V_{m+1} > V_m$ for all values of m .

Chapter 26: Analog-to-Digital Conversion

26.1 Introduction

We have seen how a digital-to-analog converter allows a digital circuit to generate continuously varying voltages and so communicate with its users in a variety of natural ways. It can generate sounds and show pictures, can control motor speeds, and can move the heads of a hard disk. Many of these modes of communication should be two-way. The circuit should be able to take a voltage signal from the outside world and convert it into a digital signal for processing. This is the job of the **Analog-to-Digital converter** or **ADC**. Such converters are found in a huge variety of devices from digital cell-phones to sampling keyboards, from the temperature measuring side of an electronic thermostat to the video input port on some modern computers, from radar signal processing systems to the humble digital voltmeter.

The whole process of conversion from an analog signal to a digital one is an information destroying process. You start off with a signal that can take any real number value within some range and end up with one of a finite, usually small, number of discrete output numbers. Thus the output will almost be only an approximation to the input. It can be useful to think of a simple geometrical process that mimics the behavior of analog-to-digital conversion. We start with a line that represents all the possible input voltages. Here is a line for a converter that can handle voltages between 0V and 5V.



Figure 26-1 0-5V Conversion Line

We divide this line up with a ruler into a set of N equal subdivisions. For example, we might divide the line into 8 parts, like this.

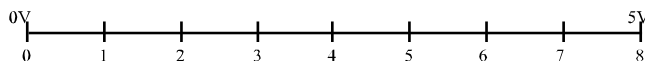


Figure 26-2 0-5V Line with 3-bit conversion

Now we can perform a conversion. We take the incoming voltage and plot it as a point on the line. So an input of 2.37V looks like this.

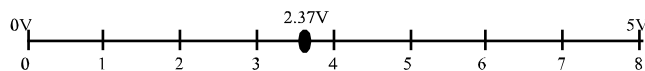


Figure 26-3 3-bit Conversion of 2.37V

Then we look for the ruling that is, in some sense, nearest the point and return the number on that ruling as the output of the conversion. In this case the nearest ruling to the point is ruling number 4. So the conversion takes an input of 2.37V and produces an output of 4. This makes an error, but we can be sure that error is no larger than $1/2$ the distance between rulings. We can make the error as small as we like by making the number of rulings large enough. We say that the more rulings there are, the higher is the **resolution** of the converter. In normal practice we make the number of division a power of 2 and describe the resolution by giving the power. Thus, we speak of an 8-bit converter that has 256 different possible outputs and makes a maximum error of only $1/512$ of the total conversion range.

There are two rather different kinds of ADC, one for high-speed applications requiring thousands of conversion every second and the other for low speed, high accuracy applications requiring only a few conversions per second. This chapter is about the fast kind, they are the ones that usually connect to computers.

26.2 Flash Conversion

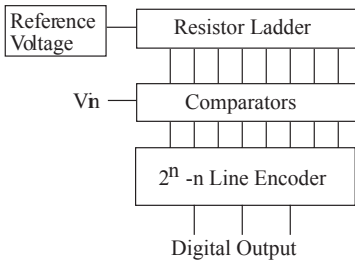


Figure 26-4 Flash ADC

The fastest ADCs use a very simple technique called **flash** conversion. This is the simplest kind of ADC to understand but it is one of the most expensive to construct, especially when the output requires a large number of bits. Figure 26-4 shows a block diagram for a flash ADC. A resistor ladder generates a set of equally spaced test voltages from a stable, precise reference voltage. A set of comparators compares the incoming voltage, V_{in} , with each of the test voltages. This results in a set of binary outputs. All of the outputs corresponding to test voltages that are less than V_{in} are logic 1s, all the rest are logic 0s. This binary bit pattern goes to a set of gates, which encode the number of the most significant 1 bit onto the output. Thus the output is a binary number representing the input voltage as a fraction of the output voltage.

This kind of converter is very fast. The time between a change in V_{in} and the corresponding change in V_{out} is only the settling time of the slowest comparator plus the gate delay through the encoder. The total can be as low as a few nS so that flash converters are found in the highest speed systems, systems such as digital oscilloscopes and video converters.

This kind of converter is also very expensive and impractical to manufacture in large numbers of bits. An n -bit converter (that is a computer that outputs an n -bit binary number) requires 2^n resistors and, much worse, 2^n comparators. Thus, to build even an 8-bit flash ADC would require 256 comparators, each of which consists of a dozen or so transistors. This makes such a system very expensive so that Flash ADCs are usually limited to no more than 8-bits.

26.3 A complete 3-bit flash ADC

In order to understand the operation more fully let us look at the complete circuit for a 3-bit flash ADC as shown in Figure 26-5.

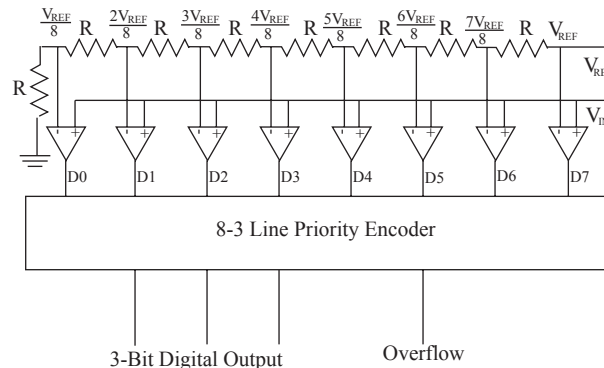


Figure 26-5 3-Bit Flash ADC

The resistive divider chain takes the reference voltage, V_{REF} and divides it into 8 equal portions giving a set of test voltages

$$\frac{V_{REF}}{8}, \frac{2V_{REF}}{8}, \frac{3V_{REF}}{8}, \frac{4V_{REF}}{8}, \frac{5V_{REF}}{8}, \frac{6V_{REF}}{8}, \frac{7V_{REF}}{8}, \frac{8V_{REF}}{8}$$

Each comparator receives one of these test voltages and compares it to the input voltage, producing a 1 if $V_{IN} >$ Test Voltage and a 0 otherwise. This places the input voltage into a category by setting all the bits that correspond voltages δV_{IN} and clearing all the bits corresponding to voltages $>V_{IN}$.

Example

If the input is $V_{IN} = 0.36 \cdot V_{REF}$ then bits D0 and D1 will be set and bits D2-D7 will be clear.

The comparator outputs go to the inputs of an 8-line to 3-line priority encoder. This device creates a 3-bit binary output corresponding to the bit number of the most significant 1 bit in its input. So, if the input is %00011111 then the output will be %101 = 5 and if the input is %00000001 then the output will be %001 et cetera. Because there are 8 input lines and only three outputs it is possible for there to be no way to represent the input and when this happens

an Overflow flag is set. If we look back at the comparator chain, then we can see that this will happen only if $V_{IN} > V_{REF}$ that is, if the input voltage is out of range.

In order to complete this example we need to look inside the priority encoder. First we need a truth table for the device (Table 26-1). This truth table is rather abbreviated because there are a lot of input states that can never occur because of the way the comparator chain works. If input D_n is a 1 then we can be certain that inputs $D_{n-1} \dots D_0$ are also 1 since if $V_{IN} > nV_{REF}/8$ then it is also greater than all voltages smaller than that!

Table 26-1: 8-3 Line Priority Encoder

D_7	D_6	D_5	D_4	D_3	D_2	D_1	D_0	Q_2	Q_1	Q_0	Q_3
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	1	1	0	0	1	0
0	0	0	0	0	1	1	1	0	0	1	1
0	0	0	0	1	1	1	1	0	1	0	0
0	0	0	1	1	1	1	1	0	1	0	1
0	0	1	1	1	1	1	1	0	1	1	0
0	1	1	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	1	1	0	0	0

This gives us the logic equations

$$\begin{aligned}
 O_v &= D_7 \\
 Q_2 &= \overline{D_7} \times D_6 + \overline{D_6} \times D_5 + \overline{D_5} \times D_4 + \overline{D_4} \times D_3 \\
 Q_1 &= \overline{D_7} \times D_6 + \overline{D_6} \times D_5 + \overline{D_3} \times D_2 + \overline{D_2} \times D_1 \\
 Q_0 &= \overline{D_7} \times D_6 + \overline{D_5} \times D_4 + \overline{D_3} \times D_2 + \overline{D_1} \times D_0
 \end{aligned}$$

which lead to the fairly simple circuit of Figure 26-6.

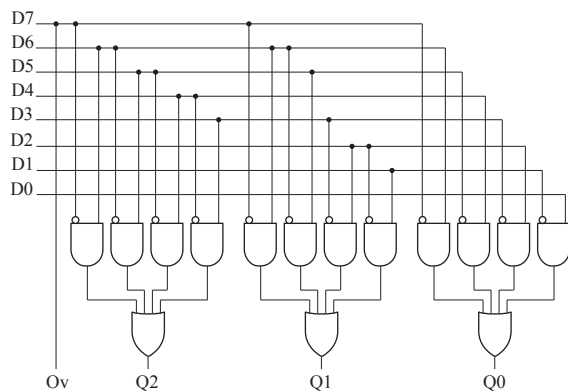


Figure 26-6 8-3 Line Priority Encoder Circuit

Obviously these are not devices that you build from discrete components but are fabricated as complete integrated circuits.

26.4 Successive Approximation Conversion

The flash ADC is simply too expensive or too limited in resolution for most applications and there are a variety of other circuits in use that trade speed for affordability and resolution. We shall examine the most common of these, the successive approximation converter.

The successive approximation converter is built around a single comparator instead of the 2^n comparators needed by the flash converter. One input to the converter is the incoming voltage and the other is taken from the output of a digital-to-analog converter. Logic within the converter adjusts the voltage from the DAC until it is as close as possible to the input voltage. Then it outputs that binary number as its best approximation to the input, see Figure 26-7.

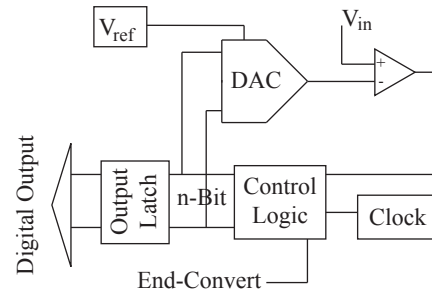


Figure 26-7 Successive Approximation ADC

Since the conversion is a multi-step process, with a series of comparisons taking place one-after-another, the control logic must be a sequential system. The ADC has an extra output to tell the rest of the world when it has finished performing a conversion and has a new sample ready to be read. This output is usually **End-Of-Convert** or **EOC**.

There are two different sequences that an ADC can use to find the best approximation to a given input voltage. The simplest, linear search, is also the slowest and is suitable only for educational use; all commercial converters use the more complex binary search.

26.4.1 Conversion by Linear Search

The most straightforward way to operate a successive approximation conversion is to start from a value of 0 and increase the value 1 step at a time until you reach a value that is just greater than the input voltage. You can tell when this happens by monitoring the output of the comparator. So long as the DAC voltage is less than V_{IN} , the comparator will output a 0. As soon as the DAC voltage exceeds V_{IN} , the comparator output will change to a 1 and you can stop the process and output the current value as the best approximation to the input. We call this a **linear search**. Thinking about our line analogy it amounts to starting at the left-hand end of the line and stepping along the line until we pass the sample point. Then we stop and return the current point as the answer. This means that the sample value is always on the high side but the average error is still 1/2 the sample spacing. If the incoming samples are spread uniformly over the whole sample line then, on average, it will take 2^{n-1} comparisons to find the sample. Thus, the conversion time is proportional to the number of sample states, 2^n for an n-bit converter.

26.4.2 Conversion by Binary Search

There is a much more efficient searching method that we can use called a **binary search**. This method works a little like the procedure that you might use to find a given page in a book. You would not start at page one and start flipping pages until you reached the correct one. Instead, you would make a guess, open the book at a reasonable place, and look at the page number. If the page you found was too low then you would jump forward some distance and try again. If the page number you found was too high then you would go back a few pages and try again. You would repeat this process until you found the page.

The binary search is a systematic version of this process. You start by comparing the input voltage with a voltage right in the middle of the range. If the test voltage is $<V_{IN}$, then you know that V_{IN} lies in the top half of the test range. Otherwise, you know that V_{IN} lies in the bottom half of the test range. Now you can repeat the process using the appropriate half range and figure out which quarter of the range holds V_{IN} . A third step would allow you to find in which eighth of the range V_{IN} lay and so on. Thus, after 3 comparisons you would have a conversion accurate to 1 part in 8, after 4 comparisons to 1 part in 16 and so on. In general, you can search 2^n sample states in only n conversions. This is **MUCH** more efficient than the linear search.

Example

An 8-bit converter requires an average of 128 conversions in a linear search but only 8 conversions in a binary search.

Commercial analog-to-digital converters use a synchronous state-machine like the ones that we studied in Chapter 17 to perform the successive approximation conversion. They are quite large machines with many states and so I will not give the details here. For example, the controller for an 8-bit ADC has 8 flip-flops, 256 states, and a 9-input, 8-output set of gates controlling the selection of the next state. Instead of looking at all the details, we will content ourselves with working our example from the introduction using the binary search algorithm. Here is our example again (Figure 28-8).

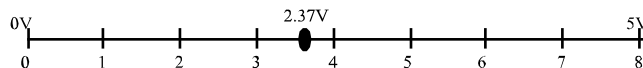


Figure 26-8 3-bit conversion example: Step 1

We start by trying the middle number. That is we send an output of $4 = 100_2$ to the DAC and compare that output voltage with V_{IN} . When we send the number 4 to the DAC it generates a voltage $= 4 \cdot 5V/8 = 2.5V$. Since $2.37V < 2.5V$ the comparator returns a 0 and we know that the sample lies in the lower half of the input line. That means that the first digit of the answer is a 0 so that the result can be written $\%0xx$ at this point.

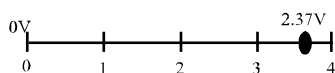


Figure 26-9 Step 2

Now we repeat the process with this lower half of the line. The center of the line is at 2 so we send 010_2 to the DAC which generates the voltage $2 \cdot 5V/8 = 1.25V$. This is $< 2.37V$ so the comparator outputs a 1. Now we know that the number lies in the second quarter of the line so that the top two bits of the number are 01_2 and the number must have the form $01x_2$.

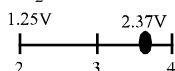


Figure 26-10 Step 3

The process repeats exactly as before. We set the next bit in the number, giving us a test value of 011_2 , and send that value to the DAC. The DAC generates a voltage of $3 \cdot 5V/8 = 1.875V$, which is then compared to V_{IN} . Since $1.875V < 2.37V$ the comparator output is 1 giving us a current approximation of 011_2 . Since there are only three bits in the converter we have finished and this number is the binary search approximation to $2.37V$. Note that it is NOT the same answer that we got from the linear search. The linear search will always give an answer that is too large, the binary search will give an answer that is too large half the time and too small half the time. The average error remains $1/2$ the sample interval.

26.4.3 Sample-and-Hold

The successive approximation process suffers from one major problem. If the input value changes during the time that the conversion takes, then the answer will be wrong. We must arrange that the input voltage remains stable during the conversion or we will get answers that are less accurate than the converter resolution would suggest. The answer is a circuit called a **Sample-and-Hold**, which does just what its name suggests; it takes a sample of the input voltage and holds it steady for some length of time. It is an analog memory.

At the heart of a sample-and-hold circuit is a capacitor (Figure 26-11). The circuit charges the capacitor with the input voltage and then disconnects the capacitor from the input. So long as there is no path for the charge on the capacitor to leak away, the capacitor will hold the charge, and thus the voltage, forever. We can read off the voltage without drawing charge from the capacitor using a low-bias current op-amp. The process is controlled by a digital signal driving an FET switch. When the signal is high (Sample), the switch is closed and the capacitor charges to V_{IN} . When the signal goes low, the switch opens and the capacitor is isolated from the input. It now holds the voltage at the non-inverting input of the op-amp steady and so keeps the output voltage equal to the voltage that was captured from the input. Because real

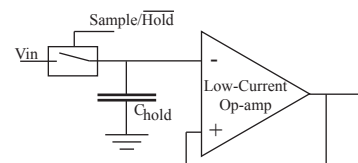


Figure 26-11 Sample-and-Hold

op-amps have non-zero bias currents and real FET switches have non-zero leakage currents the stored voltage will slowly drift as charge leaks off the capacitor. However, this process can be made very small over the time of interest, usually only a few μS . Although you can buy special purpose sample-and-hold chips and can build the circuits from discrete components, you rarely need to. They are so important to the operation of successive approximation ADCs that it is common practice to build the sample-and-hold onto the same chip as the ADC.

26.5 Imperfections of ADCs

Analog-to-digital converters are as subject to error as any other electronic circuit. In particular, they are subject to errors arising from the internal DAC and no ADC can be more accurate than the DAC at its heart. Even a flash ADC is not immune to this effect. Although it has no explicit DAC inside it, the resistor ladder is essentially a multi-output DAC and suffers from the same error problems that all DACs do.

Most of the specifications of DACs apply equally to ADCs so that the accuracy of an ADC is specified in terms of its linearity error and differential non-linearity. In addition, you will see the term **no missing codes** used of ADCs. This corresponds to the monotonicity of the underlying DAC. If the internal DAC is not monotonic, then at some point the output voltage falls as the input number rises. At that point, there will be a region of digital codes that can never be output by the ADC. In most cases this is merely another manifestation of the error but in some applications it is a real problem. In such an instance, you must choose an ADC that has no missing codes.

Apart from the error, an ADC suffers from two other limitations. First, there is the conversion speed. Except for flash ADCs, this is much less than the speed of a DAC of similar cost and it gets worse in higher resolution devices. Every bit added to the output code is another comparison time added to the conversion time and so high-resolution devices tend to be slower than lower resolution ones. Although DACs can operate happily at millions of samples per second, ADCs are normally at least ten times as slow.

26.6 IC Converters

The market for analog-to-digital converters is large and growing all the time, particularly in the 16-bit audio converter segment. I am going to look at just a couple of examples, a simple 8-bit device suitable for most low-accuracy moderate speed needs and a 16-bit converter optimized for use in audio systems such as digital audio tape recorders and high-end digital samplers.

26.6.1 A low-cost 8-bit converter: the ADC0801

This is the most accurate (and the most expensive) of a family of similar 8-bit devices, the ADC0801, ADC0802, ADC0803, ADC0804, and ADC0805. Each member can perform a conversion in $100\mu\text{S}$, very slow compared to a low-priced DAC, but they differ in accuracy, temperature range, and price. The ADC0801 is the most expensive device (\$20) and offers $\pm 1/4$ LSB accuracy over the -40°C to $+85^\circ\text{C}$ temperature range. Other devices offer reduced accuracies of $\pm 1/2$ LSB and ± 1 LSB and reduced temperature range, 0°C to 70°C but they cost less; the cheapest costing only \$5.

The accuracy specification deserves a little more explanation. We know that an ideal ADC makes errors of up to $1/2$ LSB just because of the information loss in the conversion to a fixed number of bits. This may be clearer if we look at Figure 26-12. Figure 26-12a shows the transfer function for an ideal ADC, that is, the digital output plotted as function of the analog input voltage. It clearly shows how each digital output code corresponds to a small range of analog inputs and is only correct at the center of each little range. If we plot the error as a function of input voltage then we get the graph of Figure 26-12b.

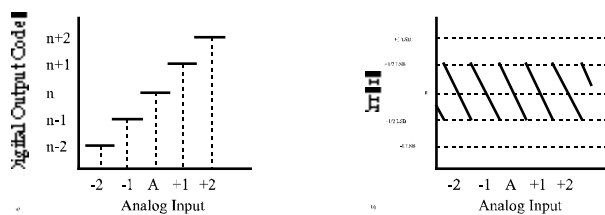


Figure 26-12 Transfer Function and Error for Ideal ADC

In Figure 26-12b we see that the error ranges from $-1/2$ LSB to $+1/2$ LSB round each central value so that the maximum error we ever make is only $1/2$ LSB. A non-ideal ADC has input steps of slightly different widths, corresponding to the imperfections in the values of the resistors in the DAC chain. Its transfer function looks more like Figure 26-13a. The uneven spacing of the steps leads to a non-uniformity in the error distribution producing a plot like Figure 26-13b. Here we see how the uneven widths of the steps leads to errors greater than $1/2$ LSB. The accuracy specification sets a limit on the Extra Error term. Thus the largest extra error that an ADC0801 will make is $1/4$ LSB.

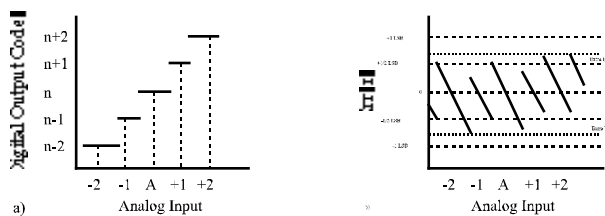


Figure 26-13 Transfer Function and Error for Real ADC

The ADC080x converters do not have an internal reference voltage so that you must supply a reference or derive one from the power supply voltage. Using an external reference guarantees higher absolute accuracy but increases the cost and is a poor choice in some applications so that you can use the power supply voltage as a reference. This means that you need a good, stable, quiet power supply but has advantages when working with a signal that is derived from the power supply voltage.

Example

You can make a simple temperature sensor from a resistor whose value varies with temperature (a **thermistor**).

If you connect the thermistor in series with a stable resistor, one whose value is insensitive to temperature, as in the figure, then the voltage V_o varies with the temperature and can be used to measure the temperature. The trouble is that V_o also varies with the power supply voltage V_+ . We can use the voltage divider equation to see that

$$V_o = \frac{R_{Th}}{R_{Th} + R} \times V_+$$

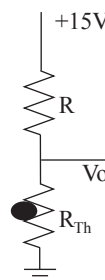


Figure 26-14 Temperature sensing circuit

If we use an absolute voltage reference in ADC then variations in the power supply voltage will lead to variations in the output even through the temperature stays constant. If we use the power supply as our reference source then the variations in power supply voltage will change ADC steps in exactly the same way that they change the input voltage and so the errors will cancel. This is called **ratiometric conversion** and is the correct way of measuring any signal that is a fraction of the power supply voltage.

The ADC080x converters are **unipolar** devices. That is, they can convert only voltages of one polarity; positive. The input range is controlled by the choice of reference voltage. The total input range will be twice the voltage applied to the $V_{REF}/2$ pin. The most common case is to set $V_{REF}/2$ to 2.5V, half the power supply voltage, either with a resistive divider from the power supply for ratiometric use or with a 2.5V voltage reference IC. In this case the conversion range is 0-5V.

The ADC080x converters do not have sample-and-hold built in and so are suitable only for use with slowly varying signals. Given that it takes 100 μ S to make a single conversion, you can see that a signal should not vary more than a few percent per second. Signals such as DC voltages and room temperatures are easy to measure but anything moving much faster will need an external sample-and-hold.

The ADC080x converters are designed specifically to work with microprocessors and have extra digital circuitry to make it easy to interface the converter to a small microcomputer. Full details are found in the data sheet.

26.6.2 A 16-bit audio converter: the PCM78

At the other end of the price and performance spectrum from the \$5 ADC0804 is the PCM78, a \$50 16-bit converter that can perform 200,000 conversions per second. This is a bipolar device, with a conversion range from -3V @ +3V.

Instead of having its error described only by an absolute error range, the PCM78 has a rather complex set of error characteristics that are designed to describe its performance in generating audio signals. The overall accuracy for a single sample is described by giving the Gain and Offset errors, which seem quite high in comparison to the ADC080x specification. The total gain error is given as $\pm 2\%$ compared to the total error of the ADC0801, **which** was $\pm 1/4$ LSB or $\pm 0.1\%$. The converter is not designed for great absolute accuracy—nobody cares whether an audio signal volume level is in error by 2%—instead it is designed to have the code-to-code errors as small and as evenly distributed as possible. That makes the quality of the output sounds as high as possible.

Most of the error data describe the quality of the sine waves that the device can produce under various circumstances. For example, the **Total Harmonic Distortion** is given as -90dB for converting a 10kHz sine wave at a rate of 200,000 samples per second. That means that if you split the output into an ideal value and an error then the energy associated with the error is 90dB smaller than the energy associated with the ideal signal. That is, the error energy is only 1 billionth of the signal energy!

The output of the PCM78 is not a set of 16 digital data lines. Instead, it is a two-line serial interface. The bits representing the answer are sent one-at-a-time as the conversion proceeds with a clock signal to tell the receiving computer when each bit is ready (a synchronous serial interface, see Computer Chapter 9). This is perfect for a successive approximation converter since the converter generates the bits one at a time as it searches through the sample space. Each bit is sent to the computer as soon as it is ready and does not have to wait for the conversion to complete. One advantage of this scheme is that you can trade resolution for conversion speed. If you can accept a 12-bit representation of the signal instead of a 16-bit one, then you do not have to wait until the 16th bit is converted. You can tell the converter to quit and start another sample as soon as you have received the 12th sample. This is called **short cycling** the converter and can gain a small increase in conversion speed. For example, while a 16-bit conversion usually takes 4 μ S to perform, a 12-bit conversion could be performed in only 3 μ S.

26.7 ADC Examples

The uses of ADC are so varied that it is hard to do more than hint at the variety of applications in a single chapter. However, here are two examples of the sort of thing that you can do with an ADC. I have chosen one example for each of the converters that we just looked at.

26.7.1 Electronic Thermostat

A thermostat is a device for regulating temperature by turning on and off a heating or cooling device (or both). Originally these devices used an ingenious scheme wherein temperature changes caused a mechanical device, a bi-metal strip, to bend and so make and break and electrical contact. If the room was too warm then the strip bent more and opened the contact, turning the furnace off. If the room was too cool then the strip straightened, closing the contact

and turning the furnace back on. Not only was this device somewhat inaccurate, it was also fixed. If you wanted to alter the room temperature on a regular basis then you had to remember to reset the thermostat every time.

A modern thermostat is an all electronic device. It has a temperature sensor, an ADC, a switch for the furnace, and a small computer controlling the whole device. The advantage is that, with a handful of switches and a small LCD display, you can program the computer to alter the temperature automatically at various times. A good quality electronic thermostat may have four or five temperature settings every day. It may even have a different set for each day of the week so that it can turn the heat down on Thursday evenings, because you go out to choir that night, but not on any other day. It can turn the heat on at 6:30 am during the week but leave it turned down until 8 am on weekends if you usually sleep late on weekends. It is much more flexible.

It does not take an elaborate computer to perform this sort of task. A small, slow, ultra-low power computer with a few input/output lines will suffice. Many such small computers are made, lots of them with their own ADCs built-in, but I will assume the use of a processor that does not have its own ADC so I can use an ADC0804 for my ADC. Figure 26-15 is a block diagram of the system.

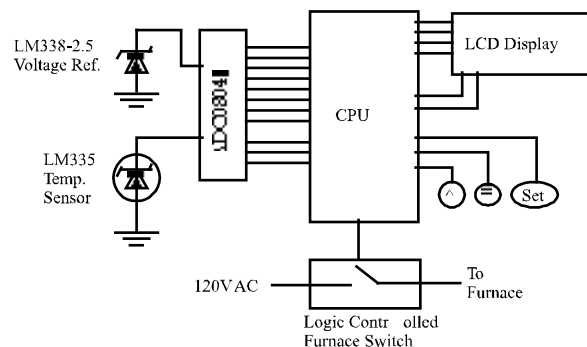


Figure 26-15 Thermostat Block Diagram

The Computer sits reading the current temperature and comparing it to the current set point. If the temperature is too low, then it turns the furnace on. If the temperature is too high, it turns the furnace off. In order to reduce wear and tear on the furnace, the temperature at which the furnace turns on is set a few degrees cooler than the temperature at which the furnace turns off. This gives the system a degree of hysteresis, providing a range of temperature over which the furnace is turned on so that the furnace runs for a few minutes to raise the temperature then shuts off for some length of time while the house cools down.

At the same time as it is controlling the furnace, the computer shows the current time-of-day and current temperature, and possibly the set temperature, on the LCD. It also sits waiting for action on the input switches. When a switch is pushed, the computer responds by entering a command state in which it allows the user to alter the internal settings. The whole system can be powered by a small battery and run for months without needing new batteries.

26.7.2 Sampling Keyboard

The past twenty years has seen a huge rise in the popularity and availability of low-to-moderate priced electronic keyboards. One of the more recent features of such keyboards is sampling. In this mode, the keyboard makes a digital recording of a sound and then plays the sound back at various different frequencies when a key is pressed. At the low end, this results in keyboards that allow children to play simple tunes with dog barks. At the high end it allows sophisticated keyboards to use samples from real grand pianos to sound just like the real thing while occupying a tenth the space and costing a fortieth the price.

At the heart of a sampling keyboard are an ADC and DAC controlled by a small computer. The block diagram, Figure 28-15, is quite simple. A voltage representing the input sound comes in the line marked Audio In. This might come from a microphone or could come from

This is an example of an Anti-Aliasing filter.

another audio source such as a CD player. The sound is passed through a low-pass filter to remove any frequency components that are too high for the ADC to record. The ADC converts the sound into a set of numbers that are sent to the digital signal processor (a microcomputer that has some special support for operations on time varying signals such as audio signals). The digital signal processor (DSP) stores the numbers in its internal memory as a “sample”.

The keyboard microcontroller spends its time looking at the keys (and at any other control buttons on the keyboard; I left them off the diagram) and when the user presses keys on the keyboard tells the DSP which keys have been pressed. The DSP takes the information on which keys are held down at any instant and uses them to decide how to play back the sample. If a low note key is pressed, then the DSP plays the samples back out of memory in such a way as to play the sampled sound at a low frequency. If the user pressed a high note, then the DSP plays the samples back out at a high frequency.

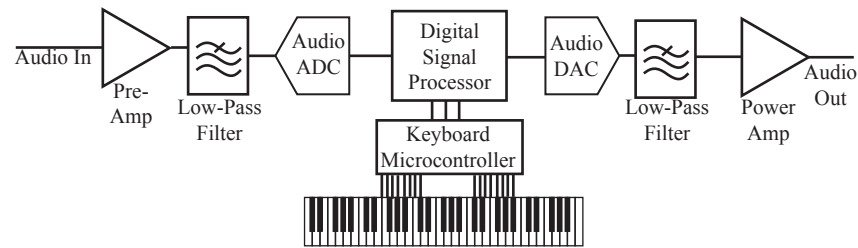


Figure 26-16 Sampling Keyboard Block Diagram

A good DSP can handle the simultaneous playback of several samples at once, mixing the signals together to get a compound sound. Thus the user can play chords and hear all the notes sound at once because the DSP can manipulate the numbers fast enough. The output of the DSP is a stream of numbers representing the output signal level. This stream is sent over a serial link to the Audio DAC, which converts the digital numbers back into voltages. The second low-pass filter removes the little steps that result from the digitization processes and then sends the signal to an amplifier and so to the outside world where it is either replayed over a speaker, played over headphones, or mixed with other signals to make up a larger sound.

A filter used this way on the output of a DAC is often called a **reconstruction filter**.

Summary

An **analog-to-digital converter (ADC)** is the opposite of a digital-analog converter. It takes an analog voltage as an input and outputs a digital number. The number produced is proportional to the incoming voltage.

A **flash ADC** uses a large set of comparators and some digital logic to perform extremely fast conversions. An N bit converter uses 2^N comparators to split the analog voltage range up into 2^N slices and simultaneously compares the incoming voltage with all of the slices. Such a converter is very fast but difficult to make with more than 4 bits and impractical beyond 8-bits. Such a converter can make 10^7 - 10^9 conversions per second corresponding to conversion times as low as 1nS.

A **successive approximation ADC** uses a single comparator, an N bit DAC, and some control logic. It compares the incoming voltage to a series of DAC voltages and outputs the closest approximation. It takes N individual comparisons to convert a single voltage to an N -bit number using a binary search implemented in the control logic.

ADCs suffer from the same sort of error problems that DACs do, largely because they use a DAC to generate the comparison voltages.

Gain Error: This is the difference between the real slope of the gain curve (A') and the ideal value (A). It is expressed as a percentage of A .

Offset Error: This is the constant error term O . It is expressed as a percentage of the full scale output.

Differential Linearity Error: This is the difference between the ideal size of a step and the real size. It involves both the gain error and the step error, E_m . It is usually expressed as a fraction of the step size.

Linearity Error: This is the maximum difference between the real output voltage and the ideal value. It is the best measure of the overall quality of the ADC. It is expressed as a fraction of the step size.

Monotonicity: This is not a separate measurement but a convenient rough summary. An ADC is monotonic if $V_{m+1} > V_m$ for all values of m .

Chapter 27: Power Switches

27.1 Power Switches

Despite their greater size and power handling, power switches are typically easier to understand than logic switches, so we shall look at them first. At its simplest, an FET power switch consists of a single FET connected in series with a load and driven directly from the input signal, as we saw in Chapter 12. We shall look at the limitations of this circuit (Figure 27-1) in some detail so that we can see what improvements need to be made.

The first limitation is simply the limitation of the FET that we have chosen. According to the data sheet, the 2N7000 can withstand a drain-source voltage of at most 60V, can carry a drain current up to 200mA continuously, and can dissipate up to 350mW. This clearly limits the size of load that it can drive. If we want to switch larger loads then we must choose a more powerful FET. Some examples of the range of power FETs that is available is shown in the sidebar.

<Sidebar on power FETs>

The second limitation is the drive—the voltage available to turn the FET on. In this case, we had plenty of voltage to saturate the FET since the saturation current for a 2N7000 with a 5V gate-source voltage is 380mA, well above the 100mA required. However, if we tried to drive this circuit with a TTL signal, a common kind of logic signal, which is only guaranteed to get up to 2.4V we would be in trouble since that is below the threshold voltage for this FET. In that case we would have to use an FET with a much lower threshold voltage and that could be at odds with the choice of a high power FET since they often need quite large voltages to get high saturation currents. For example, the 4A, 100V IRF510 power MOSFET is rated with a threshold of 4V and requires $V_{GS} = 10V$ for full power operation. How would we build a switch driven by our 0-5V signal that controls a 60V, 2A heater?

27.2 Moderate Power Switch

We have to use a small-signal FET switch to switch a separate gate drive voltage for the power FET. Figure 27-2 shows a circuit that would work.

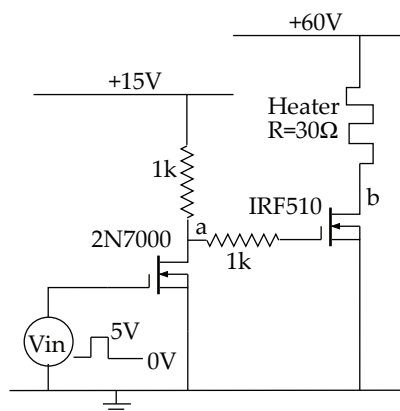


Figure 27-2 Moderate Power Switch

When the input voltage is 0V, the 2N7000 is turned off so that the voltage at point a rises to +15V, no current flowing into the gate of the IRF510. That is sufficient gate drive to turn the IRF510 fully on. Since the saturation current of the IRF510 is 4A and thus greater than the current drawn by the heater resistance, the FET acts as a resistance of about 0.6Ω and the voltage at point b falls to nearly 0, turning on the heater. When the input voltage is +5V, the 2N7000 turns on fully so the voltage at point a falls nearly to 0V and turns off the IRF510. Now the

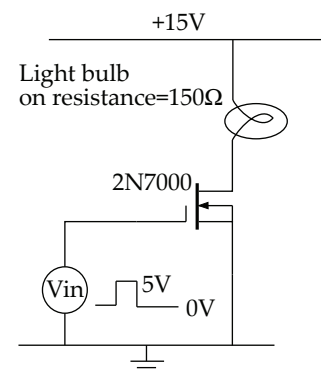


Figure 27-1 Simple Power Switch

Note The 1k resistor in the gate lead of the IRF510 is often included to protect the FET from excess current or voltage arising from static charges. This is something that MOSFETs are *very* sensitive to. A single static zap will destroy a power MOSFET so it is very common to try to protect the gate. In fact, you will often see a resistor of about 100k connected between the gate and source as well to give a static charge another path to take.

current through the heater falls to zero. This circuit controls the much higher-power load from the same signal, though at the cost of inverting the signal. Where 0V originally meant OFF, it now means on and 5V now means OFF. As we shall see when we look at logic circuits, this inversion is not usually a problem.

Note that in this circuit the IRF510 is carrying a fairly significant current and dissipates a power of $I^2R = 2^2 \cdot 0.6 = 2.4\text{W}$. That is high enough that the transistor would need to be connected to a metal **heat sink** to help carry away the heat or it would get too hot and be damaged.

<Sidebar on heat dissipation and heat sinks.>

The last limitation is also associated with driving the FET. Ideally the load will turn on and off instantaneously when the input signal turns on and off. In practice it takes time to turn the FET on and off. The principle cause of this is the capacitance between the gate and the drain and source. When the gate voltage changes current must flow in the gate lead to charge these capacitors and it takes time for that current to flow through the gate resistor and the internal resistance of the signal source. For example, when the 2N7000 turns off, the voltage at point a has to rise to 15V and the gate capacitance of the IRF510, 180pF, has to charge through the total of 2k of resistance. That process has a time constant of $RC = 0.36\mu\text{S}$ so that the power FET will take a few times that, about $1\mu\text{S}$, to turn on. Moreover, at the start of the turn-on when the gate is still at 0V, a current of $15\text{V}/2\text{k} = 7.5\text{mA}$ flows in the input resistors which contrasts strongly with our usual thought that the gate of FET draws no current. During the switching process the gate of an FET may draw a significant current and the signal source has to be able to supply that current!

Info If we look at the turn-on process in more detail we find that we have to take account of the gate-drain capacitance as well as the gate-source capacitance. When the gate voltage changes current has to flow in the gate-source capacitance and we can calculate the current easily,

$$I_g = V_{GS} \cdot dV_{GS}/dt$$

We can roughly calculate the average gate current by saying that it takes about $1\mu\text{S}$ to turn on the FET so that the gate voltage rises 15V in $1\mu\text{S}$ so that

$$I_{gs} = 180\text{pF} \cdot 15\text{V}/\mu\text{S} = 180\text{pF} \cdot 15,000,000\text{V/s} = 2.7\text{mA}.$$

At the same time current must flow to the gate-drain capacitance but this is more complicated because the drain voltage changes at the same time. In this case, in the time that the gate voltage goes from 0V to 15V, the drain voltage goes from 60V to 0V. This means that the gate-drain voltage goes from -60V to +15V, a change of 75V. Thus the average charging current associated with the gate-drain capacitance of 15pF is

$$I_{gd} = 15\text{pF} \cdot 75\text{V}/\mu\text{S} = 15\text{pF} \cdot 75,000,000 = 1.1\text{mA}$$

This means that, although the gate-drain capacitance is usually quite a lot smaller than the gate-source capacitance the current needed to charge it is quite comparable and the total current is often twice the current that you would predict from the gate-source capacitance alone.

27.3 Switching capacitive and inductive loads

When the load of a power switch is not purely resistive there are extra problems to consider. If the load has a capacitive portion then the current needed to charge that capacitance must be included in the current handling of the FET. If the load has an inductive portion then self-inductive effects can make the voltage across the FET do some remarkably nasty things.

First we will look at the effect of capacitive loads, for example long wires connecting the switch to the load will add a lot of capacitance. Generally there are two effects. First the switching time is increased by the time it takes to charge the capacitance. Second the extra charging current may increase the size of FET required. Let us look at the effect of moving our light bulb a long way away so that we add 200pF of capacitance across the lamp (Figure 27-3).

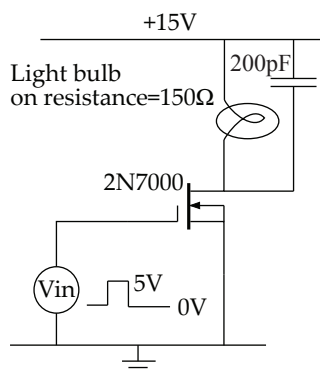


Figure 27-3

The switching time for a 2N7000 is listed on the data sheet as 10nS. So that, if the driving signal is of low enough impedance to charge the input capacitance in this time, then the transistor will turn on in 10nS, that is its resistance will go from infinity to about 5Ω in that time. The lead capacitance will then have to charge through the channel resistance with a time constant $RC = 5\Omega \cdot 200\text{pF} = 1\text{nS}$. Thus, the leads do not slow down the turn-on. However, when the FET turns off, the lead capacitance has to discharge through the resistor with a time constant of $RC = 150\Omega \cdot 200\text{pF} = 30\text{nS}$ and the turn-off is significantly slowed by the lead capacitance. Moreover, during turn-on, the peak charging current is $15\text{V}/5\Omega = 3\text{A}$, which is 6 times the maximum transient current that the device can withstand so we would need to either slow the switching down deliberately or use a much larger switching FET. So when switching capacitive loads we always have to compute the current required to charge the FET and add that to the load current to find the peak current rating of the switching FET.

An inductive load presents a different problem. Just as a capacitor resists changes in the voltage across it, so an inductor resists changes in the current flowing through it. In fact, an inductor generates a voltage

$$V = -L \times \frac{dI}{dt}$$

between its ends when the current flowing through it changes. In this L is a measure of the strength of the inductor that depends on the number of turns and their geometry. Its unit is the Henry. It is quite easy to see this effect. When you turn off a light switch there is often a little flash of light at the switch itself. That happens because you are trying to turn off the current flowing in the inductance formed by the lamp and all the wires leading to it. The inductance responds by generating a large voltage and a spark jumps across the contacts of the switch producing a flash of light. It is not unusual to see spikes of several thousand volts produced this way.

This behavior of an inductor has two effects in a switching circuit. First, it limits the speed with which you can turn on the current through an inductor. If you have a power supply of V volts and an inductive load of L Henries then the minimum time to turn on a current I through the inductor is $t = L \cdot I/V$. There is a second more important effect. When you turn off the FET the current in the FET falls very quickly and this creates a large voltage across the FET. Moreover, the voltage is in the reverse direction so that the FET end of the inductor may drop to tens of volts negative with respect to ground. This can profoundly stress the FET; it is essential to stop it happening. We usually protect the FET by putting a diode in parallel with the inductor as shown in Figure 27-4.

The protection diode must be chosen to be able to handle the peak current from the inductor. That is given by the peak voltage/inductor resistance. If the peak current is too high then a resistor can be placed in series with the diode at the expense of a slower turn-off. The time constant for the current to decay in a circuit with a resistor and an inductor is L/R where R is the sum of the external resistance, the diode forward resistance, and the internal resistance of the inductor.

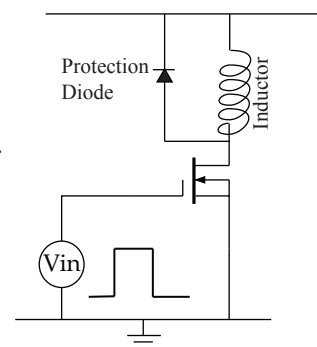


Figure 27-4

